CHAPTER 12

Measures of Association and Item Analysis

OBJECTIVE

The main objective of this chapter is to review some special agreement coefficients that are used in item analysis, or used to quantify the extent of agreement among raters with respect to the rankings of subjects. I discuss Cronbach's alpha, known to measure internal consistency of items in scale development. I also discuss widely-used measures of association such as the Kendall's coefficient of concordance, the Spearman's correlation coefficient, among others. I present the assumptions underlying these techniques as well as the practical situations in which their use is recommended.

CONTENTS

12.1 Overview

In this chapter, I will review the Cronbach's alpha coefficient frequently used in item analysis. Also discussed are other measures of association aimed at evaluating the extent of agreement among raters with respect to the ranking of subjects. These special agreement coefficients do not fall into the category of chance-corrected measures such as Kappa, nor into the category of intraclass correlation coefficients. Nevertheless, they are often used by practitioners to address specific inter-rater reliability problems that cannot be resolved with the methods discussed in previous chapters.

The Cronbach's alpha, which is discussed in section 12.2, is deemed useful by researchers involved in scale development. It allows you to select an adequate set of questions for developing a scale needed to measure a particular construct of interest. In sections 12.3, 12.4, and 12.5 I present additional methods, which are recommended for evaluating the extent of agreement among raters when the subject's rank relative to other subjects is more relevant than its actual score. In many inter-rater reliability problems, the exact score assigned to subjects is not needed. What matters, is how each subject ranks with respect to other subjects. I previously gave an example of government examiners scoring proposals submitted by contractors in response to a pre-solicitation notice. Although each individual proposal is scored, it is the ranking that determines the winner. Any inter-rater reliability experiment designed to improve the extent of agreement among government examiners for example, will produce data that should be analyzed with rank-based methods.

12.2 Cronbach's Alpha

Cronbach's Alpha[1] (α) is a measure of internal consistency that is popular in the field of psychometrics. This measure was originally developed in a context where a set of questions (also called items) are asked to a group of individuals with the objective of measuring a specific construct such as risk aversion, extraversion or introversion. The extent to which all questions contribute positively towards measuring the same concept is known as internal consistency. This is a key element for evaluating the quality of the overall score. Cronbach's alpha is one of the most widely-used measure of internal consistency.

Items that are internally consistent can be seen as raters that agree about the

[1] The term alpha (α) is used very often in the study of reliability as we saw in the past few chapters. We previously investigated Krippendorff's alpha as well as Aickin's alpha. Cronbach's alpha however, proposed by Cronbach (1951) does not have much in common with the previous alpha coefficients.

12.4.3 *p-Value for Kendall's Tau*

Let z be a statistic defined as follows:

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}, \tag{12.4.3}$$

where τ is the sample-based Kendall's tau, and n the number of subjects that participated in the experiment. Under the assumption that the "true" value of τ is actually 0, the statistic z of equation 12.4.3 follows the standard Normal distribution, and can be used for computing the p-value.

The p-value, which represents the probability that the absolute value of the random variable z exceeds its observed value, can be calculated with Excel as follows:

=2⋆(1-NORM.S.DIST(z_{obs},TRUE)), *for Excel 2010/2011 or a more recent version,*
=2⋆(1-NORMSDIST(z_{obs})), *for Excel 2007/2008 or an earlier version.*

Equation 12.4.3 will generally be valid for sample sizes as small as 10, and may be used with Kendall's tau whether it is tie-adjusted or not.

Although the number of subjects in Example 12.4 is only 6, using equation 12.4.3 for illustration purposes will yield an observed z-value of,

$$z_{obs} = \frac{3 \times 0.4667\sqrt{6 \times (6-1)}}{2 \times (2 \times 6 + 5)} = 1.3152.$$

This observed z-value leads to a p-value of = 2*(1-NORM.S.DIST(1.3152,TRUE)) = 0.1885, which is too high for Kendall's tau to be considered statistically significant[6]. Kendall's tau, Spearman's and Pearson's correlations are all bivariate measures of association that I discussed in the past few sections. In the next section, I will expand the study of measures of association to include the analysis of three raters or more. The two (very similar) measures to be described are the Kendall's Coefficient of Concordance (KCC), and the Friedman's Chi-Square Statistic (KCS).

12.5 Kendall's Coefficient of Concordance (KCC)

Kendall's coefficient of concordance (KCC) quantifies the extent of agreement among three raters or more with respect to their ranking of the same group of

[6] Note that tau would be considered statistically significant in this context, if the p-value is below 0.05. This would mean that the "true" value of tau with no sampling errors can be considered to exceed 0 in absolute 0.

subjects. It is often denoted by W and varies from 0 to 1, 0 representing total absence of agreement and 1 representing complete agreement. The fact that W never takes a negative value is not a problem, as long as the KCC is used to measure agreement among three raters or more. The notion of negative association or negative correlation (i.e. ratings changing in opposite directions) does not apply to a group of three raters or more, even though it is often relevant in the case of 2 raters.

