bivariate analysis: The statistical analysis of the relationship between two variables.

cell frequency: The number of cases in a cell of a cross-tabulation (contingency table).

chi-square (χ^2) test for independence: A test of statistical significance used to assess the likelihood that an observed association between two variables could have occurred by chance.

consistency checking: A data-cleaning procedure involving checking for unreasonable patterns of responses, such as a 12-year-old who voted in the last US presidential election.

correlation coefficient: A statistical measure of the strength and direction of a linear relationship between two variables; it may vary from −1 to 0 to +1.

data cleaning: The detection and correction of errors in a computer datafile that may have occurred during data collection, coding, and/or data entry.

data matrix: The form of a computer datafile, with rows as cases and columns as variables; each cell represents the value of a particular variable (column) for a particular case (row).

data processing: The preparation of data for analysis.

descriptive statistics: Procedures for organizing and summarizing data.

dummy variable: A variable or set of variable categories recoded to have values of 0 and 1. Dummy coding may be applied to nominal- or ordinal-scale variables for the purpose of regression or other numerical analysis.

frequency distribution: A tabulation of the number of cases falling into each category of a variable.

histogram: A graphic display in which the height of a vertical bar represents the frequency or percentage of cases in each category of an interval/ratio variable.

imputation: A procedure for handling missing data in which missing values are assigned based on other information, such as the sample mean or known values of other variables.

inferential statistics: Procedures for determining the extent to which one may generalize beyond the data at hand.

listwise deletion: A common procedure for handling missing values in multivariate analysis that excludes cases which have missing values on any of the variables in the analysis.

marginal frequencies: Row and column totals in a contingency table (cross-tabulation) that represent the univariate frequency distributions for the row and column variables.

mean: The average value of a dataset, calculated by adding up the individual values and dividing by the total number of cases.

measures of association: Descriptive statistics used to measure the strength and direction of a bivariate relationship.

median: The midpoint in a distribution of interval- or ratio-scale data; indicates the point below and above which 50 percent of the values fall.

missing data: Refers to the absence of information on a variable for a given case.

mode: The value or category of a frequency distribution having the highest frequency; the most typical value.

multiple regression: A statistical method for determining the simultaneous effects of several independent variables on a dependent variable.

***N*:** An abbreviation representing the number of observations on which a statistic is based (e.g., $N = 753$).

null hypothesis: The hypothesis, associated with tests of statistical significance, that an observed relationship is due to chance; a test that is significant rejects the null hypothesis at a specified level of probability.

outliers: Unusual or suspicious values that are far removed from the preponderance of observations for a variable.

partial regression coefficient: Coefficients in a multiple-regression equation that estimate the effects of each independent variable on the dependent variable when all other variables in the equation are held constant. Also called *partial slope*.

partial table: A table in elaboration analysis which displays the original two-variable relationship for a single category of the control variable, thereby holding the control variable constant.

percentage distribution: A norming operation that facilitates interpreting and comparing frequency distributions by transforming each frequency to a common yardstick of 100 units (percentage points) in length; the number of cases in each category is divided by the total and multiplied by 100.

***R^2*:** A measure of fit in multiple regression that indicates approximately the proportion of the variation in the dependent variable predicted or “explained” by the independent variables.

range: The difference between the lowest and highest values in a distribution, which is usually reported by identifying these two extreme values.

regression analysis: A statistical method for analyzing bivariate (simple regression) and multivariate (multiple regression) relationships among interval- or ratio-scale variables.

regression line: A geometric representation of a bivariate regression equation that provides the best linear fit to the observed data by virtue of minimizing the sum of the squared deviations from the line; also called the *least squares line*.

residuals: The difference between observed values of the dependent variable and those predicted by a regression equation.

scatterplot: A graph plotting the values of two variables for each observation.

slope: A bivariate regression statistic indicating how much the dependent variable increases (or decreases) for every unit change in the independent variable; the slope of a regression line. Also called *regression coefficient*.

standard deviation: A measure of variability or dispersion that indicates the average “spread” of observations about the mean.

standardized regression coefficients: Coefficients obtained from a norming operation that puts partial-regression coefficients on common footing by converting them to the same metric of standard deviation units.

test of statistical significance: A statistical procedure used to assess the likelihood that the results of a study could have occurred by chance.

univariate analysis: The statistical analysis of one variable at a time.

wild-code checking: A data-cleaning procedure involving checking for out-of-range and other “illegal” codes among the values recorded for each variable.

***Y*-intercept:** The predicted value of the dependent variable in regression when the independent variable or variables have a value of zero; graphically, the point at which the regression line crosses the *Y*-axis.