Multivariate Analyses

The word "multivariate" in the term multivariate analysis has been defined variously by different authors and has no single definition. Most statistics books on multivariate statistics define multivariate statistics as tests that involve multiple dependent (or response) variables together (Pituch & Stevens, 2105; Tabachnick & Fidell, 2013; Tatsuoka, 1988). But common usage by some researchers and authors also include analysis with multiple independent variables, such as multiple regression, in their definition of multivariate statistics (e.g., Huberty, 1994). Some analyses, such as principal components analysis or canonical correlation, really have no independent or dependent variable as such, but could be conceived of as analyses of multiple responses. A more strict statistical definition would define multivariate analysis in terms of two or more random variables and their multivariate distributions (e.g., Timm, 2002), in which joint distributions must be considered to understand the statistical tests. I suspect the statistical definition is what underlies the selection of analyses that are included in most multivariate statistics books. Multivariate statistics texts nearly always focus on continuous dependent variables, but continuous variables would not be required to consider an analysis to be multivariate.

Huberty (1994) describes the goals of multivariate statistical tests as including prediction (of continuous outcomes or group membership), understanding association among one or more independent variables with a set of dependent variables, testing for group differences, or discovering the inherent structure behind the associations of a set of measures. The rationale for needing multivariate statistics has included: a) reduction of Type I error by avoiding multiple univariate statistical tests, b) avoiding variation in results from individual univariate tests of different measures due to random chance, c) combining similar variables into a linear composite has the potential for increasing the likelihood of finding group differences or relationships among predictors and outcomes, and d) data reduction (see Pituch & Stevens, 2015, and Tabachnick and Fidell, 2013 for discussions).

Below is a brief description of most of the statistical analyses that have traditionally been considered under the term “multivariate analysis.” We have already considered two of these is some detail (for more on these, see the "Multivariate Analysis of Variance" and "Principal Components Analysis" handouts), but I've included their brief summaries here for completeness.

Multivariate Analysis of Variance (MANOVA)

MANOVA makes groups comparisons on a set of related dependent variables (e.g., a comparison of whether anxiety, depression, and perceived stress differ between two psychotherapy methods). Two or more groups can be compared and MANOVA can be extended to factorial designs (multiple independent variables) and to include covariates (MANCOVA).

Principal Components Analysis (PCA)

PCA seeks to discover a smaller set of more general constructs that underlie a larger set of response variables. For example, an educational test may have 100 questions, and PCA could be used to discover whether there are several types of abilities that are responsible for the patterns for the answers to the full set of questions (e.g., mathematical abilities, verbal abilities, analytic abilities). The goal can be data reduction (as with analysis of "big data"), measurement development or psychometric evaluation. PCA is the initial step used for exploratory factor analysis methods to decide the number of factors to extract. The mathematical method of deriving linear composite weights used with principal components is central to most traditional multivariate statistics and also integral to many big data approaches.

There has been a major resurgence of interest in PCA because it has become a central feature of big data and machine learning (Fan, Sun, Zhou, & Zhu, 2014). Its relatively simple computation and goal of finding a fewer set of meaningful components is incorporated into the data driven approaches to deal with the huge complexities and wealth of information gathered electronically. Spectral analysis or spectral decomposition, used in physics, X-rays, medical applications, astronomy, also has PCA at its core.

Factor Analysis
Factor analysis may have the same goals as PCA of data reduction, measurement development, and psychometric evaluation, but differs in the statistical and theoretical underpinnings. PCA is typically the initial step in exploratory factor analysis in which the researcher chooses how many factors to extract. There are a number of decision points in factor analysis, including the number of factors to extract, whether the factors should be considered correlated or not, the algorithm used for scaling the loadings, the interpretation of the factors, which items belong to which factors, and whether to eliminate items. A major distinction is between exploratory factor analyses (EFA) and confirmatory factor analysis (CFA). Exploratory factor analysis is generally recommended when the researcher does not have *a priori* hypotheses about the factor structure of a set of items, whereas testing confirmatory factor analysis requires specification of a hypothesized factor structure to conduct. Although the process and software used for conducting these two general types of factor analysis differ, either method may involve approaches and goals that vary along a continuum of exploratory to confirmatory. How exploratory one process is depends more on the extent to which the researcher follows an *a priori* process, such as whether he or she has initial hypotheses about the factor structure when conducting an EFA or how many modifications are made when conducting a CFA.

