StatNews #31

Assessing Agreement Between Raters
October 1998

In many fields researchers would like to assess agreement between raters. An example would be when two people have to separately rate, on a three-point scale, fruits according to the presence or absence of fungus. These data could be presented in a 3 x 3 table as follows.

	Rater A		
Rater B	None	Some	Extensive
None	3	5	6
Some	9	4	5
Extensive	2	3	8

Similar data are often found in food tasting studies.

As you can see from the table above, these raters exactly agree on 3+4+8=15 out of 45 fruits, and completely disagree on 2+6=8 fruits. We might want to know whether the extent of agreement is stronger than what would be expected by chance.

The most common way of assessing agreement is to use the kappa statistic. Cohen's kappa will measure the excess of agreement between raters over the level of agreement that would have been obtained by chance alone. The kappa coefficient will equal 1 if there is perfect agreement, whereas 0 is what would be expected by chance alone. Along with the kappa measure of agreement we can also test the hypothesis that the kappa coefficient equals 0. Several statistical packages will allow you to obtain kappa along with the test: SPSS (in Crosstabs), Stata and SAS (Proc Freq) to name the most commonly used here on campus. Be aware though that the test will not be accurate if your crosstabulation is not square. This can happen when one rater never used one rating category. SPSS has a macro available at their web site to address this specific problem.

Besides complete agreement or complete disagreement, one might also be interested in the level of disagreement. The seriousness of the disagreement might in some cases depend on the difference between the ratings. For example, if Rater A says that a particular fruit has no fungus and Rater B says that the same fruit has extensive fungus, this is a stronger disagreement than if Rater B said the fruit had some fungus. A similar situation can occur even in a case where the ratings are not ordered. When it becomes necessary to take the closeness of agreement into consideration one could do this by using weights. The weighted kappa, also called generalized kappa, can be calculated using SAS (proc Freq) or STATA.

A multirater kappa is available for when there are more than two raters rating a group of objects/ subjects. You can implement it very easily with STATA. Macros to implement this in SPSS also exist and can be found at the company's web site http://www.spss.com/corpinfo/?source=homepage&hpzone=footer.

A good general reference for comparison of raters is: Fleiss JL. Statistical Methods for Rates and Proportions. New York: John Wiley and Sons. 1981

If you have data which compare raters or other similar data, and would like help with the analysis, please contact the Office of Statistical Consulting.

Author: Francoise Vermeylen

Back to StatNews Table of Contents