Throughout this section I will assume that the ratings to be analyzed are organized columnwise as shown in Table 12.10. This table shows numeric scores that 4 judges assigned to 6 subjects.

Table 12.10: Ratings of Six Subjects by Judge1, Judge2, Judge3, and Judge4

Subject	Judge1	Judge2	Judge3	Judge4
A	9	2	5	8
B	6	1	3	2
C	8	4	6	8
D	7	1	2	6
E	10	5	6	9
F	6	2	4	7

To formally define Kendall's coefficient of concordance, I assume that n subjects were rated by k judges (for Table 12.10, $n = 6$, and $k = 4$). Since KCC is a measure that is based on ranks, each column of Table 12.10 (except the first column of subject labels) must first be ranked in ascending order from 1 to the number of subjects. Each rating in a particular tie series will receive the same average rank. For example the "Judge1" column of Table 12.10 contains the tie series $\{6, 6\}$, and these numbers are the 2 lowest of all ratings by Judge1. Their respective ranks would normally be 1, and 2, leading to an average rating of $(1 + 2)/2 = 1.5$. For the purpose of calculating the KCC, both numbers will receive the same ranking of 1.5. Let R_{ij} be the abstract rank associated with subject i and rater j.

The KCC represents the ratio of the variance associated with the subject marginal sums of ranks R_i to the maximum possible variance given the number of subjects and the number of judges. More formally, the KCC denoted by W is calculated using one of the following two equations[7]:

$$\boxed{W = \frac{12\mathrm{S}}{k^2n(n^2-1) - kT},} \tag{12.5.1}$$

[7] Both ways of calculating Kendall's coefficient of concordance are found in textbooks, and lead to the same result. However, some statistical tables often used to evaluate the statistical significance of KCC are based on the value of S, which cannot be obtained from equation 12.5.2

$$\boxed{W = \frac{12\mathrm{S}^\star - 3k^2n(n+1)^2}{k^2n(n^2-1) - kT},} \tag{12.5.2}$$

where S is the sum over all subjects, of the squared differences from the marginal sums of ranks R_i $(i = 1, \cdots, n)$ to their overall mean $\overline{R}$ (i.e. $(R_i - \overline{R})^2$), and $\mathrm{S}^\star$ the sum over all subjects, of the squared R_i (i.e. R_i^2). T is the tie-correction factor that is defined as follows:

$$T = \sum_{l=1}^{m}(t_l^3 - t_l), \tag{12.5.3}$$

where m is the number of tie series in the data table (e.g. there are 5 tie series having 2 subjects each, in Table 12.10), and t_l is the number of subjects associated with a specific tie series l (e.g. for each tie series l of Table 12.10, $t_l = 2$). The following example illustrates the calculation of Kendall's coefficient of concordance using Table 12.10 data.

Example 12.6

Table 12.11 shows 6 subjects and their rankings by 4 judges based on Table 12.10 ratings. The ratings are ranked independently for each judge. The "R_i" column contains the subject marginal sums of ranks (or row sums), while the "R_i^2" column contains the associated squared values. The "S" row contains the column sum of squared differences $(R_i - \overline{R})^2$, while T shows the result of equation 12.5.3 for each column and for all columns combined.

Table 12.11: Ratings of Six Subjects by Judge1, Judge2, Judge3, and Judge4

Subject	Judge1	Judge2	Judge3	Judge4	R_i	R_i^2
1	5	3.5	4	4.5	17	289
2	1.5	1.5	2	1	6	36
3	4	5	5.5	4.5	19	361
4	3	1.5	1	2	7.5	56.25
5	6	6	5.5	6	23.5	552.25
6	1.5	3.5	3	3	11	121
Total	21	21	21	21	84	1415.5
S	17	16.5	17	17	239.5	
T	6	12	6	6	30	

Using equation 12.5.1, you can compute Kendall's coefficient of concordance as follows:

$$W = \frac{12 \times 239.5}{4^2 \times 6 \times (6^2 - 1) - 4 \times 30} = 0.887.$$

The same result would be obtained using equation 12.12 as follows,

$$W = \frac{12 \times 1,415.5 - 3 \times 4^2 \times 6 \times (6+1)^2}{4^2 \times 6 \times (6^2 - 1) - 4 \times 30} = 0.887.$$

In this example, Kendall's coefficient of concordance is high and close to 1, which suggests high inter-judge reliability among the 4 judges.

