CHAPTER 5

Analysis Techniques for Categorical Ratings

Contents

5.1 Introduction

This chapter addresses 2 specific and important issues related to the analysis of categorical ratings. The first issue is that of testing the difference of 2 chance-corrected agreement coefficients for statistical significance. The second issue is the benchmarking of chance-corrected agreement coefficients on existing benchmark scales proposed in the literature. Benchmarking consists of evaluating the magnitude of an agreement coefficient against standard levels so that researchers can label them as "low", "moderate" or "high." The statistical test of significance is discussed in section 5.2, whereas the benchmarking procedure is described in section 5.3.

5.2 Testing Differences for Statistical Significance

The problem addressed in this section is that of comparing 2 agreement coefficients and wanting to formally test their difference for statistical significance. The extent of agreement among raters could be measured on two occasions. On the first occasion for example, agreement would be measured before the raters receive a formal training, and be measured again after training. The difference between the before-training and the after-training agreement coefficients tells us something about the effectiveness of the training program. Therefore, this difference provides useful information and must be carefully analyzed.

What is statistical significance? In a nutshell, an observed difference is deemed statistically significant if its magnitude cannot be explained by statistical errors alone. A typical agreement coefficient is based on a random sample of subjects and will therefore carry a sampling or statistical error. Its estimated value will differ from the "true" value by a certain margin. Consequently, a sizeable difference can be observed between 2 agreement coefficients due to statistical errors, when in reality there is no difference between the "true" values. Testing a difference for statistical significance amounts to evaluating the likeli-

hood that the observed difference was caused by statistical errors alone. If this likelihood is "small", then the difference is deemed statistical significant.

5.2.1 *The Problem*

A difference between 2 agreement coefficients is deemed "statistically" significant when its magnitude exceeds the maximum value that statistical errors alone are expected to produce. Therefore, an observed difference between agreement coefficients that appears meaningful (i.e. has practical value) to you, should still be tested for statistical significance to ensure that it was not caused by "statistical noise." Statistical noise is typically quantified by the variance associated with the agreement coefficient. Only a meaningful difference that is also statistically significant can ultimately be useful.

When evaluating the difference of 2 agreement coefficients, 2 scenarios must be considered. The first scenario is one where the 2 agreement coefficients are uncorrelated. In the second scenario, the coefficients are deemed correlated. For all practical purposes, 2 agreement coefficients are deemed correlated if they are based on 2 overlapping rosters of raters, 2 overlapping groups of subjects or a mixture of both. On the other hand, 2 agreement coefficients will be uncorrelated if they are based on 2 independent groups of subjects and 2 independent rosters of raters.

Testing uncorrelated agreement coefficients for statistical significance does not pose any problem in particular and relies on standard and well-documented statistical procedures. I will not discuss this scenario in this book. However, interested readers may get more information on these procedures in Gwet (2021b, chap 9-section 9.3.2). But, you should be able to implement all of these procedures in SAS using the SAS/IML libraries discussed in the past few chapters.

Testing the difference of correlated coefficients is an entirely different problem, the solution of which is briefly reviewed in section 5.2.2. The techniques

used to address this problem were initially introduced by Gwet (2016), and further expanded in Gwet (2021b).

5.2.2 *The Solution*

The general procedure for testing the difference $\widehat{\kappa}_2 - \widehat{\kappa}_1$ between 2 chance-corrected agreement coefficients, consists of first computing the associated standardized difference T (also known as the *Pivot* in statistical jargon) defined by,

$$T = \left(\widehat{\kappa}_2 - \widehat{\kappa}_1\right) / \sqrt{V\left(\widehat{\kappa}_2 - \widehat{\kappa}_1\right)}. \tag{5.2.1}$$

If the pivot's absolute value exceeds a certain threshold then you can conclude that the difference is statistically significant. The challenge of equation 5.2.1 is the computation of the denominator. How do you evaluate the variance of the difference? For uncorrelated coefficients, the variance of the difference equals the sum of individual variances. For correlated coefficients, things may not be trivial. The approach used in this section is that of Gwet (2016), which is based on the "linearization" method . Interested readers may get all the technical details pertaining to this method in Gwet (2021b). In the next paragraph however, I will give you a general flavour of what this technique is about, before presenting the SAS program that implements it.

