Chapter 3: Kappa Statistics

David Eubanks

Table of contents

1 Introduction

The formula for chance-corrected measure of agreement (generically a "kappa") compares observed match rates to the expectation of random match rates. The kappas vary in how they estimate the random match rates. For two ratings to match, two raters j, k of the same subject i must agree in their assignment of either Class 1 or Class 0 classifications. In other words, the binary random variables must agree: $C_{ij} = C_{ik}$. A generic formula that includes the most common kappas is

$$\kappa = \frac{m_o - m_c}{1 - m_c},$$

where m_o is the observed proportion of agreements and m_c is the expected proportion of agreements under chance. The assumption about m_c is a defining feature of the various kappa statistics. The most general treatment of such statistics is the Krippendorff alpha (Krippendorff, 2018, pp. 221–250)

The various kappas differ in the assumption made about the chance correction probability m_c. Commonly, the assumption is that $m_c = x^2 + \bar{x}^2$ for some probability x. This simple formulation makes sense when both raters are guessing, but the actual case is more complicated

because a match "by chance" could be a case where one rating was accurate and the other was a guess. This distinction isn't generally made in the derivations of the kappas, although the AC1 paper discusses the issue, and hints at a full three-parameter model. It's ironic that the confusion about kappas is disagreement about the probability of agreement by chance.

The Fleiss kappa (Fleiss, 1971) uses the fraction of Class 1 ratings c to create $m_c = c^2 + \bar{c}^2$. The S statistic Bennett et al. (1954) (also called the Guilfords' G or G-index Holley & Guilford (1964)) is a kappa that assumes $m_c = 1/2$ when there are two categories. The AC1 kappa has a different form, assuming that $m_c = 2c\bar{c}$ Gwet (2008). The Cohen kappa is a variation where each rater gets a guessing distribution, so $m_c = x_1 x_2 + \bar{x_1}\bar{x_2}$ Cohen (1960).

Consider two raters classifying an observation. In the t-a-p model we can express the expected value of observed matches m_o as the sum of three kinds of agreement: (1) m_a is when both raters are accurate (and hence agree), (2) m_i when both raters are inaccurate (guessing) and agree, and (3) m_x is the mixed case when one rater is accurate and the other is inaccurate but they agree. The second two of these have expressions that include the guessing rate m_c. Following that thinking we have the following expectations for rates:

$$
\begin{aligned}
m_a &= a^2 && \text{(both accurate)} \\
m_r &= p^2 + \bar{p}^2 && \text{(random ratings)} \\
m_i &= \bar{a}^2 m_r = a^2 m_r - 2a m_r + m_r && \text{(both inaccurate)} \\
m_x &= 2a\bar{a}(tp + \bar{t}\bar{p}) && \text{(mixed accurate and inaccurate)} \\
m_o &= m_a + m_i + m_x && \text{(observed match rate)} \\
&= a^2 + a^2 m_r + m_r - 2a m_r + 2a\bar{a}(tp + \bar{t}\bar{p})
\end{aligned}
\tag{1}
$$

For m_a, both ratings must be accurate, in which case they automatically agree. For m_i, both must be inaccurate (probability $\bar{a}^2$) and then match randomly (probability m_r). For m_x, one rater must be accurate and the other inaccurate, in which case they agree if the accurate rater chooses the category that the inaccurate rater guesses. The various kappa derivations usually ignore these mixed matches in favor of using m_r as the chance match rate, which we called m_c in the kappa formula. This amounts to choosing p since $m_r = p^2 + \bar{p}^2$.

The various match rates in Equation 1 create a vocabulary for understanding some of the kappa statistics. The easiest one to analyze is the S-statistic (it is sometimes called the G-index).

2 Naive Raters: The S-Statistic

Recall that rater agreement for a binary choice is when there are two equal sized groups of raters that assign each of the two categories. This would be the case, on average, if the raters were flipping coins to assign categories. In the t-a-p model, this 50% chance for each category defines the p parameter. The S-statistic Bennett et al. (1954) makes the assumption that

if there are two categories to choose from, then the chance of a random match between two raters is $m_c = m_r = 1/2$. If we assume this means two *random* raters in the t-a-p model, we have $p = 1/2$. I call this the naive rater assumption, because it assumes that inaccurate raters are not influenced by the actual proportions of the two categories. For example, if an inexperienced doctor repeatedly diagnosed patients as having a very rare condition, this would be "naive" in the meaning here.

