Selecting Statistics Glossary

Glossary

ADDITIVE. A situation in which the best estimate of a dependent variable is obtained by simply adding together the appropriately computed effects of each of the independent variables. Additivity implies the absence of interactions. See also INTERACTION.

AGREEMENT. Agreement measures the extent to which two sets of scores (e.g., scores obtained from two raters) are identical. Agreement involves a more stringent matching of two variables than does covariation, which implicitly allows one to change the mean (by adding a constant) and/or to change the variance (by multiplying by a constant) for either or both variables before checking the match.

BIAS. The difference between the expected value of a statistic and the population value it is intended to estimate. See EXPECTED VALUE. BIASED ESTIMATOR. A statistic whose expected value is not equal to the population value. See EXPECTED VALUE.

BIVARIATE NORMALITY. A particular form of distribution of two variables that has the traditional “”bell”” shape (but not all bell-shaped distributions are normal). If plotted in three- dimensional space, with the vertical axis showing the number of cases, the shape would be that of a three-dimensional bell (if the variances on both variables were equal) or a “”fireman’s hat”” (if the variances were unequal). When perfect bivariate normality obtains, the distribution of one variable is normal for each and every value of the other variable. See also NORMAL DISTRIBUTION.

BRACKETING. The operation of combining categories or ranges of values of a variable so as to produce a small number of categories. Sometimes referred to as ‘collapsing’ or ‘grouping.’

CAPITALIZATION ON CHANCE. When one is searching for a maximally powerful prediction equation, chance fluctuations in a given sample act to increase the predictive power obtained; since data from another sample from the same population will show different chance fluctuations, the equation derived for one sample is likely to work less well in any other sample.

CAUSAL MODEL. An abstract quantitative representation of real-world dynamics (i.e., of the causal dependencies and other interrelationships among observed or hypothetical variables).

COMPLEX SAMPLE DESIGN. Any sample design that uses something other than simple random selection. Complex sample designs include multi-stage selection, and/or stratification, and/or clustering.

COVARIATE. A variable that is used in an analysis to correct, adjust, or modify the scores on a dependent variable before those scores are related to one or more independent variables. For example, in an analysis of how demographic factors (age, sex, education, etc.) relate to wage rates, monthly earnings might first be adjusted to take account of (i.e., remove effects attributable to) number of hours worked, which in this example would be the covariate. COVARIATION. Covariation measures the extent to which cases (e.g., persons) have the same relative positions on two variables. See also AGREEMENT. DEPENDENT VARIABLE. A variable which the analyst is trying to explain in terms of one or more independent variables. The distinction between dependent and independent variables is typically made on theoretical grounds-in terms of a particular causal model or to test a particular hypothesis. Synonym: criterion variable.

DESIGN MATRIX. A specification, expressed in matrix format, of the particular effects and combinations of effects that are to be considered in an analysis. DICHOTOMOUS VARIABLE. A variable that has only two categories. Gender (male/female) is an example. See also TWO-POINT SCALE. DUMMY VARIABLE. A variable with just two categories that reflects only part of the information actually available in a more comprehensive variable. For example, the four-category variable Region (Northeast Southeast, Central, West) could be the basis for a two-category dummy variable that

would distinguish Northeast from all other regions. Dummy variables often come in sets so as to reflect all of the original information. In our example, the four-category region variable defines four dummy variables (1) Northeast vs. all other; (2) Southeast vs. all other; (3) Central vs. all other; and, (4) West vs. all other. Alternative coding procedures (which are equivalent in terms of explanatory power but which may produce more easily interpretable estimates) are effect coding and orthogonal polynomials.

EXPECTED VALUE. A theoretical average value of a statistic over an infinite number of samples from the same population.

HETEROSCEDASTICITY. The absence of homogeneity of variance. See HOMOGENEITY OF VARIANCE.

HIERARCHICAL ANALYSIS. In the context of multidimensional contingency table analysis, a hierarchical analysis is one in which inclusion of a higher order interaction term implies the inclusion of all lower order terms. For example, if the interaction of two independent variables is included in an explanatory model, then the main effects for both of those variables are also included in the model.

