UNIT - IV

Discriminant Analysis

Discriminant Analysis

- Discriminant analysis is the appropriate statistical techniques when the dependent variable is a categorical (nominal or nonmetric) variable and the independent variables are metric variables.
- In many cases, the dependent variable consists of two groups or classifications, for example, male versus female or high versus low.
- In other instances, more than two groups are involved, such as low, medium, and high classifications.
- Discriminant analysis is capable of handling either two groups or multiple (three or more) groups.
- When two classifications are involved, the technique is referred to as two-group discriminant analysis.
- When three or more classifications are identified, the technique is referred to as multiple discriminant analysis (MDA).
- Logistic regression is limited in its basic form to two groups, although other formulations can handle more groups.

Discriminant Analysis

- **Discriminant Analysis**
- Discriminant analysis involves deriving a variate.
- The discriminant variate is the linear combination of the two (or more) independent variables that will discriminate best between the objects (persons, firms, etc.) in the groups defined a priori.
- Discrimination is achieved by calculating the variate's weights for each independent variable to maximize the differences between the groups (i.e., the between-group variance relative to the within-group variance).
- The variate for a discriminant analysis, also known as the discriminant function, is derived from an equation much like that seen in multiple regression.
- It takes the following form:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

where

Z_{jk} = discriminant Z score of discriminant function j for object k
a = intercept
W_i = discriminant weight for independent variable i
X_{ik} = independent variable i for object k

Discriminant Analysis

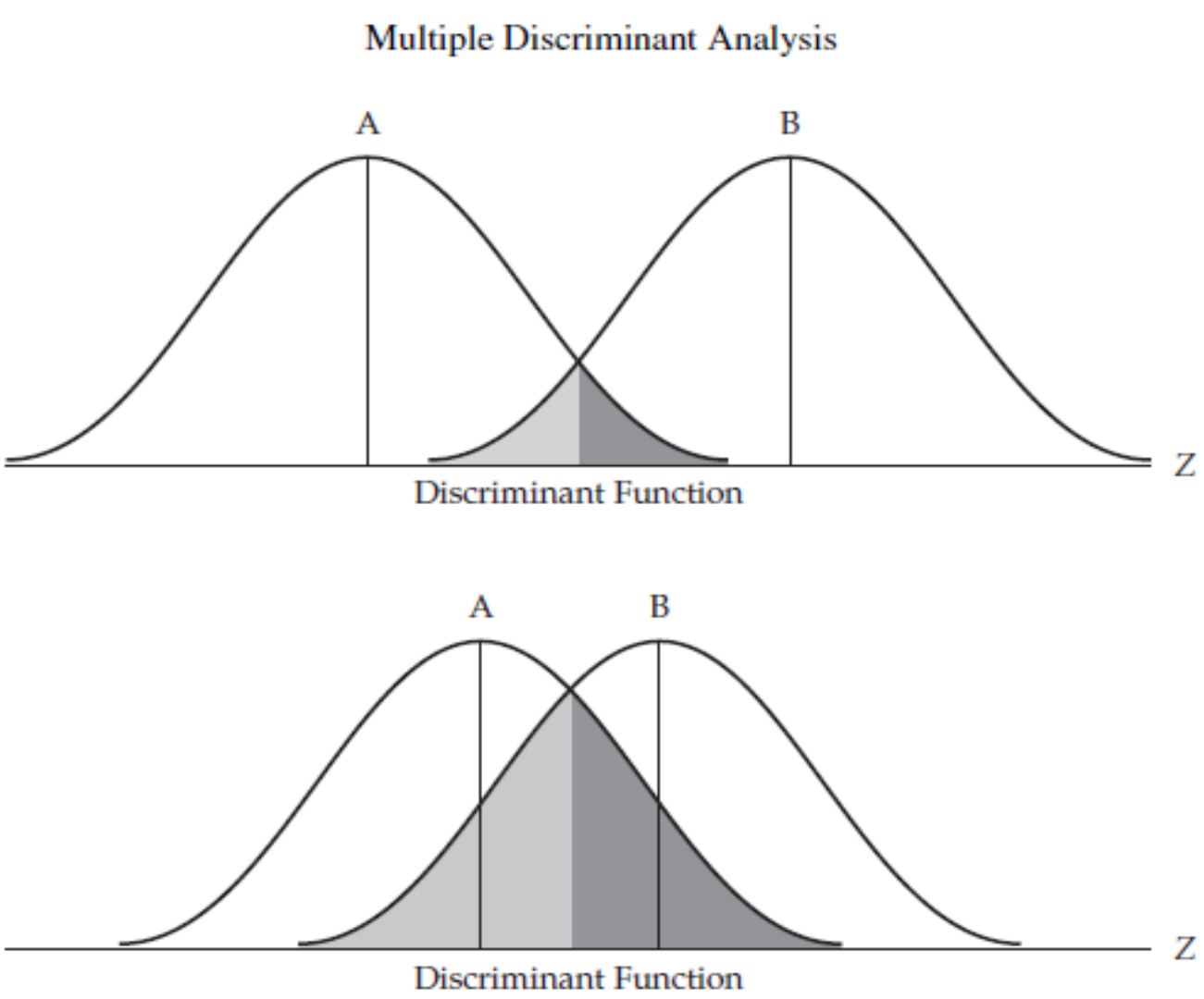
- **Discriminant Analysis**
- As with the variate in regression or any other multivariate technique we see the discriminant score for each object in the analysis (person, firm, etc.) being a summation of the values obtained by multiplying each independent variable by its discriminant weight.
- What is unique about discriminant analysis is that more than one discriminant function may be present, resulting in each object possibly having more than one discriminant score.
- Discriminant analysis is the appropriate statistical technique for testing the hypothesis that the group means of a set of independent variables for two or more groups are equal.
- By averaging the discriminant scores for all the individuals within a particular group, we arrive at the group mean.
- This group mean is referred to as a centroid.
- When the analysis involves two groups, there are two centroids; with three groups, there are three centroids; and so forth.