Substituting $p = 1/2$ into the formulas of Equation 1 results in the kappa

$$\begin{aligned}\kappa_s &= \frac{m_o - m_c}{1 - m_c} \\ &= \frac{a^2 + a^2/2 + 1/2 - a + 2a\bar{a}(t/2 + \bar{t}/2) - 1/2}{1 - 1/2} \\ &= 2(3a^2/2 + 1/2 + -a + a - a^2 - 1/2) \\ &= a^2\end{aligned}$$

In this case, the intuition from the introductory chapter that we're interested in something like the square root of rater agreement is exactly right. If the raters really do assign random ratings as if flipping a coin, then the resulting kappa derived from the data will have as its expectation the accuracy squared. Actual results will also have estimation error, depending on sample size and how unlucky you are.

3 Unbiased Raters: the Fleiss Kappa

The Fleiss kappa is designed to work with ratings where the number of raters per subject can vary, and with a rating scale of arbitrary length. It assumes an asymptotic form for chance correction, so is most appropriate for large samples. I will only consider the binary scale case here for simplicity.

The baseline for random ratings for Fleiss is if we took all the ratings and randomly shuffled them between subjects. In this case, the match rate for two raters is given by the proportions of the ratings for the two classes. For example, if Class 1 ratings comprise 20% of the total, then the random match rate is $m_c = .2^2 + .8^2$. That's the probability that either two random Class 1 ratings match or two random Class 0 ratings do. If c is the expected proportion of Class 1 ratings, then $m_c = c^2 + \bar{c}^2$. From the t-a-p diagram, we can see that $c = t(a + \bar{a}p) + \bar{t}\bar{a}p = ta + \bar{a}p$. If inaccurate ratings assign Class 1 at the true rate so that $t = p$, I'll describe the raters as "unbiased." In that case $c = pa + p\bar{a} = p = t$; the rating proportions of Class 1 reflect the true rates, because the raters assign proportionate "guesses" for inaccurate ratings. Under this assumption, with $t = p$ and $m_c = m_r = p^2 + \bar{p}^2$, kappa becomes

$$
\begin{aligned}
\kappa_f &= \frac{m_o - m_c}{1 - m_c} \\
&= \frac{a^2 + a^2 m_r + m_r - 2am_r + 2a\bar{a}(p^2 + \bar{p}^2) - m_r}{1 - m_r} \\
&= \frac{a^2 + m_r(a^2 + 1 - 2a + 2a - 2a^2 - 1)}{1 - m_r} \\
&= \frac{a^2 - a^2 m_r}{1 - m_r} \\
&= a^2.
\end{aligned}
$$

For the Fleiss kappa, it is also true that the expectation of kappa is the accuracy squared, this time if the condition $t = p$ is met by the raters represented in the data you have.

A review of the properties of Fleiss kappa can be found in Fleiss et al. (2013), chapter 18, including kappa's equivalence to an intraclass correlation coefficient, defined as ICC(1,1) in Shrout & Fleiss (1979). Under the $t = p$ "unbiased" condition, rater accuracy a is the correlation between the ratings and the true classifications: $\sqrt{\kappa_f} = a = cor(C, T)$. Additionally, the Fleiss kappa is the intraclass correlation of the ratings. Derivations of these results are found in Appendix A, where there is also an alternative derivation of the $a = \sqrt{\kappa_f}$ result.

If the $t = p$ assumption is not true for your raters, then the resulting kappa will be biased. In some cases, kappa may be negative. This is a general problem for the kappas, since they make assumptions about the distribution of random raters without testing those assumptions.

4 AC 1

The AC1 version of kappa developed in Gwet (2008) uses a disaggregation of rating agreements found in Table 4, page 36, where the author distinguishes between ratings that are certain (the same as "for cause" as found in Landis & Koch (1977)) versus random. The stated goal is to estimate the probability that two raters match when they both rate accurately ("with certainty"). This is a^2 in the t-a-p model (equation 16 in the paper). Random matches are those in which at least one of raters rates randomly. In terms of the t-a-p model, this assumes that the probability of a by-chance agreement is $m_c = m_i + m_x$, the cases where either both raters match randomly or at least one makes an inaccurate rating and they match. The acknowledgement of partially inaccurate matches is a philosophical advance over previous derivations of kappa.

