Measure of Association

G-H Huang, National Chiao Tung University, Hsinchu, Taiwan

Glossary

Agreement – The measure of agreement intends to quantify the reproducibility of the same variable measured more than once.

Cochran–Mantel–Haenszel procedure – When the association between two discrete variables is affected by a third factor, one should look at the odds ratio of the two variables in separate strata by the third factor. The Cochran–Mantel–Haenszel procedure is used to combine the odds ratios for separate strata into an overall summary estimate.

Continuous variables – A variable is called a continuous variable if it can take on any values within a given interval. There are no gaps in the possible values of the variable.

Discrete ordinal/nominal variables – A discrete variable can take values form a discrete set of numbers. When the numbers reflect only the relative order but not specific numerical values, the discrete variable is called a discrete ordinal variable. If the numbers are used merely to identify categories and the categories have no specific ordering, the discrete variable is called a discrete nominal variable.

Kappa stastistic – The Kappa statistic is used to measure the degree of nonrandom agreement beween two discrete variables. Here, the nonrandom agreement means that two variables agree more often than expected by chance.

Kendall's tau – The Kendall's tau statistic is a nonparametric measure of association based on the number of concordances and discordances in paired observations. The statistic can be applied to the cases where the variables are continuous and/or discrete ordinal.

Odds ratio – The odds ratio is a basic measure of association in a 2x2 table formed from nominal variables. The odds ration is defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group.

Partial correlation – The partial correlation is a measure that assesses the degree of association between two variables after controlling for other variables. It can be calculated as the Pearson correlation coefficient for the residuals from linear regressions of interested variables on controlled variables.

Rank correlation – The rank correlation coefficient uses the ranks of the data, instead of the actual observed values, to compute a correlation coefficient. It is very useful when there exist extreme values in one or both variables, where the Pearson correlation coefficient will be greatly affected.

Researchers often wish to measure the strength of relationship or association between two variables. A high degree of association indicates that changes in one variable tend to be accompanied by changes in the other, and a low level of association would indicate two variables to be almost independent of each other. There are many indices that characterize the association between two variables. This article divides the indices according to the type of variable we are measuring – namely, whether variables are continuous, discrete ordinal, or discrete nominal. In the following discussion, we assume that a set of n bivariate observations $(x_1, y_1), \ldots, (x_n, y_n)$ is measured, and x_i's and y_i's can be one of the three variable types. We are interested in assessing the association between X and Y.

Association for Continuous Variables

Consider two continuous variables; for example, the score of a math test versus the score of a science test. The Pearson product–moment correlation coefficient (Pearson, 1896; Fisher, 1915) can be used to measure the degree to which two variables are linearly related, and can be calculated as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x_i's and y_i's, respectively. The value of r lies between -1 and $+1$. If $r > 0$, then the variables are positively correlated – as x increases (decreases), so does y. If $r < 0$, then the variables are negatively correlated – as x increases (decreases), y tends to decrease (increase). The values of r close to zero mean that there is no linear relationship between the variables; possible reasons can be that (1) the two variables are independent (e.g., knowledge of the math score in no way improves the prediction of the science score), or (2) the two variables have a nonlinear