Table of Contents

Introduction

1 Multivariate Statistics: Why?

Multivariate statistics are increasingly popular techniques used for analyzing complicated data sets. They provide analysis when there are many independent variables (IVs) and/or many dependent variables (DVs), all correlated with one another to varying degrees. Because of the difficulty in addressing complicated research questions with univariate analyses and because of the availability of canned software for performing multivariate analyses, multivariate statistics have become widely used. Indeed, a standard univariate statistics course only begins to prepare a student to read research literature or a researcher to produce it.

But how much harder are the multivariate techniques? Compared with the multivariate methods, univariate statistical methods are so straightforward and neatly structured that it is hard to believe they once took so much effort to master. Yet many researchers apply and correctly interpret results of intricate analysis of variance before the grand structure is apparent to them. The same can be true of multivariate statistical methods. Although we are delighted if you gain insights into the full multivariate general linear model, we have accomplished our goal if you feel comfortable selecting and setting up multivariate analyses and interpreting the computer output.

Multivariate methods are more complex than univariate by at least an order of magnitude. However, for the most part, the greater complexity requires few conceptual leaps. Familiar concepts such as sampling distributions and homogeneity of variance simply become more elaborate.

Multivariate models have not gained popularity by accident—or even by sinister design. Their growing popularity parallels the greater complexity of contemporary research. In psychology, for example, we are less and less enamored of the simple, clean, laboratory study, in which pliant, first-year college students each provides us with a single behavioral measure on cue.

1.1 The Domain of Multivariate Statistics: Numbers of IVs and DVs

Multivariate statistical methods are an extension of univariate and bivariate statistics. Multivariate statistics are the *complete* or general case, whereas univariate and bivariate statistics are special cases of the multivariate model. If your design has many variables, multivariate techniques often let you perform a single analysis instead of a series of univariate or bivariate analyses.

Variables are roughly dichotomized into two major types—independent and dependent. Independent variables (IVs) are the differing conditions (treatment vs. placebo) to which you expose your subjects, or the characteristics (tall or short) that the subjects themselves bring into the

From Chapter 1 of *Using Multivariate Statistics*, Sixth Edition. Barbara G. Tabachnick, Linda S. Fidell.

research situation. IVs are usually considered predictor variables because they predict the DVs—the response or outcome variables. Note that IV and DV are defined within a research context; a DV in one research setting may be an IV in another.

Additional terms for IVs and DVs are predictor-criterion, stimulus-response, task-performance, or simply input–output. We use IV and DV throughout this chapter to identify variables that belong on one side of an equation or the other, without causal implication. That is, the terms are used for convenience rather than to indicate that one of the variables caused or determined the size of the other.

The term *univariate statistics* refers to analyses in which there is a single DV. There may be, however, more than one IV. For example, the amount of social behavior of graduate students (the DV) is studied as a function of course load (one IV) and type of training in social skills to which students are exposed (another IV). Analysis of variance is a commonly used univariate statistic.

Bivariate statistics frequently refers to analysis of two variables, where neither is an experimental IV and the desire is simply to study the relationship between the variables (e.g., the relationship between income and amount of education). Bivariate statistics, of course, can be applied in an experimental setting, but usually they are not. Prototypical examples of bivariate statistics are the Pearson product–moment correlation coefficient and chi-square analysis.

With multivariate statistics, you simultaneously analyze multiple dependent and multiple independent variables. This capability is important in both nonexperimental (correlational or survey) and experimental research.

1.2 Experimental and Nonexperimental Research

A critical distinction between experimental and nonexperimental research is whether the researcher manipulates the levels of the IVs. In an experiment, the researcher has control over the levels (or conditions) of at least one IV to which a subject is exposed by determining what the levels are, how they are implemented, and how and when cases are assigned and exposed to them. Further, the experimenter randomly assigns subjects to levels of the IV and controls all other influential factors by holding them constant, counterbalancing, or randomizing their influence. Scores on the DV are expected to be the same, within random variation, except for the influence of the IV (Campbell & Stanley, 1966). If there are systematic differences in the DV associated with levels of the IV, these differences are attributed to the IV.

For example, if groups of undergraduates are randomly assigned to the same material but different types of teaching techniques, and afterward some groups of undergraduates perform better than others, the difference in performance is said, with some degree of confidence, to be caused by the difference in teaching technique. In this type of research, the terms *independent* and *dependent* have obvious meaning: the value of the DV depends on the manipulated level of the IV. The IV is manipulated by the experimenter and the score on the DV depends on the level of the IV.

