Appendix A: Statistical Details

David Eubanks

Table of contents

1 Correlation of Ratings between Raters **1**
1.1 Correlation Between Ratings and True Values 2

2 Alternate Derivation of Fleiss Kappa Relationship **3**

1 Correlation of Ratings between Raters

Given two distinct raters i and j with common accuracy a and guess probability p, what's the correlation between their ratings? Let $c = E[C_i] = E[C_j] = ta + p\bar{a}$. Capital letters denote random binary variables, so that A_i is one if the first rater made an accurate assessment and zero if not. T is the true value of a common subject being rated. The covariance between the two raters' ratings is

$$
\begin{aligned}
\operatorname{Cor}(C_i, C_j) &= \frac{\operatorname{Cov}(C_i, C_j)}{\sqrt{\operatorname{Var}(C_i)\operatorname{Var}(C_j)}} \\
&= \frac{E[(TA_i + \bar{A}_i P_i)(TA_j + \bar{A}_j P_j)] - c^2}{\operatorname{Var}(C)} \\
&= \frac{ta^2 + 2ta\bar{a}p + \bar{a}^2 p^2 - (ta + p\bar{a})^2}{(ta + p\bar{a})\overline{(ta + p\bar{a})}} \qquad \text{(since } T^2 = T\text{)} \\
&= \frac{ta^2 + 2ta\bar{a}p - t^2 a^2 - 2tap\bar{a}}{(ta + p\bar{a})\overline{(ta + p\bar{a})}} \\
&= \frac{a^2 t\bar{t}}{c\bar{c}}
\end{aligned}
$$

Rater accuracy can be obtained via

$$a^2 = \frac{c\bar{c}}{t\bar{t}}\text{Cor}(C_a, C_b) = \frac{c\bar{c}}{t\bar{t}}\kappa_{fleiss} \tag{1}$$

The correlation between two raters' ratings of the same subject is the intraclass correlation coefficient (ICC) for a two-way random effects model @shrout_intraclass_1979, which has been shown to be equivalent to the Fleiss kappa as described in @fleiss2013statistical, p. 611-12. Under the $t = p$ proficient rater assumption, $c = ta + \bar{a}p = p$, so that the Fliess kappa is (again) shown to be a^2 under that condition. The relation Equation 1 suggests that the Fliess kappa could be adjusted for cases when $t \neq p$ by making assumptions about those two parameters. For example, maybe the true rate is known from other information. The overall rate of Class 1 ratings c can be estimated directly from the data, but estimating t requires either prior knowledge of the context or using the full t-a-p estimation process, in which case there's no need to compute the Fliess kappa.

1.1 Correlation Between Ratings and True Values

It is of interest to find the correlation between T_i the truth value of subject i and the resulting classification C_i. Note that both of the random variables T_i and C_i take only values of zero or one, so squaring them doesn't change their values. This fact simplifies computations, for example $E[C_i^2] = E[C_i] = ta + p\bar{a}$. The variance of C is therefore

$$\begin{aligned}
\text{Var}(C) &= E[C^2] - E^2[C] \\
&= c - c^2 \\
&= c\bar{c} \\
&= (ta + p\bar{a})\overline{(ta + p\bar{a})}.
\end{aligned}$$

Similarly, $Var(T) = t\bar{t}$. The correlation between true values and ratings is then

$$\begin{aligned}
\text{Cor}(T, C) &= \frac{\text{Cov}(T, C)}{\sqrt{\text{Var}(T)\text{Var}(C)}} \\
&= \frac{E[T(Ta + p\bar{a})] - t(ta + p\bar{a})}{\sqrt{t\bar{t}c\bar{c}}} \\
&= \frac{t(a + p\bar{a}) - t(ta + p\bar{a})}{\sqrt{t\bar{t}c\bar{c}}} \\
&= a\frac{\sqrt{t\bar{t}}}{\sqrt{c\bar{c}}} \\
&= a\frac{\sigma_T}{\sigma_C}.
\end{aligned}$$

Where σ is the standard deviation (square root of variance). The relationship in **?@eq-cor-tc** can also be seen as $a = \text{Cor}(T, C)\frac{\sigma_C}{\sigma_T}$, which means a can be interpreted as the slope of the regression line $C = \beta_0 + \beta_1 T + \varepsilon$, i.e. $a = \beta_1$. In the proficient rater case $p = t$, $\sigma_C = \sigma_T$ and so $\mathrm{Cor}(T, C) = a$. It can also be shown that for a t-a_1, a_0-p model, the $t = p$ assumption leads to $a = \sqrt{a_1 a_0}$. See @eubankscause.

The two correlations derived here are related by $\mathrm{Cor}^2(T, C) = \mathrm{Cor}(C_i, C_j)$.