HOMOGENEITY OF VARIANCE. A situation in which the variance on a dependent variable is the same (homogeneous) across all levels of the independent variables. In analysis of variance applications, several statistics are available for testing the homogeneity assumption (see Kirk, 1968, page 61); in regression applications, a lack of homogeneity can be detected by examination of residuals (see Draper and Smith, 1966, page 86). In either case, a variance-stabilizing transformation may be helpful (see Kruskal, 1978, page 1052). Synonym: homoscedasticity. Antonym: heteroscedasticity. HOMOSCEDASTICITY. See HOMOGENEITY OF VARIANCE.

INDEPENDENT VARIABLE. A variable used to explain a dependent variable. Synonyms:

predictor variable, explanatory variable. See also DEPENDENT VARIABLE. INTERACTION. A situation in which the direction and/or magnitude of the relationship between two variables depends on (i.e., differs according to) the value of one or more other variables. When interaction is present, simple additive techniques are inappropriate; hence, interaction is sometimes thought of as the absence of additivity. Synonyms: nonadditivity, conditioning effect, moderating effect, contingency effect. See also PATTERN VARIABLE, PRODUCT VARIABLE.

INTERVAL SCALE. A scale consisting of equal-sized units (dollars, years, etc.).

On an interval scale the distance between any two positions is of known size. Results from analytic techniques appropriate for interval scales will be affected by any non-linear transformation of the scale values. See also SCALE OF MEASUREMENT.

INTERVENING VARIABLE. A variable which is postulated to be a predictor of one or more dependent variables, and simultaneously predicted by one or more independent variables. Synonym: mediating variable.

KURTOSIS. Kurtosis indicates the extent to which a distribution is more peaked or flat-topped than a normal distribution.

LINEAR. The form of a relationship among variables such that when any two variables are plotted, a straight line results. A relationship is linear if the effect on a dependent variable of a change of one unit in an independent variable is the same for all possible such changes.

MATCHED SAMPLES. Two (or more) samples selected in such a way that each case (e.g., person) in one sample is matched-i.e., identical within specified limits-on one or more preselected characteristics with a corresponding case in the other sample. One example of matched samples is having repeated measures on the same individuals. Another example is linking husbands and wives. Matched samples are different from independent samples, where such case-by-case matching on selected characteristics has not been assured.

MEASURE OF ASSOCIATION. A number (a statistic) whose magnitude indicates the degree of correspondence-i.e., strength of relationship-between two variables. An example is the Pearson product-moment correlation coefficient. Measures of association are different from statistical tests of association (e.g., Pearson chi-square, F test) whose primary purpose is to assess the probability that the strength of a relationship is different from some preselected value (usually zero). See also STATISTICAL MEASURE, STATISTICAL TEST. MISSING DATA. Information that is not available for a particular case (e.g., person) for which at least some other information is available. This can occur for a variety of reasons, including a person's refusal or inability to answer a question, nonapplicability of a question, etc. For useful discussions of how to overcome problems caused by missing data in surveys see Hertel (1976) and Kim and Curry (1977).

MULTIVARIATE NORMALITY. The form of a distribution involving more than two variables in which the distribution of one variable is normal for each and every combination of categories of all other variables. See Harris (1975, page 231) for a discussion of multivariate normality. See also NORMAL DISTRIBUTION. NOMINAL SCALE. A classification of cases which defines their equivalence and non-equivalence, but implies no quantitative relationships or ordering among them. Analytic techniques appropriate for nominally scaled variables are not affected by any one-to-one transformation of the numbers assigned to the classes. See also SCALE OF MEASUREMENT.

NONADDITIVE. Not additive. See ADDITIVE, INTERACTION. NORMAL DISTRIBUTION. A particular form for the distribution of a variable which, when plotted, produces a ""bell"" shaped curve-symmetrical, rising smoothly from a small number of cases at both extremes to a large number of cases in the middle. Not all symmetrical bell-shaped distributions meet the definition of normality. See Hays (1973, page 296).

NORMALITY. See NORMAL DISTRIBUTION.