In nonexperimental (correlational or survey) research, the levels of the IV(s) are not manipulated by the researcher. The researcher can define the IV, but has no control over the assignment of subjects to levels of it. For example, groups of people may be categorized into geographic area of residence (Northeast, Midwest, etc.), but only the definition of the variable is under researcher control. Except for the military or prison, place of residence is rarely subject to manipulation by a researcher. Nevertheless, a naturally occurring difference like this is often considered an IV and is used to predict some other nonexperimental (dependent) variable such as income. In this type of research, the distinction between IVs and DVs is usually arbitrary and many researchers prefer to call IVs *predictors* and DVs *criterion variables.*

In nonexperimental research, it is very difficult to attribute causality to an IV. If there is a systematic difference in a DV associated with levels of an IV, the two variables are said (with some degree of confidence) to be related, but the cause of the relationship is unclear. For example, income as a DV might be related to geographic area, but no causal association is implied.

Nonexperimental research takes many forms, but a common example is the survey. Typically, many people are surveyed, and each respondent provides answers to many questions, producing a large number of variables. These variables are usually interrelated in highly complex ways, but univariate and bivariate statistics are not sensitive to this complexity. Bivariate correlations between all pairs of variables, for example, could not reveal that the 20 to 25 variables measured really represent only two or three "supervariables."

If a research goal is to distinguish among subgroups in a sample (e.g., between Catholics and Protestants) on the basis of a variety of attitudinal variables, we could use several univariate *t* tests (or analyses of variance) to examine group differences on each variable separately. But if the variables are related, which is highly likely, the results of many *t* tests are misleading and statistically suspect.

With the use of multivariate statistical techniques, complex interrelationships among variables are revealed and assessed in statistical inference. Further, it is possible to keep the overall Type I error rate at, say, 5%, no matter how many variables are tested.

Although most multivariate techniques were developed for use in nonexperimental research, they are also useful in experimental research, in which there may be multiple IVs and multiple DVs. With multiple IVs, the research is usually designed so that the IVs are independent of each other and a straightforward correction for numerous statistical tests is available. With multiple DVs, a problem of inflated error rate arises if each DV is tested separately. Further, at least some of the DVs are likely to be correlated with each other, so separate tests of each DV reanalyze some of the same variance. Therefore, multivariate tests are used.

Experimental research designs with multiple DVs were unusual at one time. Now, however, with attempts to make experimental designs more realistic, and with the availability of computer programs, experiments often have several DVs. It is dangerous to run an experiment with only one DV and risk missing the impact of the IV because the most sensitive DV is not measured. Multivariate statistics help the experimenter design more efficient and more realistic experiments by allowing measurement of multiple DVs without violation of acceptable levels of Type I error.

One of the few considerations not relevant to choice of *statistical* technique is whether the data are experimental or correlational. The statistical methods "work" whether the researcher manipulated the levels of the IV or not. But attribution of causality to results is crucially affected by the experimental–nonexperimental distinction.

1.3 Computers and Multivariate Statistics

One answer to the question "Why multivariate statistics?" is that the techniques are now accessible by computer. Only the most dedicated number cruncher would consider doing real-life-sized problems in multivariate statistics without a computer. Fortunately, excellent multivariate programs are available in a number of computer packages.

Two packages are demonstrated in this chapter. Examples are based on programs in IBM SPSS and SAS.

If you have access to both packages, you are indeed fortunate. Programs within the packages do not completely overlap, and some problems are better handled through one package than the other. For example, doing several versions of the same basic analysis on the same set of data is particularly easy with IBM SPSS, whereas SAS has the most extensive capabilities for saving derived scores from data screening or from intermediate analyses.

Recent versions of the multivariate programs are implemented in Windows. All multivariate techniques may be implemented through syntax, and syntax itself is generated through menus. Then you may add or change syntax as desired for your analysis. For example, you may "paste" menu choices into a syntax window in IBM SPSS, edit the resulting text, and then run the program. Also, syntax generated by IBM SPSS menus is saved in the "journal" file (statistics.jnl), which may also be accessed and copied into a syntax window. Syntax generated by SAS menus is recorded in a "log" file. The contents may then be copied to an interactive window, edited, and run. Do not overlook the help files in these programs. Indeed, SAS and IBM SPSS now provide the entire set of user manuals on CD, often with more current information than is available in printed manuals.

With commercial computer packages, you need to know which version of the package you are using. Programs are continually being changed, and not all changes are immediately implemented at each facility. Therefore, many versions of the various programs are simultaneously in use at different institutions; even at one institution, more than one version of a package is sometimes available.

Program updates are often corrections of errors discovered in earlier versions. Occasionally, though, there are major revisions in one or more programs or a new program is added to the package. Sometimes defaults change with updates, so that the output looks different although syntax is the same. Check to find out which version of each package you are using. Then be sure that the manual you are using is consistent with the version in use at your facility. Also check updates for error correction in previous releases that may be relevant to some of your previous runs.

Information on availability and versions of software, macros, books, and the like changes almost daily. We recommend the Internet as a source of "keeping up."

1.4 Garbage In, Roses Out?

The trick in multivariate statistics is not in computation. This is easily done as discussed above. The trick is to select reliable and valid measurements, choose the appropriate program, use it correctly, and know how to interpret the output. Output from commercial computer programs, with their beautifully formatted tables, graphs, and matrices, can make garbage look like roses.

