CHAPTER 1

The SAS Solution and Its Problems

Contents

1.1 Introduction

My book entitled "Handbook of Inter-Rater Reliability: *Volume 1: Chance-corrected Agreement Coefficients for Categorical Ratings*," (see Gwet, 2021a) is essentially about methodology. It discusses the fundamentals of several techniques for evaluating the extent of agreement among raters based on categorical ratings. In this book, I have decided to shift my focus, from presenting the methods to producing numbers. Because the focus is on production, I will confine myself to presenting a non-mathematical review of the techniques where appropriate, and to concentrate on one technology, which is the SAS software. SAS is a massive and expensive system, which is typically licensed to institutions such as universities, pharmaceutical companies, financial institutions or government agencies. It is therefore assumed that you already know SAS and can access it through your institution.

It is worth mentioning that an interesting learning edition of SAS known as "SAS OnDemand for Academics" is available for free to everybody. You can use the link https://welcome.oda.sas.com to set up a SAS profile and be able to use this learning edition. Setting up a SAS profile is a straightforward process. This product is very convenient since it is in the cloud and is therefore accessible with a browser from wherever you have Internet access. Although, SAS OnDemand for Academics is limited in terms of the amount of data it can processed, I still use it today to write new programs and test them on small datasets.

Most inter-rater reliability experiments involve small datasets. Therefore, many researchers in the field of inter-rater reliability should be able to process most of their datasets with SAS OnDeman for Academics. I like this software and invite you to check it out if you have not yet done so.

To respond to the growing demand from researchers for software products that can compute inter-rater reliability coefficients, particularly Cohen's Kappa (see Cohen, 1960) and Gwet's AC_1 (see Gwet, 2008a), SAS now includes options

in the FREQ procedure for calculating kappa, its weighted version proposed by Cohen (1968), Gwet's AC_1 as well as the Prevalence-Adjusted Bias-Adjusted Kappa of Byrt et al. (1993). In addition to computing these coefficients, the FREQ procedure also computes the associated precision measures, such as the standard errors, P-values, and confidence intervals. Researchers who already use SAS can now take advantage of these features. However, the solution proposed by SAS with its FREQ procedure is imperfect and may lead to potentially serious problems if not used with caution. It also has some limitations that you must be aware of. These problems and limitations are discussed in sections 1.2 and 1.3. This chapter will attempt to clarify and document many of them, and to propose solutions.

1.2 Limitations of the SAS Solution

There are 2 main limitations associated with the built-in SAS solution to the problem of inter-rater reliability.

- The first limitation is with the number of raters that must equal 2 if you are using the FREQ procedure.

- The second limitation is with the choice of agreement coefficients you want to use. Currently, the FREQ procedure[1] offers 3 agreement coefficients. These coefficients are Cohen's kappa of Cohen (1960), Gwet's AC_1 of Gwet (2008a) and the Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) of Byrt et al. (1993).

  You cannot yet compute Gwet's AC_2 with any weight, nor any weighted PABAK coefficient. PABAK is a special version of the more general Brennan-Prediger coefficient of Brennan and Prediger (1981) , and its weighted version is discussed by Gwet (2021a) . Other agreement coefficients not implemented in the FREQ procedure are Scott's Pi of Scott (1955) or Krippendorff's alpha.

[1] I am referring to SAS/STAT version 15.2.

These 2 limitations are further discussed in the next 2 sections.

1.2.1 *Number of Raters Limited to 2*

The implementation of Kappa in the FREQ procedure is almost entirely based on the content of Fleiss et al. (2003)[2], and is limited to 2 raters only. Many inter-rater reliability experiments involve 3 raters or more. Although several agreement coefficients have been proposed in the literature for multiple raters and categorical ratings, none is yet implemented in SAS. However, the SAS Institute's support group has developed the magree.sas macro program that could be downloaded using the following link:

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

This macro program is also based entirely on Fleiss et al. (2003) recommendations. However, some of these recommendations, particularly those related to the standard error of Fleiss' generalized kappa are questionable. Gwet (2021c) proposed a more accurate expression for calculating the standard error of Fleiss' generalized kappa coefficient. For categorical ratings, the `magree.sas` macro can compute Fleiss' generalized kappa as well as Gwet's AC_1/AC_2. Alternative solutions for multiple raters including other agreement coefficients such as Conger's kappa (see Conger, 1980) or Krippendorff's alpha (see Krippendorff, 2004) are discussed in chapter 3.

1.2.2 *Limited choice of coefficients*

As previously mentioned, the FREQ procedure of SAS/STAT® version 15.2 implements Cohen's kappa, Gwet's AC_1 and PABAK coefficients. The recent

[2]This book is actually a revision of Fleiss (1981) that was published by others.

implementation of both PABAK and AC_1 is unquestionably a welcome addition to the host of statistical techniques that can now be handle by the FREQ procedure. Since kappa is vulnerable to the paradoxes discussed by Cicchetti and Feinstein (1990) and Feinstein and Cicchetti (1990), having more paradox-resistant alternatives to kappa such as PABAK and AC_1 can only be appreciated. However, a few other agreement coefficients not yet implemented in SAS have been proposed and are currently being used when the number of raters is limited to 2. Two such coefficients are Krippendorff's alpha by Krippendorff (1970) and Scott's Pi proposed by Scott (1955). Krippendorff's alpha is mainly used in the field of Content analysis. The use of Scott's Pi is uncommon and its interest lies in that the more popular Fleiss' generalized kappa proposed by Fleiss (1971) reduces to it when the number of raters is 2.

In the case of multiple raters (3 or more), researchers may want to explore alternative generalized coefficients due to Conger (1980), Fleiss (1971), Brennan and Prediger (1981) , Gwet (2008a). SAS does not offer any built-in solution for these coefficients for an arbitrarily large number of raters. In chapters 3 I will present several SAS/IML function modules that implement all of these agreement coefficients, and a few more, and will show you step by step how they can be used.

1.2.3 *On the weighted kappa*

Agreement coefficients can be weighted when the categorical ratings being analyzed are of ordinal type. The concept of weighting is not used here in pure traditional sense as known in statistical science. Cohen (1968) introduced it as a way to give partial agreement credit to disagreement not deemed too serious. The FREQ procedure allows you to compute the weighted kappa coefficient. However, there is no weighted version for the other 2 agreement coefficients (i.e. PABAK and AC_1) implemented in the FREQ procedure. Consequently, if you want to compute a weighted PABAK or weighted AC_1 (i.e. AC_2), you will