The AC1 kappa is derived from the usual formula (total match rate - estimated random match rate) / (1 - estimated random match rate) (see equation 17 in the paper). The problem, as usual with this approach, is to estimate the (partially) random match rate $m_c = m_i + m_x$. Of course, neither m_i nor m_x are directly observable, so the author derives an approximation in

two steps. First the probability of agreement between two raters, at least one of whom made an inaccurate rating, is assumed to be 1/2. In the t-a-p model, this amounts to

$$Pr[\text{match}|\text{random}] = \frac{m_i + m_x}{1 - a^2} \approx 1/2. \tag{2}$$

To remove the denominator requires multiplying by the probability of at least one random rating, which is then approximated by $4c\bar{c}$, where c is the probability of a rater assigning Class 1 to a subject, which can be estimated directly from the data. Since from the t-a-p diagram of conditional probabilities, $c = ta + \bar{a}p$, the estimation entails assuming that

$$Pr[\text{random}] = 1 - a^2 \approx 4(ta + \bar{a}p)(1 - ta - \bar{a}p)$$

Multiplying these gives (page 37) $\hat{m_c} = 2c\bar{c} = 2(ta + \bar{a}p)(1 - ta - \bar{a}p)$. The approach here is ingenious and philosophically rich, but the limitations of a one-parameter index for rater agreement limit how much can be done.

In our notation here, the Fleiss kappa assumes $m_c = c^2 + \bar{c}^2$, and the AC1 assumes that $m_c = 2c\bar{c}$. The AC1 version is the complement of the Fleiss version: they sum to one since $1 = (c + \bar{c})^2 = c^2 + 2c\bar{c} + \bar{c}^2$. In some sense, the assumptions about the two kappas are opposite: what Fleiss considers random, AC1 considers non-random, and vice-versa, at least in expectation.

5 Discussion

Both the worst-case match rate for binary ratings, $p = 1/2$, and the proportional (unbiased) rate $t = p$ lead to kappas that have a nice relationship to accuracy in the t-a-p model, but only when the respective assumption about raters is true. Generally we don't know what p is for a given data set, however, so assuming either of those conditions is a leap of faith.

We might wonder if there are other kappas that have the nice property that accuracy is the square root. We can attempt to choose m_c so that $\kappa = a^2$ via $m_o - m_c = a^2(1 - m_c)$. Solving for m_c and using $m_o - a^2 = m_i + m_x$ leads to

$$\begin{aligned} m_c^* &= \frac{m_i + m_x}{1 - a^2} \\ &= \frac{\bar{a}^2 m_r + 2a\bar{a}(tp + \bar{t}\bar{p})}{(1 + a)\bar{a}} \\ &= \frac{\bar{a}(p^2 + \bar{p}^2) + 2a(tp + \bar{t}\bar{p})}{1 + a} \end{aligned} \tag{3}$$

where the asterisk denotes the choice of the chance correction formula m_c that makes $\kappa = a^2$. In the first line of Equation 3, the numerator is the expected proportion of matches where there is at least one inaccurate rating, and the denominator is the the rate of non-perfect rating pairs, where at least one of the raters is inaccurate (they may or may not match). We saw this above in the derivation of AC1 in Equation 2. It turns out that the correct choice of m_c is the conditional probability of a match given that at least one of the raters is inaccurate, rather than the unconditional probability of a match given that at least one of the raters is inaccurate.

The chance correction is therefore accounting for the accurate ratings by taking them out of the data altogether and then calculating inaccurate matches out of all rating pairs as the probability of by-chance matching.

The formula in Equation 3 is useful for testing properties of kappa assumptions. We can use it to verify that $p = 1/2$ (naive raters) and $t = p$ (unbiased raters) works as shown earlier.

In practice, it's better to just derive all three of the t-a-p parameters instead of making assumptions that have to be tested (by deriving all the parameters).

Bennett, E. M., Alpert, R., & Goldstein, A. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, *18*(3), 303–308.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. john wiley & sons.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Holley, J. W., & Guilford, J. P. (1964). A note on the g index of agreement. *Educational and Psychological Measurement*, *24*(4), 749–753.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.