Second, when you use multivariate statistics, you rarely get as close to the raw data as you do when you apply univariate statistics to a relatively few cases. Errors and anomalies in the data that would be obvious if the data were processed by hand are less easy to spot when processing is entirely by computer. But the computer packages have programs to graph and describe your data in the simplest univariate terms and to display bivariate relationships among your variables. These programs provide preliminary analyses that are absolutely necessary if the results of multivariate programs are to be believed.

There are also certain costs associated with the benefits of using multivariate procedures. Benefits of increased flexibility in research design, for instance, are sometimes paralleled by increased ambiguity in interpretation of results. In addition, multivariate results can be quite sensitive to which analytic strategy is chosen (cf. Section 2.4) and do not always provide better protection against statistical errors than their univariate counterparts. Add to this the fact that occasionally you still cannot get a firm statistical answer to your research questions, and you may wonder if the increase in complexity and difficulty is warranted.

Frankly, we think it is. Slippery as some of the concepts and procedures are, these statistics provide insights into relationships among variables that may more closely resemble the complexity of the "real" world. And sometimes you get at least partial answers to questions that could not be asked at all in the univariate framework. For a complete analysis, making sense of your data usually requires a judicious mix of multivariate and univariate statistics.

The addition of multivariate statistical methods to your repertoire makes data analysis a lot more fun. If you liked univariate statistics, you will love multivariate statistics![1]

2 Some Useful Definitions

In order to describe multivariate statistics easily, it is useful to review some common terms in research design and basic statistics. Distinctions were made between IVs and DVs and between experimental and nonexperimental research in preceding sections.

2.1 Continuous, Discrete, and Dichotomous Data

In applying statistical techniques of any sort, it is important to consider the type of measurement and the nature of the correspondence between the numbers and the events that they represent. The distinction made here is among continuous, discrete, and dichotomous variables; you may prefer to

[1] Don't even think about it.

substitute the terms *interval* or *quantitative* for *continuous* and *nominal, categorical* or *qualitative* for *dichotomous* and *discrete.*

Continuous variables are measured on a scale that changes values smoothly rather than in steps. Continuous variables take on any values within the range of the scale, and the size of the number reflects the amount of the variable. Precision is limited by the measuring instrument, not by the nature of the scale itself. Some examples of continuous variables are time as measured on an old-fashioned analog clock face, annual income, age, temperature, distance, and grade point average (GPA).

Discrete variables take on a finite and usually small number of values, and there is no smooth transition from one value or category to the next. Examples include time as displayed by a digital clock, continents, categories of religious affiliation, and type of community (rural or urban).

Sometimes discrete variables are used in multivariate analyses as if continuous if there are numerous categories and the categories represent a quantitative attribute. For instance, a variable that represents age categories (where, say, 1 stands for 0 to 4 years, 2 stands for 5 to 9 years, 3 stands for 10 to 14 years, and so on up through the normal age span) can be used because there are a lot of categories and the numbers designate a quantitative attribute (increasing age). But the same numbers used to designate categories of religious affiliation are not in appropriate form for analysis with many of the techniques[2] because religions do not fall along a quantitative continuum.

Discrete variables composed of qualitatively different categories are sometimes analyzed after being changed into a number of dichotomous or two-level variables (e.g., Catholic vs. non-Catholic, Protestant vs. non-Protestant, Jewish vs. non-Jewish, and so on until the degrees of freedom are used). Recategorization of a discrete variable into a series of dichotomous ones is called *dummy variable coding.* The conversion of a discrete variable into a series of dichotomous ones is done to limit the relationship between the dichotomous variables and others to linear relationships. A discrete variable with more than two categories can have a relationship of any shape with another variable, and the relationship is changed arbitrarily if the assignment of numbers to categories is changed. Dichotomous variables, however, with only two points, can have only linear relationships with other variables; they are, therefore, appropriately analyzed by methods using correlation in which only linear relationships are analyzed.

The distinction between continuous and discrete variables is not always clear. If you add enough digits to the digital clock, for instance, it becomes for all practical purposes a continuous measuring device, whereas time as measured by the analog device can also be read in discrete categories such as hours or half hours. In fact, any continuous measurement may be rendered discrete (or dichotomous) with some loss of information, by specifying cutoffs on the continuous scale.

The property of variables that is crucial to the application of multivariate procedures is not the type of measurement so much as the shape of distribution. Non-normally distributed continuous variables and dichotomous variables with very uneven splits between the categories present problems to several of the multivariate analyses.

Another type of measurement that is used sometimes produces a rank order (ordinal) scale. This scale assigns a number to each subject to indicate the subject's position vis-à-vis other subjects along some dimension. For instance, ranks are assigned to contestants (first place, second place, third place, etc.) to provide an indication of who is the best—but not by how much. A problem with

[2] Some multivariate techniques (e.g., logistic regression, SEM) are appropriate for all types of variables.

ordinal measures is that their distributions are rectangular (one frequency per number) instead of normal, unless tied ranks are permitted and they pile up in the middle of the distribution.