The typical agreement coefficient is a twisting function of individual subject ratings, making it near impossible to untangle the correlation structure of 2 correlated agreement coefficients. However, as the number of subjects increases, the agreement coefficient gets closer to the "true" value it approximates, and becomes more stable around that value. It can be shown mathematically that once an agreement coefficient reaches the vicinity of its "true" value, it can reliably be expressed as a linear function of individual subject-level values. Using this linear expression as a surrogate for the agreement coefficient to calculate the variance of the difference between two coefficients is the technique that underlies the linearization method proposed by Gwet (2016).

Program 5.1 shows how you can use the SAS/IML library of function modules `pairedttest.sas` for testing the difference of 2 agreement coefficients for statistical significance. This library is not used alone. It must be used along with the `weights.sas` library, which I already discussed in previous chapters. These 2 SAS/IML libraries can be downloaded using the following 2 links:

https://agreestat.com/books/sas2/chap5/pairedttest.sas
https://agreestat.com/books/sas2/weights.sas

The segment of Program 5.1 defined by lines `#01` through `#39` reads the 2 datasets used to produce the 2 agreement coefficients being compared. Although both datasets have the same number of raters, this is not a requirement. However, the number of subjects is expected to be same. If not, then only subjects rated by both groups of raters would be used in the paired test. Subjects rated by only one group of raters should be removed. After all, the primary goal for testing the different of 2 correlated is to evaluating the change in rater agreement for a given sample of subjects.

Lines `#43` and `#44` include the two SAS/IML libraries of functions `weights.sas` and `pairedttest.sas` into the program. You will want to change directory names from my directory `c:\kgwet\chap3` to the directory where these libraries are stored. It is in lines `#53` through `#58` that all tests of significance are performed for the 5 agreement coefficients under investigation. Here are the 5 SAS/IML functions that accomplished this task:

- **Fleiss' generalized kappa.** The testing of the difference between 2 correlated Fleiss' kappa coefficients is accomplished with function `ttest_fleiss`, which is defined as follows:

```
start ttest_fleiss(g1_ratings,g2_ratings,
        weights="unweighted",conflev=0.95,Npop=10**7);
```

- **Gwet's AC_2 or AC_1 coefficient.** The testing of the difference between 2 correlated AC_2 (or AC_1) coefficients[1] is done using function `ttest_ac2`. This function is defined as follows:

```
start ttest_ac2(g1_ratings,g2_ratings,
        weights="unweighted",conflev=0.95,Npop=10**7);
```

- **Krippendorff's alpha coefficient.** Function `ttest_alpha` is used for testing the difference between 2 correlated Krippendorff's alpha coefficients, and is defined as follows:

```
start ttest_alpha(g1_ratings,g2_ratings,
        weights="unweighted",conflev=0.95,Npop=10**7);
```

- **Conger's generalized alpha coefficient.** Function `ttest_conger` is used for testing the difference between 2 correlated Conger's generalized kappa coefficients[2], and is defined as follows:

```
start ttest_conger(g1_ratings,g2_ratings,
        weights="unweighted",conflev=0.95,Npop=10**7);
```

- **Brennan-Prediger coefficient.** Function `ttest_bp` is used for testing the difference between 2 correlated Brennan-Prediger coefficients, and is defined as follows:

```
start ttest_bp(g1_ratings,g2_ratings,
        weights="unweighted",conflev=0.95,Npop=10**7);
```

[1] Note that AC_1 can be seen as AC_2 based on the set of identity weights.

[2] Remember that Conger's generalized kappa reduces to Cohen's kappa when the number of raters is 2.