INTRACLASS CORRELATION FORMULAS

THOMAS BRASCHLER

1. Aim of this document

This document provides an elementary literature overview and some detailed calculations for the implementation of the intraclass correlation coefficient in ICC.R of this package (function ICC).

2. Fisher's intraclass correlation coefficient

In the simplest scenario of clustering, we dispose of a sample z of a random variable Z. The sample of n realization is grouped into N groups also referred as clusters. The group sizes are n_g, where g is the group index, and where by definition $\sum_{g=1}^{g=N} n_g = n$. For convenience, we note the i-th realization in group g as z_{ig}. For this one-level hierarchy Fisher's intraclass correlation coefficient can be written as follows[1]:

$$\rho_{\text{Fisher}} = \frac{\sum_g \sum_{i \neq j} (z_{ig} - \bar{z})(z_{jg} - \bar{z})}{V(z_{ig}) \cdot \sum_g n_g (n_g - 1)} \tag{1}$$

where z_{ig} denotes the i-th value in group (cluster) g. $V(z_{ig})$ denotes the variance of the sample values regardless of the group structure:

$$V(z_{ig}) = \frac{\sum_g \sum_i (z_{ig} - \bar{z})^2}{n} \tag{2}$$

Likewise, the sample mean $\bar{z}$ is calculated regardless of the group structure as:

$$\bar{z} = \frac{\sum_g \sum_i z_{ig}}{n} \tag{3}$$

In eq. 1, the squared terms are explicitly excluded by the condition $i \neq j$. For calculation purposes, and particularly with the aggregation methods in R, it is more convenient to sum over all combinations and subtract squared terms instead. In this sense, one notes that $\sum_g \sum_{i \neq j} (z_{ig} - \bar{z})(z_{jg} - \bar{z}) = \sum_g \left[\sum_i (z_{ig} - \bar{z}) \sum_j (z_{jg} - \bar{z}) \right] - \sum_g \sum_i (z_{ig} - \bar{z})^2$. With the group means $\overline{z_g} = \frac{\sum_i z_{ig}}{n_g}$, one further notes that $\sum_i (z_{ig} - \bar{z}) = n_g (\overline{z_g} - \bar{z})$. By the definition of the overall variance in eq. 2, we also have $\sum_g \sum_i (z_{ig} - \bar{z})^2 = n \cdot V(z_{ig})$, and so eq. 1 can be equivalently rewritten as:

$$\rho_{\text{Fisher}} = \frac{1}{\sum_g n_g (n_g - 1)} \cdot \left[\frac{\sum_g n_g^2 \left(\overline{z_g} - \bar{z}\right)^2}{V(z_{ig})} - n \right] \tag{4}$$

As it is easier to evaluate eq. 4 in R than eq. 1, we used eq. 4 to facilitate the evaluation of ρ_{Fisher} (i.e. R script ICC.R, method "unbiased").

3. ANOVA-based intraclass correlation coefficients

For equal group sizes $n_g = K = \text{const}$, eq. 4 can be further simplified:

$$\rho_{\text{Harris}} = \frac{1}{(K-1)} \cdot \left[\frac{K \sum_g \left(\overline{z_g} - \bar{z}\right)^2}{N \cdot V(z_{ig})} - 1 \right] \tag{5}$$

where we have used of $n = NK$ for equal group sizes. Eq. 5 is well-known[2]. Indeed, aside from the biais correction term $\frac{1}{K-1}$, it explicitly relates intraclass variance (the $\left(\overline{z_g} - \bar{z}\right)^2$ terms) to overall residual variance $V(z_{ig})$. This is why in analysis of variance ANOVAapproaches, the intraclass correlation coefficients are typically derived from intra- and intergroup variance, as in[1]:

$$\rho_{\text{ANOVA}} = \frac{\sigma^2_{\text{inter-group}}}{\sigma^2_{\text{inter-group}} + \sigma^2_{\text{intra-group}}} \tag{6}$$

The statistical properties such as expectation values and confidence intervals of such ANOVA-based intraclass correlation coefficients have been studied in great detail[3]. For completeness, the intraclass correlation calculation method ICC in the moultonTools package (file ICC.R) allows evaluation of such ANOVA-based intraclass correlation by invoking functions basic ANOVA on a standard R linear model (lm).

4. Range of values and biais of intraclass correlation coefficients

An intraclass correlation coefficient should ideally range from 0 (for random sample values that are not correlated to group identity g) to 1 for full clustering (i.e. when the z_{ig} values are constant in each group such that $z_{ig} = \overline{z_g}$).

