

measure of association

measure of association, in statistics, any of various factors or coefficients used to quantify a relationship between two or more variables. Measures of association are used in various fields of research but are especially common in the areas of epidemiology and psychology, where they frequently are used to quantify relationships between exposures and diseases or behaviours.

A measure of association may be determined by any of several different analyses, including correlation analysis and regression analysis. (Although the terms *correlation* and *association* are often used interchangeably, *correlation* in a stricter sense refers to linear correlation, and *association* refers to any relationship between variables.) The method used to determine the strength of an association depends on the characteristics of the data for each variable. Data may be measured on an interval/ratio scale, an ordinal/rank scale, or a nominal/categorical scale. These three characteristics can be thought of as continuous, integer, and qualitative categories, respectively.

Methods of analysis

Pearson's correlation coefficient

A typical example for quantifying the association between two variables measured on an interval/ratio scale is the analysis of relationship between a person's height and weight. Each of these two characteristic variables is measured on a continuous scale. The appropriate measure of association for this situation is Pearson's correlation coefficient, r (rho), which measures the strength of the linear relationship between two variables on a continuous scale. The coefficient r takes on the values of -1 through $+1$. Values of -1 or $+1$ indicate a perfect linear relationship between the two variables, whereas a value of 0 indicates no linear relationship. (Negative values simply indicate the direction of the association, whereby as one variable increases, the other decreases.) Correlation coefficients that differ from 0 but are not -1 or $+1$ indicate a linear relationship, although not a perfect linear relationship. In practice, ρ (the population correlation coefficient) is estimated by r , which is the correlation coefficient derived from sample data.

Although Pearson's correlation coefficient is a measure of the strength of an association (specifically the linear relationship), it is not a measure of the significance of the association. The significance of an association is a separate analysis of the sample correlation coefficient, r , using a t -test to measure the difference between the observed r and the expected r under the null hypothesis.

Spearman rank-order correlation coefficient

The Spearman rank-order correlation coefficient (Spearman rho) is designed to measure the strength of a monotonic (in a constant direction) association between two variables measured on an ordinal or ranked scale. Data that result from ranking and data collected on a scale that is not truly interval in nature (e.g., data obtained from Likert-scale administration) are subject to Spearman correlation analysis. In addition, any interval data may be transformed to ranks and analyzed with the Spearman rho, although this results in a loss of information. Nonetheless, this approach may be used, for example, if one variable of interest is measured on an interval scale and the other is measured on an ordinal scale. Similar to Pearson's correlation coefficient, Spearman rho may be tested for its significance. A similar measure of strength of association is the Kendall tau, which also may be applied to measure the strength of a monotonic association between two variables measured on an ordinal or rank scale.

As an example of when Spearman rho would be appropriate, consider the case where there are seven substantial health threats to a community. Health officials wish to determine a hierarchy of threats in order to most efficiently deploy their resources. They ask two credible epidemiologists to rank the seven threats from 1 to 7, where 1 is the most significant threat. The Spearman rho or Kendall tau may be calculated to measure the degree of association between the epidemiologists' rankings, thereby indicating the collective strength of a potential action plan. If there is a significant association between the two sets of ranks, health officials may feel more confident in their strategy than if a significant association is not evident.

Chi-square test

The chi-square test for association (contingency) is a standard measure for association between two categorical variables. The chi-square test, unlike Pearson's correlation coefficient or Spearman rho, is a measure of the significance of the association

rather than a measure of the strength of the association.

A simple and generic example follows. If scientists were studying the relationship between gender and political party, then they could count people from a random sample belonging to the various combinations: female-Democrat, female-Republican, male-Democrat, and male-Republican. The scientists could then perform a chi-square test to determine whether there was a significant disproportionate membership among those groups, indicating an association between gender and political party.

Relative risk and odds ratio

Specifically in epidemiology, several other measures of association between categorical variables are used, including relative risk and odds ratio. Relative risk is appropriately applied to categorical data derived from an epidemiologic cohort study. It measures the strength of an association by considering the incidence of an event in an identifiable group (numerator) and comparing that with the incidence in a baseline group (denominator). A relative risk of 1 indicates no association, whereas a relative risk other than 1 indicates an association.

As an example, suppose that 10 out of 1,000 people exposed to a factor X developed liver cancer, while only 2 out of 1,000 people who were never exposed to X developed liver cancer. In this case, the relative risk would be $(10/1000)/(2/1000) = 5$. Thus, the strength of the association is 5, or, interpreted another way, people exposed to X are five times more likely to develop liver cancer than people not exposed to X. If the relative risk was less than 1 (perhaps 0.2, for example), then the strength of the association would be equally evident but with another explanation: exposure to X reduces the likelihood of liver cancer five-fold, indicating that X has a protective effect. The categorical variables are exposure to X (yes or no) and the outcome of liver cancer (yes or no). This calculation of the relative risk, however, does not test for statistical significance. Questions of significance may be answered by calculation of a 95% confidence interval. If the confidence interval does not include 1, the relationship is considered significant.

Similarly, an odds ratio is an appropriate measure of strength of association for categorical data derived from a case-control study. The odds ratio is often interpreted the same way that relative risk is interpreted when measuring the strength of the

association, although this is somewhat controversial when the risk factor being studied is common.

Additional methods

There are a number of other measures of association for a variety of circumstances. For example, if one variable is measured on an interval/ratio scale and the second variable is dichotomous (has two outcomes), then the point-biserial correlation coefficient is appropriate. Other combinations of data types (or transformed data types) may require the use of more specialized methods to measure the association in strength and significance.

Other types of association describe the way data are related but are usually not investigated for their own interest. Serial correlation (also known as autocorrelation), for instance, describes how in a series of events occurring over a period of time, events that occur closely spaced in time tend to be more similar than those more widely spaced. The Durbin-Watson test is a procedure to test the significance of such correlations. If the correlations are evident, then it may be concluded that the data violate the assumptions of independence, rendering many modeling procedures invalid. A classical example of this problem occurs when data are collected over time for one particular characteristic. For example, if an epidemiologist wanted to develop a simple linear regression for the number of infections by month, there would undoubtedly be serial correlation: each month's observation would depend on the prior month's observation. This serial effect (serial correlation) would violate the assumption of independent observations for simple linear regression and accordingly render the parameter estimates for simple linear regression as not credible.

Inferring causality

Perhaps the greatest danger with all measures of association is the temptation to infer causality. Whenever one variable causes changes in another variable, an association will exist. But whenever an association exists, it does not always follow that causation exists. In epidemiology, the ability to infer causation from an association is often weak because many studies are observational and subject to various alternative explanations for their results. Even when randomization has been applied, as in clinical trials, inference of causation is often limited.