In practice, we often treat variables as if they are continuous when the underlying scale is thought to be continuous, but the measured scale actually is ordinal, the number of categories is large—say, seven or more, and the data meet other assumptions of the analysis. For instance, the number of correct items on an objective test is technically not continuous because fractional values are not possible, but it is thought to measure some underlying continuous variable such as course mastery. Another example of a variable with ambiguous measurement is one measured on a Likert-type scale, in which consumers rate their attitudes toward a product as "strongly like," "moderately like," "mildly like," "neither like nor dislike," "mildly dislike," "moderately dislike," or "strongly dislike." As mentioned previously, even dichotomous variables may be treated as if continuous under some conditions. Thus, we often use the term "*continuous*" throughout the remainder of this chapter, whether the measured scale itself is continuous or the variable is to be treated as if continuous. We use the term "*discrete*" for variables with a few categories, whether the categories differ in type or quantity.

2.2 Samples and Populations

Samples are measured to make generalizations about populations. Ideally, samples are selected, usually by some random process, so that they represent the population of interest. In real life, however, populations are frequently best defined in terms of samples, rather than vice versa; the population is the group from which you were able to randomly sample.

Sampling has somewhat different connotations in nonexperimental and experimental research. In nonexperimental research, you investigate relationships among variables in some predefined population. Typically, you take elaborate precautions to ensure that you have achieved a representative sample of that population; you define your population, and then do your best to randomly sample from it.[3]

In experimental research, you attempt to create different populations by treating subgroups from an originally homogeneous group differently. The sampling objective here is to ensure that all subjects come from the same population before you treat them differently. Random sampling consists of randomly assigning subjects to treatment groups (levels of the IV) to ensure that, before differential treatment, all subsamples come from the same population. Statistical tests provide evidence as to whether, after treatment, all samples still come from the same population. Generalizations about treatment effectiveness are made to the type of subjects who participated in the experiment.

2.3 Descriptive and Inferential Statistics

Descriptive statistics describe samples of subjects in terms of variables or combinations of variables. Inferential statistical techniques test hypotheses about differences in populations on the basis of measurements made on samples of subjects. If reliable differences are found, descriptive statistics are then used to provide estimations of central tendency, and the like, in the population. Descriptive statistics used in this way are called *parameter estimates.*

[3] Strategies for random sampling are discussed in many sources, including Levy and Lemenshow (1999), Rea and Parker (1997), and de Vaus (2002).

Use of inferential and descriptive statistics is rarely an either–or proposition. We are usually interested in both describing and making inferences about a data set. We describe the data, find reliable differences or relationships, and estimate population values for the reliable findings. However, there are more restrictions on inference than there are on description. Many assumptions of multivariate statistical methods are necessary only for inference. If simple description of the sample is the major goal, many assumptions are relaxed.

2.4 Orthogonality: Standard and Sequential Analyses

Orthogonality is a perfect nonassociation between variables. If two variables are orthogonal, knowing the value of one variable gives no clue as to the value of the other; the correlation between them is zero.

Orthogonality is often desirable in statistical applications. For instance, factorial designs for experiments are orthogonal when two or more IVs are completely crossed with equal sample sizes in each combination of levels. Except for use of a common error term, tests of hypotheses about main effects and interactions are independent of each other; the outcome of each test gives no hint as to the outcome of the others. In orthogonal experimental designs with random assignment of subjects, manipulation of the levels of the IV, and good controls, changes in value of the DV can be unambiguously attributed to various main effects and interactions.

Similarly, in multivariate analyses, there are advantages if sets of IVs or DVs are orthogonal. If all pairs of IVs in a set are orthogonal, each IV adds, in a simple fashion, to prediction of the DV. Consider income as a DV with education and occupational prestige as IVs. If education and occupational prestige are orthogonal, and if 35% of the variability in income may be predicted from education and a different 45% is predicted from occupational prestige, then 80% of the variance in income is predicted from education and occupational prestige together.

Orthogonality can easily be illustrated in Venn diagrams, as shown in Figure 1. Venn diagrams represent shared variance (or correlation) as overlapping areas between two (or more) circles. The total variance for income is one circle. The section with horizontal stripes represents the part of income predictable from education, and the section with vertical stripes represents the part predictable from occupational prestige; the circle for education overlaps the circle for income 35% and the circle for occupational prestige overlaps 45%. Together, they account for 80% of the variability in income because education and occupational prestige are orthogonal and do not themselves overlap. There are similar advantages if a set of DVs is orthogonal. The overall effect of an IV can be partitioned into effects on each DV in an additive fashion.

