Using Discriminant Analysis in Marketing Research: Part 1

Discriminant function analysis (DFA) is one of the most versatile and most often used multivariate procedures in marketing research. Five major areas of data analysis where DFA is appropriate and useful are discussed in this article, which will be in two parts. Part two will appear in the next issue.

To start, let’s first review just what DFA is. In marketing research we are often faced with a situation in which we have two or more groups (buyers–nonbuyers, age cohorts, etc.) or items (brands, firms, concepts, etc.) and we want to gain a better understanding of how these groups or items differ in terms of some set of explanatory (independent) metric variables, such as a set of attribute or performance ratings that we assume to be equal-interval. (Note that some statisticians accept the use of 0–1 dichotomous variables in the independent set, but here we restrict the discussion to interval-level variables.)

If the objective of the research is to understand how the groups or items differ, we could conduct a one-way analysis of variance (ANOVA) on each independent variable (such as a brand attribute rating scale) across the group (brand) means. But more often than not, in practical marketing research the independent variables—those brand-rating scales—are correlated to some extent. Therefore, there is a distinct possibility that a series of one-way ANOVA’s will show that many of the independent variables have group (brand) means that are significantly different, when in reality only one or two nonredundant ones do. In addition, if there is a large number of independent variables, we may see differences between groups by chance alone where there really are none, because of the accumulation of Type I error.

Discriminant function analysis addresses this problem very well. In DFA, the intercorrelation of variables is addressed by partialing (or partitioning) the correlations between independent variables. That is, when DFA uses one independent variable to explain differences between the groups, the
remaining variables are "adjusted" so that any difference they show between groups is not due to correlation those other independent variables may have with the first variable. The result is that DFA addresses only the unduplicated variance between groups.

This characteristic of DFA then leads us to its first major application in marketing research—*hypothesis testing*.

Discriminant analysis is the correct procedure to test the null hypothesis:

The group mean vectors of the set of independent (explanatory) variables of two or more *a priori* groups are equal.

For example, let's say we want to test the following hypothesis:

There is no difference between heavy users, light users, and nonusers of drive-in car wash service facilities, as measured by the relevant set of owner demographic and automobile characteristics.

Further, assume that we define *a priori* heavy users as those who use such a facility two or more times a month, light users as those who use such a facility less than two times a month but at least twice a year, and nonusers as those who use such a facility less than two times a year. These three groups are then our dependent (nominal) variable.

Continuing the example, the relevant set of independent variables that we ask each respondent in our sample could be:

1. Age of vehicle owner.
2. Annual personal income of vehicle owner.
3. Age of vehicle.
5. An index of socioeconomic status of the owner.

Submitting the data to DFA for a sufficiently large sample, we could determine whether to reject or not reject the null hypothesis, using the Wilks' lambda statistic. If we reject the null hypothesis, we may determine which of the three groups differ from which other groups by inspecting the matrix of pairwise multivariate $F$-statistics between groups. Both statistics are normally printed out in most DFA programs.

The marketing researcher can think of DFA as the multivariate extension of the one-way analysis of variance between group means. The difference is that a one-way ANOVA looks at only one variable at a time (say, age of vehicle owner), whereas DFA looks at a set of independent variables together (all five in this case) and makes adjustments for the intercorrelations between them.

The second major use of DFA in marketing research is *prediction*.

Discriminant equations derived from an *a priori* set of groups can be used to predict group membership of a subsequent sample drawn from the same population, measured on the same set of variables.

Thus, if the groups formed from a *post hoc* market segmentation study are subjected to discriminant analysis on the original basis variables (i.e., those used in the clustering), the resulting discriminant equations or a set of transforms de-
derived from the discriminant equations, called "classification equations," can be used to classify respondents from subsequent research studies into their respective market segments.

One assumes, of course, that respondents in the subsequent studies respond to the same set of questions that represent the set of independent variables in the discriminant equations. A nice byproduct of this procedure is that DFA will let you know which questions are not useful for predicting group assignments, and those questions can be eliminated in future studies.

Returning to our drive-in car wash example, let's assume that we rejected the null hypothesis and that there appeared to be a high degree of discrimination between groups. Furthermore, let's assume that variable 4, current resale value of the vehicle, did not contribute to the discrimination between the three *a priori* groups, so we need not ask that question in subsequent studies. Now we want to be able to survey another sample from the same population, asking them the four remaining questions, and predict whether they are heavy, light, or nonusers.

By using the linear classification equations generated by most DFA programs, we can make these predictions. There will be one classification equation for each group. We would simply substitute each respondent's answers to the four remaining questions into each of the three derived classification equations and calculate the three scores. The equation yielding the highest score represents the group in which that respondent is most likely to belong.

If we had known beforehand that we were going to use DFA for prediction, we would have used a holdout sample in the initial analysis. That holdout sample would have been used with the classification equations to give us an estimate of how accurate the classification equations were in predicting group memberships of subsequent samples. We will discuss the use of holdout samples in more detail in the next issue.

We also will discuss the use of DFA for structural analysis, for perceptual mapping, and as a possible substitute for regression analysis in certain situations. In addition, we will provide a list of references on DFA.
Copyright of Marketing Research is the property of American Marketing Association and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.