CHAPTER 3

Analysis of Categorical Ratings for Multiple Raters

Contents

3.1 Introduction

As mentioned in chapter 2, the FREQ procedure of SAS® can only compute the extent of agreement among two raters. When the number of raters is three or more, no SAS built-in procedure can compute the many extensions of Cohen's kappa that were proposed in the literature. However, one of these generalizations due to Fleiss (1971) can be calculated using a SAS macro called "`magree.sas`", and which is offered by SAS institute as a courtesy service to interested users. This SAS macro can be downloaded using the following link:

`https://support.sas.com/kb/25/addl/fusion_25006_12_magree.sas.txt`

This macro program is well documented[1], and implements Fleiss' generalized kappa, Gwet's AC_1 coefficients as well as their weighted versions using various sets of weights. This SAS macro is further discussed in section 3.2.

Note that several other agreement coefficients such as Conger's generalized kappa, Brennan-Prediger coefficient and Krippendorff's alpha and their weighted versions are not implemented in the `MAGREE.SAS` macro. To address these limitations, I developed two SAS/IML libraries of functions named `agreecoeff3raw.sas` and `agreecoeff3dist.sas`, which are discussed in section 3.3.

3.2 The MAGREE.SAS SAS Macro

The `MAGREE.SAS` SAS macro is best introduced with the analysis of a dataset of ratings. Consider a simple inter-rater reliability experiment where 4 raters named rater1, rater2, rater3 and rater4 must each classify 10 subjects into one of 5 possible categories labeled as 1, 2, 3, 4, and 5. You have 2 options for organizing the data. The first option is to record the raw data as shown in

[1] Interested users may find more information related to this macro using the following link https://support.sas.com/kb/25/006.html

Table 3.1. This format shows each rater and the specific category into which it classified each subject. Its main advantage is to capture all the information related to the experiment. The second option is to organize the data in the form of a distribution of raters by subject and category as shown in Table 3.2. This format has the advantage of giving you a quick view of the way the raters scored the subjects. However, you would not know how a specific rater categorized a given subject unless all 4 agreed by classifying the subject into the exact same category.

Table 3.1: Classification of 10 Subjects by 4 Raters in 5 Categories

Subject	rater1	rater2	rater3	rater4
1	5	5	5	5
2	3	1	3	1
3	5	5	5	5
4	3	1	3	1
5	5	5	4	5
6	1	2	3	3
7	3	1	1	1
8	1	1	1	3
9	3	3	4	4
10	1	3	1	1

My advice is to always have the raw data (Table 3.1) at your disposal whenever possible. Note that Table3.2 alone does not allow you to compute some of the agreement coefficients proposed in the literature and can always be recreated from Table 3.1. Moreover, if you only have Table 3.2 information available, you will not be able to use the `MAGREE.SAS` macro. You will shortly see why it is a problem.

Here are some of the most interesting features of the MAGREE.SAS macro:

- For nominal ratings, the MAGREE.SAS macro can compute the unweighted Fleiss' generalized kappa as well as Gwet's AC_1 coefficient. This

macro also has the special feature of computing Kappa statistics conditionally on the response category. That is, for each category, the conditional kappa is computed as a measure of the extent of agreement between raters with respect to that specific category. The methods for computing these conditional kappa coefficients are also discussed by Fleiss (1971).

- For ordinal ratings, the MAGREE.SAS macro can compute the weighted AC_1 coefficient also known as AC_2 Gwet (2021a). However, the weighted version of Fleiss' generalized kappa is not computed.
- For quantitative ratings, `MAGREE.SAS` can compute Kendall's coefficient of concordance (see Siegel and Castellan Jr., 1988) for more on this coefficient.

Table 3.2: Distribution of Raters by Subject and Category

Subject	1	2	3	4	5	Total
1	0	0	0	0	4	4
2	2	0	2	0	0	4
3	0	0	0	0	4	4
4	2	0	2	0	0	4
5	0	0	0	1	3	4
6	1	1	2	0	0	4
7	3	0	1	0	0	4
8	3	0	1	0	0	4
9	0	0	2	2	0	4
10	3	0	1	0	0	4

Program 3.1 shows how you can use the `MAGREE.SAS` macro for analyzing Table 3.1 data. The execution of this program produces the output shown in Figures 3.1, 3.2 and 3.3.

Note that lines `#01` through `#19` reads Table 3.1 data and reorganize it in the long format. That is, the input dataset `ratings` created by this data step

will have 3 columns. The first column is `s` (for subject), the second is `r` (for rater) and the third and last column is labeled `y` and contains the score that rater `r` assigned to subject `s`. For the first 2 subjects, this input datafile is structured as follows:

```
s   r   y
1   1   5
1   2   5
1   3   5
1   4   5
2   1   3
2   2   1
2   3   3
2   4   1
```

Figure 3.1 is simply the distribution of the 4 raters across the 5 categories for each of the 10 subjects. You could see that the row marginal is consistently equal to 4, the total number of raters. This is the case since there is no missing rating, each rater having rated all 10 subjects. Figure 3.2 shows Fleiss' generalized kappa and associated precision measures. While the estimated kappa coefficient of 0.3287 is valid, the same cannot be said for the standard error and the other related statistics.

> *Note that the standard error presented in Figure 3.2 is evaluated under the null hypothesis H_0 that there is no agreement among raters (i.e. inter-rater reliability is 0). As a matter of fact, this standard error can only be used for hypothesis testing and not for constructing confidence intervals.*

The correct standard error of Fleiss' generalized was derived by Gwet (2021c) and is implemented in the SAS/IML libraries `agreecoeff3raw.sas` and `agreecoeff3dist.sas` to be discussed in section 3.3. It is the standard error, which should be used for constructing confidence intervals.

Figure 3.3 shows the AC_1 coefficient along with its precision measures. The first column of this table defines 3 scenarios.

- The first scenario (third row from the table bottom) assumes that the group of raters is fixed and that only the subjects are randomly selected. The resulting analysis only applies to this specific group of raters that produced the ratings, but may be generalized beyond the sample of subjects these ratings were collected from.

- The second scenario (second row from the table bottom) assumes that the group of subjects is fixed and that only the raters are randomly selected from a larger universe of raters. The resulting analysis cannot be generalized beyond participating subjects, but can be generalized to the larger population of raters.

- In the last scenario, both raters and subjects are assumed to have been randomly selected from larger universes. In this case, you would expect the standard error to be larger as the analysis can be generalized to both universes of raters and subjects.

Program 3.1. SAS program that computes Fleiss' generalized kappa and Gwet's AC_1 using the MAGREE.SAS SAS macro (*Download this program with this link*: https://agreestat.com/books/sas2/chap3/prg3magree.sas)

```
01    title "Analysis of data from Fleiss (2003)";
02    data ratings;
03      do s=1 to 10;
04          do r=1 to 4;
05            input y @@;
06            output;
07      end;end;
08      datalines;
09    5 5 5 5
10    3 1 3 1
11    5 5 5 5
```