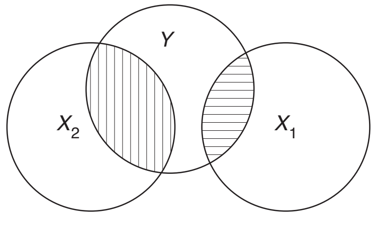


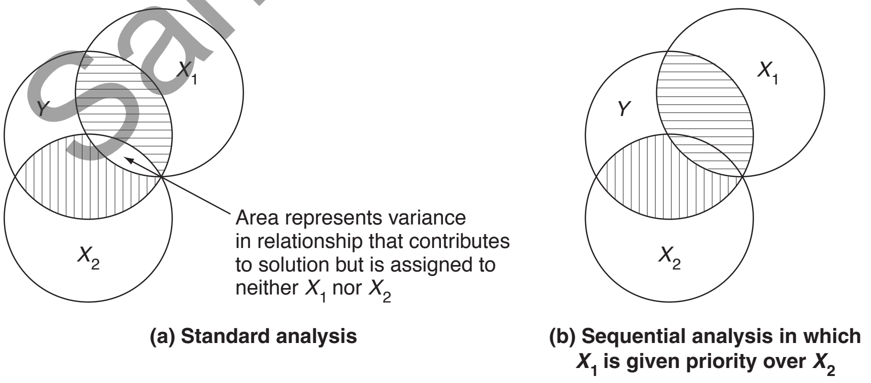
FIGURE 1 Venn diagram for Y (income), X_1 (education), and X_2 (occupational prestige).

Usually, however, the variables are correlated with each other (nonorthogonal). IVs in nonexperimental designs are often correlated naturally; in experimental designs, IVs become correlated when unequal numbers of subjects are measured in different cells of the design. DVs are usually correlated because individual differences among subjects tend to be consistent over many attributes.

When variables are correlated, they have shared or overlapping variance. In the example of Figure 2, education and occupational prestige correlate with each other. Although the independent contribution made by education is still 35% and that by occupational prestige is 45%, their joint contribution to prediction of income is not 80%, but rather something smaller due to the overlapping area shown by the arrow in Figure 2(a). A major decision for the multivariate analyst is how to handle the variance that is predictable from more than one variable. Many multivariate techniques have at least two strategies for handling it, but some have more.

In standard analysis, the overlapping variance contributes to the size of summary statistics of the overall relationship but is not assigned to either variable. Overlapping variance is disregarded in assessing the contribution of each variable to the solution. Figure 2(a) is a Venn diagram of a standard analysis in which overlapping variance is shown as overlapping areas in circles; the unique contributions of X_1 and X_2 to prediction of Y are shown as horizontal and vertical areas, respectively, and the total relationship between Y and the combination of X_1 and X_2 is those two areas plus the area with the arrow. If X_1 is education and X_2 is occupational prestige, then in standard analysis, X_1 is "credited with" the area marked by the horizontal lines and X_2 by the area marked by vertical lines. Neither of the IVs is assigned the area designated with the arrow. When X_1 and X_2 substantially overlap each other, very little horizontal or vertical area may be left for either of them, despite the fact that they are both related to Y. They have essentially knocked each other out of the solution.

Sequential analyses differ, in that the researcher assigns priority for entry of variables into equations, and the first one to enter is assigned both unique variance and any overlapping variance it has with other variables. Lower-priority variables are then assigned on entry their unique and any remaining overlapping variance. Figure 2(b) shows a sequential analysis for the same case as Figure 2(a), where X_1 (education) is given priority over X_2 (occupational prestige). The total variance explained is the same as in Figure 2(a), but the relative contributions of X_1 and X_2 have changed;



Area represents variance in relationship that contributes to solution but is assigned to neither X_1 nor X_2

(a) Standard analysis

(b) Sequential analysis in which X_1 is given priority over X_2

FIGURE 2 Standard (a) and sequential (b) analyses of the relationship between Y, X_1, and X_2. Horizontal shading depicts variance assigned to X_1. Vertical shading depicts variance assigned to X_2.

education now shows a stronger relationship with income than in the standard analysis, whereas the relation between occupational prestige and income remains the same.

The choice of strategy for dealing with overlapping variance is not trivial. If variables are correlated, the overall relationship remains the same, but the apparent importance of variables to the solution changes depending on whether a standard or a sequential strategy is used. If the multivariate procedures have a reputation for unreliability, it is because solutions change, sometimes dramatically, when different strategies for entry of variables are chosen. However, the strategies also ask different questions of the data, and it is incumbent on the researcher to determine exactly which question to ask.

3 Linear Combinations of Variables

Multivariate analyses combine variables to do useful work, such as predict scores or predict group membership. The combination that is formed depends on the relationships among the variables and the goals of analysis, but in most cases, the combination is linear. A linear combination is one in which each variable is assigned a weight (e.g., W_1), and then the products of weights and the variable scores are summed to predict a score on a combined variable. In Equation 1, Y' (the predicted DV) is predicted by a linear combination of X_1 and X_2 (the IVs).

$$Y' = W_1X_1 + W_2X_2 \tag{1}$$

If, for example, Y' is predicted income, X_1 is education, and X_2 is occupational prestige, the best prediction of income is obtained by weighting education (X_1) by W_1 and occupational prestige (X_2) by W_2 before summing. No other values of W_1 and W_2 produce as good a prediction of income.