For the ANOVA-based approaches, according to eq. 6, full correlation leads indeed to $\rho_{\text{ANOVA}} = 1$ since with full correlation, $\sigma^2_{\text{intra-group}} = 0$. However, since the σ^2 values are never negative, $\rho_{\text{ANOVA}} \geq 0$ always holds; due to non-zero variability of the sampling and thus ρ_{ANOVA}, the expectation value is generally positive: $E(\rho_{\text{ANOVA}}) > 0$. This introduces known biais in the ANOVA approach[1], and is the reason why especially for samples with only a few clusters, unbiased formulae such as eq. 1, eq. 4 or eq. 5 or continue to be used despite their more complicated definitions[1].

It is also worthwhile to consider the value range for the unbiased estimators given by eq. 1, eq. 4 and eq. 5. Eq. 5 being a specialisation of eq. 1 or equivalently eq. 4, we shall limit

our analysis to eq. 4. As the Fisher-type formuale were designed to minimize or eliminate biais for the uncorrelated case[1], we shall here focus only on the fully correlated case.

As above, in the fully correlated case, all the z_{ig} are identical per cluster g and we have $z_{ig} = \overline{z_g}$. Therefore, the variance $V_{\text{correlated}}(z_{ig})$ becomes:

$$V(z_{ig}) = \frac{\sum_g n_g \left(\overline{z_g} - \bar{z}\right)^2}{n} \tag{7}$$

and we have:

$$\rho_{\text{Fisher, correlated}} = \frac{n}{\sum_g n_g (n_g - 1)} \cdot \left[\frac{\sum_g n_g^2 \left(\overline{z_g} - \bar{z}\right)^2}{\sum_g n_g \left(\overline{z_g} - \bar{z}\right)^2} - 1 \right] \stackrel{?}{=} 1 \tag{8}$$

Ideally, $\rho_{\text{Fisher, correlated}} = 1$ regardless of the detailed z_{ig} values. This is unfortunately not generally the case, as the example $z = \{\{0, 0\}, \{1, 1, 1\}\}$ shows. Indeed, in this case, the total number of samples is $n = 5$, the group sizes are $n_g = \{n_1, n_2\} = \{2, 3\}$, and the group means are $\overline{z_1} = 0$, $\overline{z_2} = 1$, with a global mean of $\bar{z} = \frac{3}{5}$. Numerically, eq. 8 evaluates to:

$$\rho_{\text{Fisher, correlated, example}} = \frac{5}{2 \cdot 1 + 3 \cdot 2} \cdot \left[\frac{2^2 \left(-\frac{3}{5}\right)^2 + 3^2 \left(\frac{2}{5}\right)^2}{2 \cdot \left(-\frac{3}{5}\right)^2 + 3 \cdot \left(\frac{2}{5}\right)^2} - 1 \right] = \frac{7}{8} \neq 1 \tag{9}$$

While close to 1 in the present example, this shows that the desired equality to 1 in eq. 8 does not generally hold.

Upon inspection, the cause seems to be the unequal group sizes and specifically the n_g^2 terms in eq. 4 respectively the unequal number of i, j combinations in eq. 1 for variable group size. Indeed, for equal group sizes $n_g = K = \text{const}$, eq. 8 can be simplified, achieving the exact theoretical value of 1 regardless of the actual $\overline{z_g}$

$$\rho_{\text{Fisher, correlated}} = \frac{NK}{NK(K-1)} \cdot \left[\frac{K^2 \sum_g \left(\overline{z_g} - \bar{z}\right)^2}{K \sum_g \left(\overline{z_g} - \bar{z}\right)^2} - 1 \right] = 1 \tag{10}$$

Hence, we conclude that for unequal group sizes, Fisher's formula may be unbiased, but it doesn't produce $\rho = 1$ for the fully correlated case as desired. On the opposite, the ANOVA approach produces $\rho = 1$ for the fully correlated case, but is biaised.

5. Unbiased full-range estimator

Hence, the final question is whether it is possible to provide an unbiaised estimator for the intraclass correlation coefficient that evaluates to exactly 1 for the fully clustered case, yet is unbiased for a totally uncorrelated sample.

The answer to this question is yes, but only under the assumption of normality of at least the $\overline{z_g}$ and by consequence the $\bar{z}$ values. This is a relatively reasonable assumption

due to the contribution of multiple terms to these averages, which by the central theorem of statistics tend to converge towards normality if the z_{ig} are identically distributed.