ORDINAL SCALE. A classification of cases into a set of ordered classes such that each case is considered equal to, greater than, or less than every other case. Analytic techniques appropriate for ordinally scaled variables are not affected by any monotonic transformation of the numbers assigned to the classes. See also SCALE OF MEASUREMENT.

OUTLYING CASE (OUTLIER). A case (e.g., person) whose score on a variable deviates substantially from the mean (or other measure of central tendency). Such cases can have disproportionately strong effects on statistics. PATTERN VARIABLE. A nominally scaled variable whose categories identify particular combinations (patterns) of scores on two or more other variables. For example, a party-by-gender pattern variable might be developed by classifying people into the following six categories: (1) Republican males, (2) Independent males, (3) Democratic males, (4) Republican females, (5) Independent females, (6) Democratic females. A pattern variable can be used to incorporate interaction in multivariate analysis.

PRODUCT VARIABLE. An intervally scaled variable whose scores are equal to the product obtained when the values of two other variables are multiplied together. A product variable can be used to incorporate certain types of interaction in multivariate analysis.

RANKS. The position of a particular case (e.g., person) relative to other cases on a defined scale-as in '1st place,' '2nd place,' etc. Note that when the actual values of the numbers designating the relative positions (the ranks) are used in analysis they are being treated as an interval scale, not an ordinal scale. See also INTERVAL SCALE, ORDINAL SCALE. SCALE OF MEASUREMENT. As used here, scale of measurement refers to the nature of the assumptions one makes about the properties of a variable; in particular, whether that variable meets the definition of nominal, ordinal, or interval measurement. See also NOMINAL SCALE, ORDINAL SCALE, INTERVAL SCALE. SKEWNESS. Skewness is a measure of lack of symmetry of a distribution. STANDARDIZED COEFFICIENT. When an analysis is performed on variables that have been standardized so that they have variances of 1.0, the estimates that

result are known as standardized coefficients; for example, a regression run on original variables produces unstandardized regression coefficients known as b's, while a regression run on standardized variables produces standardized regression coefficients known as betas. (In practice, both types of coefficients can be estimated from the original variables.) Blalock (1967), Hargens (1976), and Kim and Mueller (1976) provide useful discussions on the use of standardized coefficients.

STANDARDIZED VARIABLE. A variable that has been transformed by multiplication of all scores by a constant and/or by the addition of a constant to all scores. Often these constants are selected so that the transformed scores have a mean of zero and a variance (and standard deviation) of 1.0.

STATISTICAL INDEPENDENCE. A complete lack of covariation between variables; a lack of association between variables. When used in analysis of variance or covariance, statistical independence between the independent variables is sometimes referred to as a balanced design.

STATISTICAL MEASURE. A number (a statistic) whose size indicates the magnitude of some quantity of interest-e.g., the strength of a relationship, the amount of variation, the size of a difference, the level of income, etc. Examples include means, variances, correlation coefficients, and many others. Statistical measures are different from statistical tests. See also STATISTICAL TEST. STATISTICAL TEST. A number (a statistic) that can be used to assess the probability that a statistical measure deviates from some preselected value (often zero) by no more than would be expected due to the operation of chance if the cases (e.g., persons) studied were randomly selected from a larger population. Examples include Pearson chi-square, F test, t test, and many others. Statistical tests are different from statistical measures. See also STATISTICAL MEASURE.

TRANSFORMATION. A change made to the scores of all cases (e.g., persons) on a variable by the application of the same mathematical operation(s) to each score. (Common operations include addition of a constant, multiplication by a constant, taking logarithms, ranking, bracketing, etc.)

TWO-POINT SCALE. If each case is classified into one of two categories (e.g., yes/no, male/female, dead/alive), the variable is a two-point scale. For analytic purposes, two-point scales can be treated as nominal scales, ordinal scales, or interval scales.

WEIGHTED DATA. Weights are applied when one wishes to adjust the impact of cases (e.g., persons) in the analysis, e.g., to take account of the number of population units that each case represents. In sample surveys weights are most likely to be used with data derived from sample designs having different selection rates or with data having markedly different subgroup response rates.