Notice that Equation 1 includes neither X_1 nor X_2 raised to powers (exponents) nor a product of X_1 and X_2. This seems to severely restrict multivariate solutions until one realizes that X_1 could itself be a product of two different variables or a single variable raised to a power. For example, X_1 might be education squared. A multivariate solution does not produce exponents or cross-products of IVs to improve a solution, but the researcher can include *X*s that are cross-products of IVs or are IVs raised to powers. Inclusion of variables raised to powers or cross-products of variables has both theoretical and practical implications for the solution. Berry (1993) provides a useful discussion of many of the issues.

The size of the *W* values (or some function of them) often reveals a great deal about the relationship between DVs and IVs. If, for instance, the *W* value for some IV is zero, the IV is not needed in the best DV–IV relationship. Or if some IV has a large *W* value, then the IV tends to be important to the relationship. Although complications prevent interpretation of the multivariate solution from the sizes of the *W* values alone, they are nonetheless important in most multivariate procedures.

The combination of variables can be considered a supervariable, not directly measured but worthy of interpretation. The supervariable may represent an underlying dimension that predicts something or optimizes some relationship. Therefore, the attempt to understand the meaning of the combination of IVs is worthwhile in many multivariate analyses.

In the search for the best weights to apply in combining variables, computers do not try out all possible sets of weights. Various algorithms have been developed to compute the weights. Most algorithms involve manipulation of a correlation matrix, a variance–covariance matrix, or a sum-of-squares and cross-products matrix. Section 6 describes these matrices in very simple terms and shows their development from a very small data set. Appendix ‘A Skimpy Introduction to Matrix Algebra’ describes some terms and manipulations appropriate to matrices.

4 Number and Nature of Variables to Include

Attention to the number of variables included in analysis is important. A general rule is to get the best solution with the fewest variables. As more and more variables are included, the solution usually improves, but only slightly. Sometimes the improvement does not compensate for the cost in degrees of freedom of including more variables, so the power of the analyses diminishes.

A second problem is *overfitting.*With overfitting, the solution is very good; so good, in fact, that it is unlikely to generalize to a population. Overfitting occurs when too many variables are included in an analysis relative to the sample size. With smaller samples, very few variables can be analyzed. Generally, a researcher should include only a limited number of uncorrelated variables in each analysis,[4] fewer with smaller samples.

Additional considerations for inclusion of variables in a multivariate analysis include cost, availability, meaning, and theoretical relationships among the variables. Except in analysis of structure, one usually wants a small number of valid, cheaply obtained, easily available, uncorrelated variables that assess all the theoretically important dimensions of a research area. Another important consideration is reliability. How stable is the position of a given score in a distribution of scores when measured at different times or in different ways? Unreliable variables degrade an analysis, whereas reliable ones enhance it. A few reliable variables give a more meaningful solution than a large number of less reliable variables. Indeed, if variables are sufficiently unreliable, the entire solution may reflect only measurement error. Further considerations for variable selection are mentioned as they apply to each analysis.

5 Statistical Power

A critical issue in designing any study is whether there is adequate power. Power, as you may recall, represents the probability that effects that actually exist have a chance of producing statistical significance in your eventual data analysis. For example, do you have a large enough sample size to show a significant relationship between GRE and GPA if the actual relationship is fairly large? What if the relationship is fairly small? Is your sample large enough to reveal significant effects of treatment on your DV(s)?

[4] The exceptions are analysis of structure, such as factor analysis, in which numerous correlated variables are measured.

Issues of power are best considered in the planning state of a study when the researcher determines the required sample size. The researcher estimates the size of the anticipated effect (e.g., an expected mean difference), the variability expected in assessment of the effect, the desired alpha level (ordinarily .05), and the desired power (often .80). These four estimates are required to determine the necessary sample size. Failure to consider power in the planning stage often results in failure to find a significant effect (and an unpublishable study). The interested reader may wish to consult Cohen (1965, 1988), Rossi (1990), or Sedlmeier and Gigerenzer (1989) for more detail.

There is a great deal of software available to help you estimate the power available with various sample sizes for various statistical techniques, and to help you determine necessary sample size given a desired level of power (e.g., an 80% probability of achieving a significant result if an effect exists) and expected sizes of relationships. One of these programs that estimates power for several techniques is NCSS PASS (Hintze, 2011). Many other programs are reviewed (and sometimes available as shareware) on the Internet.

6 Data Appropriate for Multivariate Statistics

An appropriate data set for multivariate statistical methods consists of values on a number of variables for each of several subjects or cases. For continuous variables, the values are scores on variables. For example, if the continuous variable is the GRE (Graduate Record Examination), the values for the various subjects are scores such as 500, 420, 650, and so on. For discrete variables, values are number codes for group membership or treatment. For example, if there are three teaching techniques, students who receive one technique are arbitrarily assigned a “1,” those receiving another technique are assigned a “2,” and so on.