Discriminant Function Analysis
Discriminant function analysis or discriminant analysis has the goal of determining membership based on a set of related measures (predictive) and understanding which components of a set of variables groups differ on (descriptive). As one possible example, a researcher might use a set of ability and cognitive measures to classify students into a set of learning disability categories, such as auditory processing disorders, dyslexia, visual/perceptual deficits. In one sense, discriminant analysis uses a set of variables as predictors of classes, and so could be called "MANOVA turned around" (Tabachnick & Fidell, 2001, p. 46). Weighted composites of the set of variables (known as discriminant functions, equivalent to principal component weighting) are used to discriminate among groups. Discriminant analysis can have the same goals as logistic regression, in which a set of variables is used to classify cases or predict group membership. However, discriminant analysis has equivalencies with ordinary least squares regression if used to predict a binary variable (Cohen, Cohen, West, & Aiken, 2003; Stevens, 1972)—but OLS regression is *not recommended* with categorical outcomes. When data are nonnormal and covariance matrices are not equal across classification groups, logistic regression is likely preferable for the prediction goal (Press & Wilson, 1978). Use of prediction with latent class models or structural equation modeling to investigate similar hypotheses has largely replace discriminant function analysis (Graham, 2008; Magidson & Vermunt, 2002).

Canonical Correlation
Canonical correlation or canonical analysis (or sometimes set correlation) is used to investigate the correlation between sets of variables. Linear composites, called canonical variates, are formed. For example, we might be interested in how a set of personality measures (conscientiousness, openness, neuroticism) as associated with a set of mental health measures (depression, anxiety, perceived stress). Canonical variates are much like the linear composites from PCA or MANOVA (all three were developed by Hotelling). In fact, MANOVA (see Cohen, Cohen, West, & Aiken, 2003, Chapter 16) and the multiple correlation coefficient from univariate regression (Rozboom, 1965) both can be considered as special cases of canonical correlation analysis. Instead of forming the composites to maximize group differences, as in MANOVA, however, the composites are formed to maximize the association between the sets of canonical variates. Just as with PCA multiple variates (components) may underlie a set of variables, and information about the association of each variable with the variate (component) can be obtained to make sense of the meaning of the variates and the relative importance of each variable.

Multivariate Regression
The univariate multiple regression model can be extended to prediction of a set of dependent variables. As an example, a researcher might be interested in the independent and relative effects of a set of health behaviors (e.g., smoking, exercise, diet) on a set of dependent measures of health (self-rated health,

health conditions, functioning). These models, which seem to be infrequently used in the social sciences, can be seen as a generalization of both the multiple regression model, MANOVA, and MANCOVA (Pituch & Stevens, 2015) and can be related to canonical correlation and partial set correlations (Cohen, Cohen, West, & Aiken, 2003).

Cluster Analysis
Cluster analysis is used to group or cluster sets of similar objects (see Aldenderfer & Bashfield, 1984 for an introduction). The objects could be individuals, photos, or ratings, for example. The method is popular in machine learning and pattern recognition, among other areas. Cluster analysis is not a single statistical method, per se. The term refers more to a variety of techniques that may be used for grouping based on similarity. One of the most common is based a multivariate similarity (k-means) and distances from a centroid mean (e.g., Mahalanobis distance). Principal components can also be used as a basis for clustering (Dunteman, 1989; Ding & He, 2004, show how k-means and the PCA method are related).

Multidimensional Scaling (MDS)
MDS examines paired distances among a set of objects and graphically represents those distances. Objects can be nearly anything that can be compared on a pairwise basis (similarity or dissimilarity), including perceived similarity, favorability, color perception, or physical locations. Results are usually plotted on a two-dimensional coordinate system, but other more elaborate graphical presentations can be used. There are many mathematical approaches to quantifying multiple dimensional distances, including mean centroid, median, non-Euclidean and others. Depending on the research question and the "objects" that are compared, MDS can have similar goals to principal components factor analysis (identifying meaningful clusters of relationships) and cluster analysis.