Under the assumption of normality, the sums of squares follow known χ^2 statistics, and the ratios follow F-statistics[3]. These have generally known expectation values[4] such that biais can be corrected for. Hence, by suitably adjusting the expectation value to 0 while maintaining the return value for the fully correlated case, one obtains:

$$\rho_{\text{ANOVA, unbiased}} = \frac{n-3}{n-N-2} \cdot \left[\frac{\sum_g n_g \left(\overline{z_g} - \bar{z}\right)^2}{n \cdot V\left(z_{ig}\right)} - \frac{N-1}{n-3} \right] \tag{11}$$

Eq. 11 is implemented in the ICC function in ICC.R, and can be used by providing method="

Indeed, for the totally correlated case, $\frac{\sum_g n_g (\overline{z_g} - \bar{z})^2}{n \cdot V(z_{ig})} = 1$ and so eq. 11 reduces to $\rho_{\text{ANOVA, unbiased, fully clustered}} = 1$ as it ought to be. Under the hypothesis of normally (theoretical variance: σ^2), identically distributed, independent and thus uncorrelated z_{ig}, the variance properties of the means imply that $\frac{\sum_g n_g (\overline{z_g} - \bar{z})^2}{\sigma^2}$ is a random variable with a χ^2 distribution with $N-1$ degrees of freedem. Similarly, in the denominator, $\frac{n \cdot V(z_{ig})}{\sigma^2}$ is a χ^2 variable with $n-1$ degrees of freedom. As in other F-statistics, numerator and denominator in eq. 11 are only approximately independent since they involve the same z_{ig} values grouped differently, but nevertheless, one expects the expression $\frac{n-1}{N-1} \frac{\sum_g n_g (\overline{z_g} - \bar{z})^2}{n \cdot V(z_{ig})}$ to be reasonable approximated by an F-statistic with $N-1$ degrees of freedom in the numerator and $n-1$ degrees of freedom in the denominator.

The expectation value of F-statistics is well known[4], and so:

$$\begin{aligned} E\left(\frac{n-1}{N-1} \cdot \frac{\sum_g n_g \left(\overline{z_g} - \bar{z}\right)^2}{n \cdot V\left(z_{ig}\right)} \right) &= E\left(F_{N-1,n-1}\right) = \frac{n-1}{n-3} \Rightarrow \\ E\left(\frac{\sum_g n_g \left(\overline{z_g} - \bar{z}\right)^2}{n \cdot V\left(z_{ig}\right)} \right) &= \frac{N-1}{n-3} \end{aligned} \tag{12}$$

So that ultimately:

$$E\left(\rho_{\text{ANOVA, unbiased, uncorrelated}}\right) = \frac{n-3}{n-N-2} \cdot \left[\frac{N-1}{n-3} - \frac{N-1}{n-3} \right] = 0 \tag{13}$$

as it ought to be.

6. Caution on the unbiased estimators

Two words of caution on the unbiased estimators apply. First, the F-statistics suppose normally distributed variables (at least at the level of group means), particularly for small

groups or highly non-normal individual z_{ig} values, this may not be the case. Also, just like ANOVA, it doesn't deal properly with large heteroscedasticity.

Second, and again because of the use of F-statistics for biais compensation, there are imprecisions for very small groups. Particularly, in F-statistics, numerator and denominator should be independent, which is only partially the case in practice since they are ultimately calculated from the same set of z_{ig} values. It seems advisable to proceed like in χ^2 statistics in general[5] and to apply caution if some group sizes n_g are below 5.

References

[1] J. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, Oxford: Princeton University Press, 2009.

[2] J. A. HARRIS, "ON THE CALCULATION OF INTRA-CLASS AND INTER-CLASS COEFFICIENTS OF CORRELATION FROM CLASS MOMENTS WHEN THE NUMBER OF POSSIBLE COMBINATIONS IS LARGE," *Biometrika*, vol. 9, pp. 446–472, Oct. 1913.

[3] J. L. Fleiss and P. E. Shrout, "Approximate interval estimation for a certain intraclass correlation coefficient," *Psychometrika*, vol. 43, pp. 259–262, June 1978.

[4] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. No. 55 in Applied Mathematics Series, Washington, D.C.: National Bureau of Standards, U.S. Government Printing Office, 1964, Tenth printing 1972.

[5] R. H. Riffenburgh, "Chapter 6 - Statistical Testing, Risks, and Odds in Medical Decisions," in *Statistics in Medicine (Second Edition)* (R. H. Riffenburgh, ed.), pp. 93–114, Burlington: Academic Press, Jan. 2006.