6.1 The Data Matrix

The data matrix is an organization of scores in which rows (lines) represent subjects and columns represent variables. An example of a data matrix with six subjects[5] and four variables is given in Table 1. For example, X_1 might be type of teaching technique, X_2 score on the GRE, X_3 GPA, and X_4 gender, with women coded 1 and men coded 2.

TABLE 1 A Data Matrix of Hypothetical Scores

Student	X_1	X_2	X_3	X_4
1	1	500	3.20	1
2	1	420	2.50	2
3	2	650	3.90	1
4	2	550	3.50	2
5	3	480	3.30	1
6	3	600	3.25	2

[5] Normally, of course, there are many more than six subjects.

Data are entered into a data file with long-term storage accessible by computer in order to apply computer techniques to them. Each subject starts with a new row (line). Information identifying the subject is typically entered first, followed by the value of each variable for that subject.

Scores for each variable are entered in the same order for each subject. If there are more data for each subject than can be accommodated on a single line, the data are continued on additional lines, but all of the data for each subject are kept together. All of the computer package manuals provide information on setting up a data matrix.

In this example, there are values for every variable for each subject. This is not always the case with research in the real world. With large numbers of subjects and variables, scores are frequently missing on some variables for some subjects. For instance, respondents may refuse to answer some kinds of questions, or some students may be absent the day when a particular test is given, and so forth. This creates missing values in the data matrix. To deal with missing values, first build a data file in which some symbol is used to indicate that a value on a variable is missing in data for a subject. The various programs have standard symbols, such as a dot (.), for this purpose. You can also use other symbols, but it is often just as convenient to use one of the default symbols.

6.2 The Correlation Matrix

Most readers are familiar with **R**, a correlation matrix. **R** is a square, symmetrical matrix. Each row (and each column) represents a different variable, and the value at the intersection of each row and column is the correlation between the two variables. For instance, the value at the intersection of the second row, third column, is the correlation between the second and the third variables. The same correlation also appears at the intersection of the third row, second column. Thus, correlation matrices are said to be symmetrical about the main diagonal, which means they are mirror images of themselves above and below the diagonal from top left to bottom right. Hence, it is common practice to show only the bottom half or the top half of an *R* matrix. The entries in the main diagonal are often omitted as well, since they are all ones—correlations of variables with themselves.[6]

Table 2 shows the correlation matrix for X_2, X_3, and X_4 of Table 1. The value .85 is the correlation between X_2 and X_3 and it appears twice in the matrix (as do other values). Other correlations are as indicated in the table.

Many programs allow the researcher a choice between analysis of a correlation matrix and analysis of a variance–covariance matrix. If the correlation matrix is analyzed, a unit-free result is produced. That is, the solution reflects the relationships among the variables but not in the metric in which they are measured. If the metric of the scores is somewhat arbitrary, analysis of **R** is appropriate.

TABLE 2 Correlation Matrix for Part of Hypothetical Data for Table 1

		X_2	X_3	X_4
	X_2	1.00	.85	−.13
R =	X_3	.85	1.00	−.46
	X_4	−.13	−.46	1.00

[6] Alternatively, other information such as standard deviations is inserted.

6.3 The Variance–Covariance Matrix

If scores are measured along a meaningful scale, it is sometimes appropriate to analyze a variance–covariance matrix. A variance–covariance matrix, Σ, is also square and symmetrical, but the elements in the main diagonal are the variances of each variable, and the off-diagonal elements are covariances between pairs of different variables.

Variances, as you recall, are averaged squared deviations of each score from the mean of the scores. Since the deviations are averaged, the number of scores included in computation of a variance is not relevant, but the metric in which the scores are measured is relevant. Scores measured in large numbers tend to have large numbers as variances, and scores measured in small numbers tend to have small variances.

Covariances are averaged cross-products (product of the deviation between one variable and its mean and the deviation between a second variable and its mean). Covariances are similar to correlations except that they, like variances, retain information concerning the scales in which the variables are measured. The variance–covariance matrix for the continuous data in Table 1 appears in Table 3.

6.4 The Sum-of-Squares and Cross-Products Matrix

The matrix, **S**, is a precursor to the variance–covariance matrix in which deviations are not yet averaged. Thus, the size of the entries depends on the number of cases as well as on the metric in which the elements were measured. The sum-of-squares and cross-products matrix for X_2, X_3, and X_4 in Table 1 appears in Table 4.

