East Carolina University
Department of Psychology

Inter-Rater Agreement©

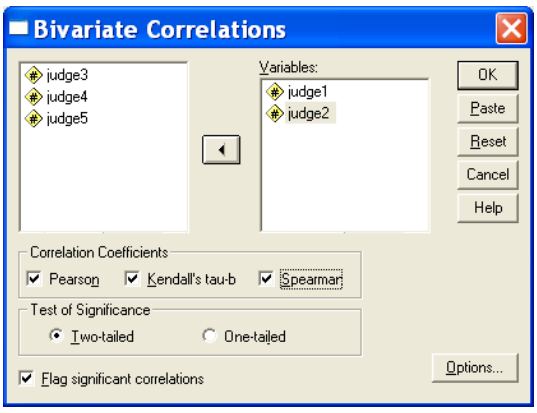
Psychologists commonly measure various characteristics by having a rater assign scores to observed people, other animals, other objects, or events. When using such a measurement technique, it is desirable to measure the extent to which two or more raters agree when rating the same set of things. This can be treated as a sort of reliability statistic for the measurement procedure.

Continuous Ratings, Two Judges

Let us first consider a circumstance where we are comfortable with treating the ratings as a continuous variable. For example, suppose that we have two judges rating the aggressiveness of each of a group of children on a playground. If the judges agree with one another, then there should be a high correlation between the ratings given by the one judge and those given by the other. Accordingly, one thing we can do to assess inter-rater agreement is to correlate the two judges' ratings. Consider the following ratings (they also happen to be ranks) of ten subjects:

Subject	1	2	3	4	5	6	7	8	9	10
Judge 1	10	9	8	7	6	5	4	3	2	1
Judge 2	9	10	8	7	5	6	4	3	1	2

These data are available in the SPSS data set **IRA-1.sav** at my SPSS Data Page. I used SPSS to compute the correlation coefficients, but SAS can do the same analyses. Here is the dialog window from Analyze, Correlate, Bivariate:



The **Pearson correlation** is impressive, $r = .964$. If our scores are ranks or we can justify converting them to ranks, we can compute the Spearman correlation coefficient or Kendall's tau. For these data **Spearman rho** is .964 and **Kendall's tau** is .867.

We must, however, consider the fact that two **judges scores could be highly correlated with one another but show little agreement**. Consider the following data:

Subject	1	2	3	4	5	6	7	8	9	10
Judge 4	10	9	8	7	6	5	4	3	2	1
Judge 5	90	100	80	70	50	60	40	30	10	20

The correlations between judges 4 and 5 are identical to those between 1 and 2, but judges 4 and 5 obviously do not agree with one another well. Judges 4 and 5 agree on the ordering of the children with respect to their aggressiveness, but not on the overall amount of aggressiveness shown by the children.

One solution to this problem is to compute the intraclass correlation coefficient. Please read my handout, The **Intraclass Correlation Coefficient**. For the data above, the intraclass correlation coefficient between Judges 1 and 2 is .9672 while that between Judges 4 and 5 is .0535.

What if we have **more than two judges**, as below? We could compute Pearson *r*, Spearman rho, or Kendall tau for each pair of judges and then average those coefficients, but we still would have the problem of high coefficients when the judges agree on ordering but not on magnitude. We can, however, compute the intraclass correlation coefficient when there are more than two judges. For the data from three judges below, the intraclass correlation coefficient is .8821.

Subject	1	2	3	4	5	6	7	8	9	10
Judge 1	10	9	8	7	6	5	4	3	2	1
Judge 2	9	10	8	7	5	6	4	3	1	2
Judge 3	8	7	10	9	6	3	4	5	2	1

The intraclass correlation coefficient is **an index of the reliability of the ratings for a typical, single judge**. We employ it when we are going to collect most of our data using only one judge at a time, but we have used two or (preferably) more judges on a subset of the data for purposes of estimating inter-rater reliability. SPSS calls this statistic the **single measure intraclass correlation**.

If what we want is **the reliability for all the judges averaged together**, we need to apply the Spearman-Brown correction. The resulting statistic is called the **average measure intraclass correlation** in SPSS and the **inter-rater reliability coefficient** by some others (see MacLennon, R. N., Interrater reliability with SPSS for Windows 5.0, *The American Statistician*, 1993, *47*, 292-296). For our data,

$\frac{j * icc}{1+(j-1)icc} = \frac{3(.8821)}{1+2(.8821)} = .9573$, where *j* is the number of judges and *icc* is the intraclass correlation coefficient. I would think this statistic appropriate when the data for our main study involves having *j* judges rate each subject.

Rank Data, More Than Two Judges

When our data are rankings, we don't have to worry about differences in magnitude. In that case, we can simply employ Spearman rho or Kendall tau if there are only two judges or Kendall's coefficient of concordance if there are three or more judges. Consult pages 320 - 321 in David Howell's *Statistics for Psychology*, 8th edition, for an explanation of Kendall's coefficient of concordance. Run the program **Kendall-Patches.sas**, from my **SAS programs page**, as an example of using SAS to compute Kendall's coefficient of concordance. The data are those from

Howell, page 310. Statistic 2 is the Friedman chi-square testing the null hypothesis that the patches differ significantly from one another with respect to how well they are liked. This null hypothesis is equivalent to the hypothesis that there is no agreement among the judges with respect to how pleasant the patches are. To convert the Friedman chi-square to Kendall's coefficient of concordance, we simply substitute into this equation:

$W = \frac{\chi^2}{J(n-1)} = \frac{33.889}{6(7)} = .807$, where *J* is the number of judges and *n* is the number of things being ranked.

If the judges gave ratings rather than ranks, you must first convert the ratings into ranks in order to compute the Kendall coefficient of concordance. An explanation of how to this with SAS is presented in my document "Nonparametric Statistics." You would, of course, need to remember that ratings could be concordant in order but not in magnitude.

Categorical Judgments

Please re-read pages 166 and 167 in David Howell's *Statistical Methods for Psychology*, 8th edition. Run the program **Kappa.sas**, from my **SAS programs page**, as an example of using SAS to compute kappa. It includes the data from page 166 of Howell. Note that Cohen's kappa is appropriate only when you have two judges. If you have more than two judges you may use Fleiss' kappa.

Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial

Return to Wuensch's Statistics Lessons Page