2 Alternate Derivation of Fleiss Kappa Relationship

This appendix gives an alternative derivation for the Fleiss kappa's relationship to rater accuracy under the proficient rater assumption.

The Fleiss kappa @fleiss1971measuring is a particular case of Krippendorf's alpha @krippendorff1978reliability and a multi-rater extension of Scott's pi @scott1955reliability. The statistic compares the overall distribution of ratings (ignoring subjects) to the average over within-subject distributions. These distributions are used to compute the number of observed matches (i.e. agreements) m_o over subjects $i = 1 \dots N$. For a two-category classification with a fixed number of raters $R > 1$ per subject the number of matched ratings for a given subject i is

$$
\begin{aligned}
m_o &= \frac{\binom{k_i}{2} + \binom{R-k_i}{2}}{\binom{R}{2}} \\
&= \frac{k_i(k_i - 1) + (R - k_i)(R - k_i - 1)}{R(R-1)} \\
&= \frac{2k_i^2 - 2k_i R + R^2 - R}{R(R-1)}
\end{aligned}
$$

where k_i is the count of Class 1 ratings for the ith subject. The match rates are averaged over the subjects to get $\mathrm{E}[m_o]$ and then a chance correction is applied with

$$
\kappa = \frac{\mathrm{E}[m_i] - \mathrm{E}[m_c]}{1 - \mathrm{E}[m_c]},
$$

where $\mathrm{E}[m_c]$ is the expected number of matches due to chance. Recall that different agreement statistics make different assumptions about this chance. Using the t-a-p model, and assuming $t = p$, the true rate of Class 1 t is assumed to be $\mathrm{E}[c_{ij}]$, so $\mathrm{E}[m_c] = t^2 + (1-t)^2$, the asymptotic expected match rate for independent Bernoulli trials with success probability t.

By replacing p with t in the t-a-p model's mixture distribution for the number k of Class 1 ratings a subject is assigned we obtain

$$
Pr(k) = t\binom{R}{k}(a + \bar{a}t)^k(\bar{a}\bar{t})^{R-k} + \bar{t}\binom{R}{k}(\bar{a}t)^k(1 - \bar{a}t)^{R-k}
$$

so it suffices for large N to write the expected match rate as

$$
\begin{aligned}
\mathrm{E}[m(a)] &= \sum_{k=0}^{R} \frac{2k^2 - 2kR + R^2 - R}{R(R-1)} \Pr(k; a, t) \\
&= \sum_{k=0}^{R} \frac{2k^2 - 2kR + R^2 - R}{R(R-1)} \left[t \binom{R}{k} (a + \bar{a}t)^k (\bar{a}\bar{t})^{R-k_i} + \bar{t} \binom{R}{k} (\bar{a}t)^k (1 - \bar{a}t)^{R-k} \right] \\
&= \frac{2}{R(R-1)} \sum_{k=0}^{R} k^2 \left[t \text{ Binom}(R, k, a + \bar{a}t) + \bar{t} \text{ Binom}(R, k, \bar{a}t) \right] \\
&- \frac{2R}{R(R-1)} \sum_{k=0}^{R} k \left[t \text{ Binom}(R, k, a + \bar{a}t) + \bar{t} \text{ Binom}(R, k, \bar{a}t) \right] \\
&+ \frac{R(R-1)}{R(R-1)} \sum_{k=0}^{R} \left[t \text{ Binom}(R, k, a + \bar{a}t) + \bar{t} \text{ Binom}(R, k, \bar{a}t) \right] \\
&= \frac{2}{R(R-1)} \left[tR(a + \bar{a}t)\bar{a}\bar{t} + tR^2(a + \bar{a}t)^2 + \bar{t}R(\bar{a}t)(1 - \bar{a}t) + \bar{t}R^2(\bar{a}t)^2 \right] \\
&- \frac{2}{R-1} \left[tR(a + \bar{a}t) + \bar{t}R(\bar{a}t) \right] + 1 \\
&= 2a^2(t - t^2) + 2t^2 - 2t + 1,
\end{aligned}
$$

using the moment identities to gather the sums. Here, t and R are fixed, and $m(a)$ is the average match rate over cases, which depends on unknown a and fixed $t = \mathrm{E}[c_{ij}]$. Now we can compute the Fleiss kappa with

$$
\begin{aligned}
\kappa_{fleiss} &= \frac{\mathrm{E}[m_i] - \mathrm{E}[m_*]}{1 - \mathrm{E}[m_*]} \\
&= \frac{2a^2(t - t^2) + 2t^2 - 2t + 1 - (t^2 + (1-t)^2)}{1 - (t^2 + (1-t)^2)} \\
&= a^2.
\end{aligned}
$$

So kappa is the square of accuracy under the proficient rater assumption, with constant rater accuracy and fixed number of raters. The relationship does not depend on the true distribution t of Class 1 cases.