The entry in the major diagonal of the matrix **S** is the sum of squared deviations of scores from the mean for that variable, hence, "sum of squares," or SS. That is, for each variable, the value in the major diagonal is

$$\mathrm{SS}(X_i) = \sum_{i=1}^{N} (X_{ij} - \overline{X}_j)^2 \tag{2}$$

where $i = 1, 2, \ldots, N$
N = the number of subjects
j = the variable identifier
X_{ij} = the score on variable j by subject i
$\overline{X}_j$ = the mean of all scores on the jth variable

TABLE 3 Variance–Covariance Matrix for Part of Hypothetical Data of Table 1

		X_2	X_3	X_4
	X_2	7026.66	32.80	−6.00
$\Sigma =$	X_3	32.80	.21	−.12
	X_4	−6.00	−.12	.30

TABLE 4 Sum-of-Squares and Cross-Products Matrix for Part of Hypothetical Data of Table 1

		X_2	X_3	X_4
	X_2	35133.33	164.00	−30.00
S =	X_3	164.00	1.05	−0.58
	X_4	−30.00	−0.58	1.50

For example, for X_4, the mean is 1.5. The sum of squared deviations around the mean and the diagonal value for the variable is

$$\sum_{i=1}^{6}(X_{i4} - \overline{X}_4)^2 = (1 - 1.5)^2 + (2 - 1.5)^2 + (1 - 1.5)^2 + (2 - 1.5)^2 + (1 - 1.5)^2 + (2 - 1.5)^2$$
$$= 1.50$$

The off-diagonal elements of the sum-of-squares and cross-products matrix are the cross-products—the sum of products (SP)—of the variables. For each pair of variables, represented by row and column labels in Table 4, the entry is the sum of the product of the deviation of one variable around its mean times the deviation of the other variable around its mean.

$$\text{SP}(X_j X_k) = \sum_{i=1}^{N}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k) \tag{3}$$

where j identifies the first variable, k identifies the second variable, and all other terms are as defined in Equation 1. (Note that if $j = k$, Equation 3 becomes identical to Equation 2.)

For example, the cross-product term for variables X_2 and X_3 is

$$\sum_{i=1}^{N}(X_{i2} - \overline{X}_2)(X_{i3} - \overline{X}_3) = (500 - 533.33)(3.20 - 3.275) + (420 - 533.33)(2.50 - 3.275)$$
$$+ \cdots + (600 - 533.33)(3.25 - 3.275) = 164.00$$

Most computations start with **S** and proceed to Σ or **R**. The progression from a sum-of-squares and cross-products matrix to a variance–covariance matrix is simple.

$$\Sigma = \frac{1}{N - 1}\mathbf{S} \tag{4}$$

The variance–covariance matrix is produced by dividing every element in the sum-of-squares and cross-products matrix by $N - 1$, where N is the number of cases.

The correlation matrix is derived from an **S** matrix by dividing each sum-of-squares by itself (to produce the 1s in the main diagonal of **R**) and each cross-product of the **S** matrix by the square root of the product of the sum-of-squared deviations around the mean for each of the variables in the pair. That is, each cross-product is divided by

$$\text{Denominator}(X_j X_k) = \sqrt{\Sigma(X_{ij} - X_j)^2\Sigma(X_{ik} - X_k)^2} \tag{5}$$

where terms are defined as in Equation 3.

For some multivariate operations, it is not necessary to feed the data matrix to a computer program. Instead, an **S** or an **R** matrix is entered, with each row (representing a variable) starting a new line. Often, considerable computing time and expense are saved by entering one or the other of these matrices rather than raw data.

6.5 Residuals

Often a goal of analysis or test of its efficiency is its ability to reproduce the values of a DV or the correlation matrix of a set of variables. For example, we might want to predict scores on the GRE (X_2) of Table 1 from knowledge of GPA (X_3) and gender (X_4). After applying the proper statistical operations—a multiple regression in this case—a predicted GRE score for each student is computed by applying the proper weights for GPA and gender to the GPA, and gender scores for each student. But because we already obtained GRE scores for the sample of students, we are able to compare the predicted score with the obtained GRE score. The difference between the predicted and obtained values is known as the *residual* and is a measure of error of prediction.

In most analyses, the residuals for the entire sample sum to zero. That is, sometimes the prediction is too large and sometimes it is too small, but the average of all the errors is zero. The squared value of the residuals, however, provides a measure of how good the prediction is. When the predictions are close to the obtained values, the squared errors are small. The way that the residuals are distributed is of further interest in evaluating the degree to which the data meet the assumptions of multivariate analyses.

References

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage.

Campbell, D. R., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. New York: Rand McNally.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

De Vaus, D. (Ed.). (2002). *Social surveys* (Vol. 4 set). London: Sage.

Hintze, J. L. (2011). *PASS user's guide—II: PASS 11 power analysis and sample size for windows.* Kaysville, UT: NCSS [documentation on software CD].

Levy, P. S., & Lemeshow, S. (1999). *Sampling populations* (3rd ed.). New York: Wiley Interscience.

Rea, L. M., & Parker, R. A. (1997). *Designing and conducting survey research: A comprehensive guide*. San Francisco: Jossey-Bass.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.