CHAPTER 2

Analysis of Categorical Ratings from 2 Raters

Contents

2.1 Overview

For inter-rater reliability experiments based on 2 raters, SAS offers 2 procedures with options to compute the unweighted and weighted versions of Cohen's kappa coefficient (Cohen, 1960) as well as the unweighted Gwet's AC_1 of Gwet (2008a) and PABAK of Byrt et al. (1993)[1]. These are the FREQ and SURVEYFREQ procedures. A fundamental question here is "which of these 2 procedures should you use?" Unless your dataset of ratings comes from a statistical survey based on a complex design where some subjects are randomly selected with different selection probabilities, I would advise you to always use the FREQ procedure.

The SURVEYFREQ procedure was introduced in SAS/STAT® 13.1 to meet the needs of researchers who use statistical surveys based on complex designs. For example, in a typical survey sponsored by the US government, minority groups such as Blacks, Hispanics, Asians and Native Americans are selected with a higher probability to ensure their adequate representation in the sample. To avoid a possible bias due to these differential selection probabilities, a statistical weight assigned to each subject and used in the analysis. Although FREQ and SURVEYFREQ both compute correct agreement estimates using the WEIGHT statement, only the SURVEYFREQ can compute the correct standard error after you specify survey design information. Therefore, if the subjects are selected with varying selection probabilities, then you should use SURVEYFREQ . Otherwise, use the regular FREQ procedure.

SAS implements the computation of Cohen's Kappa and weighted Kappa statistics as an option in its FREQ procedure. However, the number of raters to be analyzed must be limited to 2. The more general scenario involving 3 raters

[1]Note that Gwet's AC_1 and PABAK are implemented in the FREQ procedure since SAS/STAT® version 14.2. If you are using the free SAS OnDemand for Academics then you should not worry about this since it includes the latest versions of each procedure.

or more cannot be analyzed using the FREQ procedure. That is, the many extensions of Kappa proposed by Fleiss (1971), Light (1971), Conger (1980) or Gwet (2008a) are not implemented in SAS at the time this book is published. However, a SAS program presented in chapter 3 can be used to compute these multiple-rater agreement coefficients.

The FREQ procedure of SAS has other shortcomings when it comes to computing agreement coefficients and must be used with caution. As indicated in section 1.3 of chapter 1, there are situations, common in practice where the FREQ procedure will either fail to produce requested agreement statistics or will produce spurious results. The 3 main problems I discussed are the Diagonal Problem, the Imbalance Problem and the Ordinal Data Problem. I am going to present possible solutions to each of these problems in section 2.3.

The use of weights with agreement coefficients to account for partial agreement is discussed in section 2.4, where I present a general overview of the weighting problem in the context of inter-rater reliability. You will see that the FREQ procedure has some limited capability, as it only allows for the weighting of Cohen's kappa. To compute alternative weighted and unweighted agreement coefficients for 2 raters not implemented in the FREQ procedure such as Gwet's AC_2, Krippendorff's alpha or Scott's Pi coefficient, you will need to use the SAS/IML functions that I developed and which are discussed in section 2.5. Other important issues related to the problem of missing ratings are common in practice, and will only be addressed in section 3.5 of chapter 3.

Section 2.2 shows you how to organize input data for the FREQ procedure.

2.2 Organizing Your Data

The input file needed to compute Kappa with the FREQ procedure can be organized in 2 ways. It could be a Contingency Table or a data file of Raw

Scores. The contingency table shows the distribution of subjects by rater and by category, whereas the dataset of raw scores shows 2 columns of ratings associated with the 2 raters.

▶ Contingency Table

The contingency table in our context is the distribution of subjects by rater and by category as shown in Table 2.1.

Table 2.1: Distribution of 15 subjects by rater and category for the 3 categories a, b, and c.

Rater 1	Rater 2		
	a	b	c
a	5	1	0
b	0	3	2
c	1	1	2

This contingency table is specified in a SAS program as shown in Program 2.1 by listing all Table 2.1 cells and associated categories. This SAS program will compute Cohen's kappa and a few other statistics shown in Figures 2.1, 2.2 and 2.3. This SAS program reads the input dataset containing Table 2.1 information in lines `01` through `12`. It is in line `18` that I request the calculation of kappa, AC_1, PABAK and other kappa-related statistics. Line `#19` requests some tests of hypothesis for the unweighted and weighted kappa coefficients only.

- The first output of this program is the contingency table of Figure 2.1. The FREQ procedure always outputs this table even when agreement coefficients are not calculated due to one of the problems previously discussed. It is essential to carefully check this table for imbalance and for the uniformity of row and column labels. If the

row labels are different from column labels, it is a red flag and an indication that your results are likely false.

- The second output shown in Figure 2.2, contains all agreement coefficients implemented in the FREQ procedure, their associated standard errors and confidence intervals. The simple kappa[2] and the weighted are always both calculated once you specify the `Agree` option. A weighted kappa would be meaningless for nominal ratings (more on this in section 2.4).
- The third output will primarily be of interest if you want to test the hypothesis that the unweighted or weighted kappa equals 0. Typically, when the p-value (i.e. $Pr > Z$) is smaller than 0.05 then the associated kappa is considered to be statistically significant.

Program 2.1. Basic SAS program for computing the kappa coefficient of a contingency table (*To download this program, use the following link:* https://agreestat.com/books/sas2/chap2/prg2contingency.sas)

```
01  data rfile;
02    input rater1$ rater2$ count;
03  datalines;
04  a a 5
05  a b 1
06  a c 0
07  b a 0
08  b b 3
09  b c 2
10  c a 1
11  c b 1
12  c c 2
13  ;
14  ods pdf file="C:\kgwet\out2x1.pdf" style=Ocean;
15    proc freq data=rfile;
16      weight count;
```

[2] I would rather refer to it as the "Unweighted Kappa" as opposed to "Weighted Kappa."

```
17      tables rater1*rater2 /
18            agree(ac1 pabak kappadetails);
19      test agree;
20    run;
21  ods pdf close;
```

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of rater1 by rater2			
	rater2			
rater1	a	b	c	Total
a	5 33.33 83.33 83.33	1 6.67 16.67 20.00	0 0.00 0.00 0.00	6 40.00
b	0 0.00 0.00 0.00	3 20.00 60.00 60.00	2 13.33 40.00 50.00	5 33.33
c	1 6.67 25.00 16.67	1 6.67 25.00 20.00	2 13.33 50.00 50.00	4 26.67
Total	6 40.00	5 33.33	4 26.67	15 100.00

Figure 2.1: Distribution of 15 subjects by category