Kendall's coefficient of concordance was developed independently by Kendall and Babington-Smith (1939) and Wallis (1939), and has a close relationship with the Spearman's correlation, which is formulated by the following equation:

$$W = \overline{r} - (\overline{r} - 1)/k, \tag{12.5.4}$$

where $\overline{r}$ is the average of all distinct pairwise Spearman's correlation coefficients. Equation 12.5.4 shows that as the number of raters grows, Kendall's coefficient of concordance tends to get closer to the average of the pairwise Spearman correlations.

p-Value of Kendall's Coefficient of Concordance

To compute the p-value associated with Kendall's coefficient of concordance, one must first compute the following statistic:

$$\chi^2 = k(n-1)W, \tag{12.5.5}$$

where W is the KCC computed with equation 12.5.1 or equation 12.5.2. The statistic χ^2 follows the chi-square distribution with $n-1$ degrees of freedom. The p-value represents the probability that the random variable χ^2 exceeds its observed value (usually denoted by χ^2_{obs}). It can be computed with MS Excel as follows:

=CHISQ.DIST.RT($\chi^2_{obs}, n-1$), *for EXCEL 2010/2011 or a more recent version,*
=CHIDIST($\chi^2_{obs}, n-1$), *for EXCEL 2007/2008 or an earlier version.*

Using the numbers from Example 12.6, the observed statistic χ^2_{obs} is calculated as follows:

$$\chi^2_{obs} = k(n-1)W_{obs} = 4 \times (6-1) \times 0.887 = 17.74,$$

which leads to a p-value of =CHISQ.DIST.RT(17.74,6-1) = 0.00329. Since this value is below 0.05, one may conclude that the KCC is statistically significant. That is, the "true" error-free coefficient it approximates is considered positive based on the observed ratings. The magnitude of the calculated inter-judge agreement cannot be due to sampling errors alone.

If the number of raters or the number of subjects appears to be too small for the chi-square distribution to provide an adequate approximation to the the sampling distribution of the statistic χ^2, then one may consider some of the methods suggested by Marascuilo and McSweeney (1977) or Siegel and Castellan (1988) who evaluate the significance of W with an adjusted chi-square value.

Relationship with Friedman's Chi-Square Statistic

The Friedman's Chi-Square statistic (denoted by χ_r^2) proposed by Friedman (1937) is sometimes mentioned in studies of inter-judge reliability. But this is essentially due to its close mathematical relationship with Kendall's coefficient (W). Friedman's statistic was developed primarily to test the hypothesis that the ratings assigned to subjects under investigation come from the same statistical population. This is an indirect way of evaluating the extent of agreement among raters.

Kendall's coefficient of concordance may be derived from Friedman chi-square statistic using the following equation:

$$W = \frac{\chi_r^2}{n(k-1)}. \tag{12.5.6}$$

Note that the above equation is valid only if χ_r^2 is calculated first, then used to compute W. If W is calculated first, then it could be used to obtain Friedman's chi-square statistic as,

$$\chi_r^2 = k(n-1)W. \tag{12.5.7}$$

This is due to the fact that in Friedman and Kendall models, k and n play reverse roles.

12.6 Concluding Remarks

In this chapter, I discussed several alternative agreement coefficients that may be more appropriate in special situations where traditional chance-corrected agreement measures or intraclass correlation coefficients are not recommended. These agreement coefficients were discussed along with methods for calculating associated p-values.

The first agreement coefficient presented was Cronbach's alpha, used primarily in the field of psychometrics. Its main objective is to quantify the extent to which a group of items (e.g. items could be questions in a test questionnaire) contribute positively towards the measurement of the same construct such as the proficiency of a nursing student in a particular aspect of patient care. There is no need to worry about agreement by pure chance in this context. Therefore, chance-corrected agreement coefficients will not resolve this problem. Even if the scores are continuous the intraclass correlation coefficient will not be useful since none of the statistical models discussed in the Part III chapters of this book apply. Consequently, the researcher is dealing with a special situation here that requires a special method.

I also presented the well-known Pearson's correlation coefficient. It was not designed to measure agreement among raters at all. Instead, it measures the extent to which the relationship between two series of scores is linear. Although agreement implies a linear relationship, a linear relationship does not imply agreement. The main benefit for using the Pearson's correlation coefficient in the context of agreement analysis lies in its simplicity, and the easiness with which it can be obtained. If this correlation coefficient is low, you know immediately that there is no agreement among raters. On the other hand, if it is high, then you may need to proceed with the use of more appropriate agreement coefficients.

In order to address the issue of agreement among raters with respect to the ranking of subjects, and not with respect to the exact scores, I discussed two coefficients: (*i*) Spearman's correlation coefficient, and (*ii*) Kendall's Tau. Both coefficients are designed to quantify the extent to which two raters agree with respect to the ranking of subjects. Although Spearman's correlation is older than Kendall's Tau, the latter coefficient is preferred by some researchers due to some interesting statistical properties it possesses. To quantify the extent of agreement among three raters or more with respect to the ranking of subjects, I also discussed Kendall's coefficient of concordance, and the related Friedman's Chi-square statistic. For the purpose of evaluating agreement with respect to rankings, the use of these coefficients is justified. It is because a disagreement over the scoring of two subjects, which results in the same rankings is considered an agreement. Spearman's correlation, Kendall's Tau, and Kendall's coefficient of concordance will work best with continuous scores, with which chance agreement is less problematic. If raters have a limited number of scores to choose from, the likelihood of chance agreement will increase, leading to an unduly high number of ties. A large number of ties in the rankings must be avoided, as it is expected to reduce the reliability of these ranking-based coefficients.