Structural Equation Modeling (SEM) for Multivariate Hypotheses
SEM, sometimes also called covariance structural analysis or causal modeling, is a combination of regression analysis and confirmatory factor analysis. A wide variety of complex path models can be tested with or without latent variables. Because of the possibility of inclusion of more than one dependent variable, SEM can be classified as a multivariate technique. Latent variables involve a confirmatory factor analysis of a set or related items or measures (often an end in itself) that estimates a common unobserved construct, which can be used to predict or predicted by other variables in a model. Multiple groups can be compared and measured variables can be continuous, binary, or ordinal. Because of its very general nature, allowing for the specification of a large variety of possible models, SEM can be used to test virtually any of the research questions of interest in the multivariate analyses described above. SEM, however, differs in one important way—*a priori* models for factor structures as well as prediction must be specified to test the model. Because of this difference in approach, Fornell (1982; Fornell & Larcker, 1987) called SEM a "second-generation multivariate analysis" approach. This difference is perhaps most clear in how EFA and CFA are conducted. By default, EFA allows for all items to correlate with (load on) all factors, whereas, CFA requires a specification of which items relate to which factors. Most multivariate statistics can be shown to be related to principal components, in which group differences or predictive relationships among linear composites, which assume no measurement error, are examined. SEM investigates these same hypotheses using latent variables, assuming measurement error when estimating the factors. SEM can be shown to include ANOVA (Green & Thompson, 2006), MANOVA (Cole, Maxwell, Arvey, & Salas, 1993), ANCOVA (Aiken, Stein, & Bentler, 1994; Sörbom, 1978), discriminant analysis (Graham, 2008), canonical correlation (Graham, 2008), and multivariate regression as special cases (where measurement errors are zero for multivariate regression), as a result. In addition to an *a priori* orientation to model testing and the ability to estimate and correct for measurement error, SEM has a number of other advantages, including model fit comparison which allows investigation or measurement equivalence across groups, flexibility for estimating complex error structures, and generalized models with non-continuous factor indicators and outcome variables.

References

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology, 62*(3), 488.

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis: Quantitative applications in the social sciences. Beverly Hills: Sage Publication.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis in the behavioral sciences (Third Edition*). Mahwah, NJ: Erlbaum.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. Psychological Bulletin, 114(1), 174.

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (p. 29). ACM.Huberty, C. J. (1994). Why multivariable analyses? *Educational and Psychological Measurement*, *54,* 620-627.

Dunteman, G. H. (1989). *Principal component analysis. Quantitative applications in the social sciences series (vol. 69)*. Thousand Oaks, CA: Sage.

Fan, J., Sun, Q., Zhou, W. X., & Zhu, Z. (2014). Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, 1-13.

Fornell, C. G. (1982). *A second generation of multivariate analysis: An overview*. In C. Fornell (Ed.), A second generation of multivariate analysis (pp. 1—21). New York: Praeger.

Fornell, C., & Larcker, D. (1987). A second generation of multivariate analysis: Classification of methods and implications for marketing research. *Review of marketing, 51*, 407-450.

Graham, J. M. (2008). The general linear model as structural equation modeling. J*ournal of Educational and Behavioral Statistics, 33*, 485-506.

Green, S. B., & Thompson, M. S. (2006). Structural equation modeling for conducting tests of differences in multiple means. *Psychosomatic medicine, 68*, 706-717.

Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research, 20*(1), 36-43.

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS.* New York: Routledge.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*(364), 699-705.

Rozeboom, W. W. (1965). Linear correlations between sets of variables. *Psychometrika, 30*, 57-71.

Stevens, J. P. (1972). Four methods of analyzing between variation for the k-group MANOVA problem. *Multivariate Behavioral Research, 7*, 499-522.

Sörbom, D. 1978. An alternative to the methodology for analysis of covariance. Psychometrika. 43:38 1-96

Tabachnick, B.G. and Fidell, L.S. (2001). *Using multivariate statistics, fourth edition*. Boston: Allyn and Bacon.

Tabachnick, B.G. and Fidell, L.S. (2013). *Using multivariate statistics, sixth edition*. Boston: Pearson.

Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research*. New York: Macmillan.

Timm, N, H. (2002). *Applied multivariate analysis.* New York: Springer.