Discriminant Analysis

- **Discriminant Analysis**
- The centroids indicate the most typical location of any member from a particular group, and a comparison of the group centroids shows how far apart the groups are in terms of that discriminant function.
- The test for the statistical significance of the discriminant function is a generalized measure of the distance between the group centroids.
- It is computed by comparing the distributions of the discriminant scores for the groups.
- Multiple discriminant analysis is unique in one characteristic among the dependence relationships.
- If the dependent variable consists of more than two groups, discriminant analysis will calculate more than one discriminant function.
- As a matter of fact, it will calculate NG – 1 functions, where NG is the number of groups.
- Each discriminant function will calculate a separate discriminant Z score.
- In the case of a three-group dependent variable, each object (respondent, firm, etc.) will have a separate score for discriminant functions one and two, which enables the objects to be plotted in two dimensions, with each dimension representing a discriminant function.
- Thus, discriminant analysis is not limited to a single variate, as is multiple regression, but creates multiple variates representing dimensions of discrimination among the groups.

Discriminant Analysis

- **Discriminant Analysis**



Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- Discriminant analysis is applicable to any research question with the objective of understanding group membership, whether the groups comprise individuals (e.g., customers versus noncustomers), firms (e.g., profitable versus unprofitable), products (e.g., successful versus unsuccessful), or any other object that can be evaluated on a series of independent variables.
- **Two-Group Discriminant Analysis: Purchasers Versus Non-purchasers**
- Suppose KitchenAid wants to determine whether one of its new products—a new and improved food mixer—will be commercially successful. In carrying out the investigation, KitchenAid is primarily interested in identifying those consumers who would purchase the new product versus those who would not.
- To assist in identifying potential purchasers, KitchenAid devised rating scales on three characteristics—durability, performance, and style—to be used by consumers in evaluating the new product.
- Rather than relying on each scale as a separate measure, KitchenAid hopes that a weighted combination of all three would better predict purchase likelihood of consumers.

Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **IDENTIFYING DISCRIMINATING VARIABLES**
- To identify variables that may be useful in discriminating between groups (i.e., purchasers versus nonpurchasers), emphasis is placed on group differences rather than measures of correlation used in multiple regression.

TABLE 1 KitchenAid Survey Results for the Evaluation of a New Consumer Product

Groups Based on Purchase Intention	*Evaluation of New Product**		
	X_1 Durability	X_2 Performance	X_3 Style
Group 1: Would purchase			
Subject 1	8	9	6
Subject 2	6	7	5
Subject 3	10	6	3
Subject 4	9	4	4
Subject 5	4	8	2
Group mean	7.4	6.8	4.0
Group 2: Would not purchase			
Subject 6	5	4	7
Subject 7	3	7	2
Subject 8	4	5	5
Subject 9	2	4	3
Subject 10	2	2	2
Group mean	3.2	4.4	3.8
Difference between group means	4.2	2.4	0.2

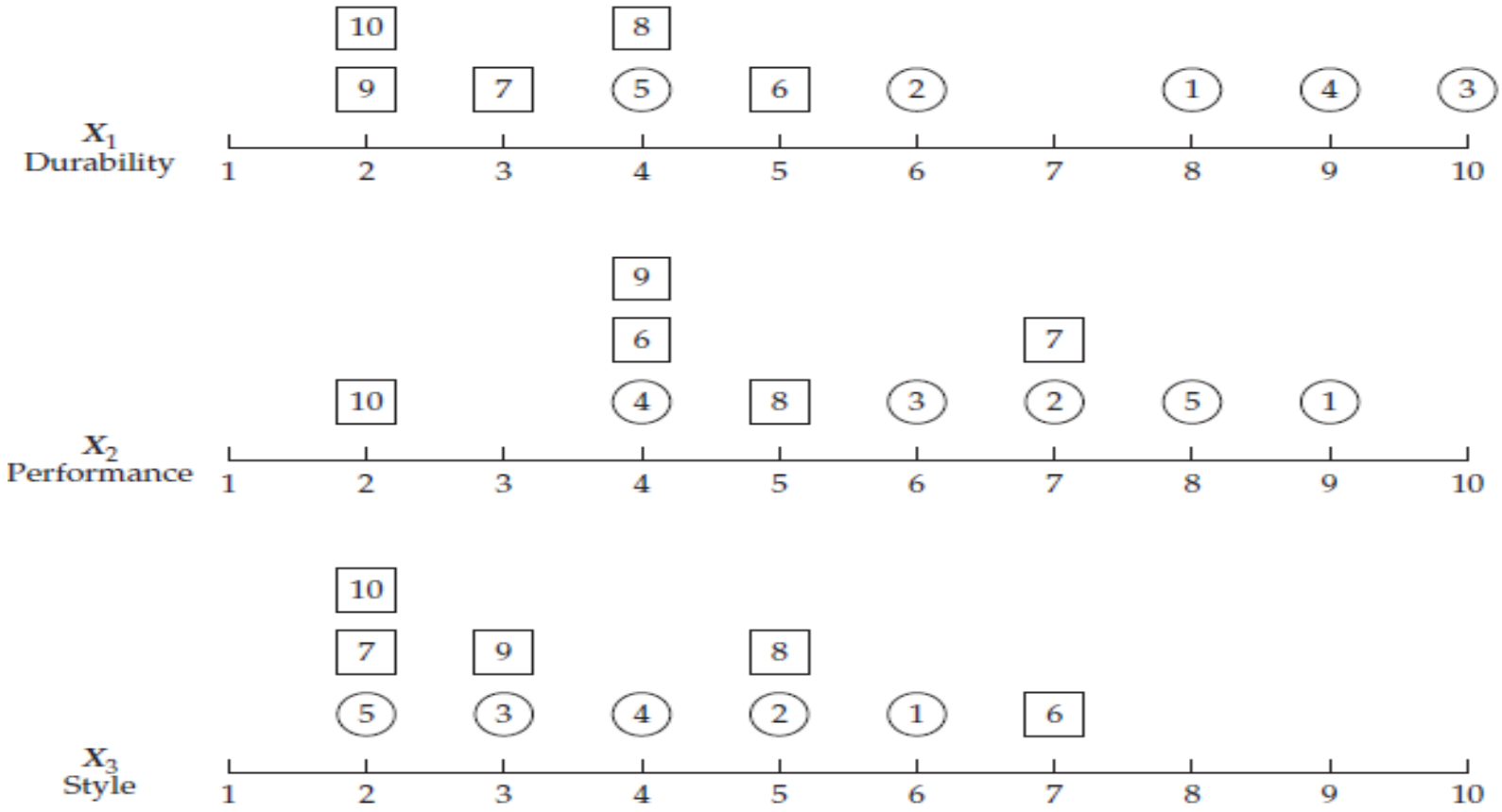
*Evaluations are made on a 10-point scale (1 = very poor to 10 = excellent).

Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **IDENTIFYING DISCRIMINATING VARIABLES**
- Because we have only 10 respondents in two groups and three independent variables, we can also look at the data graphically to determine what discriminant analysis is trying to accomplish.
- Figure 2 shows the 10 respondents on each of the three variables. The “would purchase” group is represented by circles and the “would not purchase” group by the squares. Respondent identification numbers are inside the shapes.
- • *X*1 (Durability) had a substantial difference in mean scores, enabling us to almost perfectly discriminate between the groups using only this variable. If we established the value of 5.5 as our cutoff point to discriminate between the two groups, then we would misclassify only respondent 5, one of the “would purchase” group members. This variable illustrates the discriminatory power in having a large difference in the means for the two groups and a lack of overlap between the distributions of the two groups.
- • *X*2 (Performance) provides a less clear-cut distinction between the two groups. However, this variable does provide high discrimination for respondent 5, who was misclassified if we used only *X*1. In addition, the respondents who would be misclassified using *X*2 are well separated on *X*1. Thus, *X*1 and *X*2 might be used quite effectively in combination to predict group membership.

Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **IDENTIFYING DISCRIMINATING VARIABLES**
- • *X*3 (Style) shows little differentiation between the groups. Thus, by forming a variate of only *X*1 and *X*2, and omitting *X*3, a discriminant function may be formed that maximizes the separation of the groups on the discriminant score.



Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **CALCULATING A DISCRIMINANT FUNCTION**
- With the three potential discriminating variables identified, attention shifts toward investigation of the possibility of using the discriminating variables in combination to improve upon the discriminating power of any individual variable.
- To this end, a variate can be formed with two or more discriminating variables to act together in discriminating between the groups.
- Table 2 contains the results for three different formulations of a discriminant function, each representing different combinations of the three independent variables.

Discriminant Analysis

- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **CALCULATING A DISCRIMINANT FUNCTION**

TABLE 2 Creating Discriminant Functions to Predict Purchasers Versus Nonpurchasers

	Calculated Discriminant Z Scores		
Group	**Function 1: $Z = X_1$**	**Function 2: $Z = X_1 + X_2$**	**Function 3: $Z = -4.53 + .476X_1 + .359X_2$**
Group 1: Would purchase			
Subject 1	8	17	2.51
Subject 2	6	13	.84
Subject 3	10	16	2.38
Subject 4	9	13	1.19
Subject 5	4	12	.25
Group 2: Would not purchase			
Subject 6	5	9	−.71
Subject 7	3	10	−.59
Subject 8	4	9	−.83
Subject 9	2	6	−2.14
Subject 10	2	4	−2.86
Cutting score	5.5	11	0.0

Classification Accuracy:

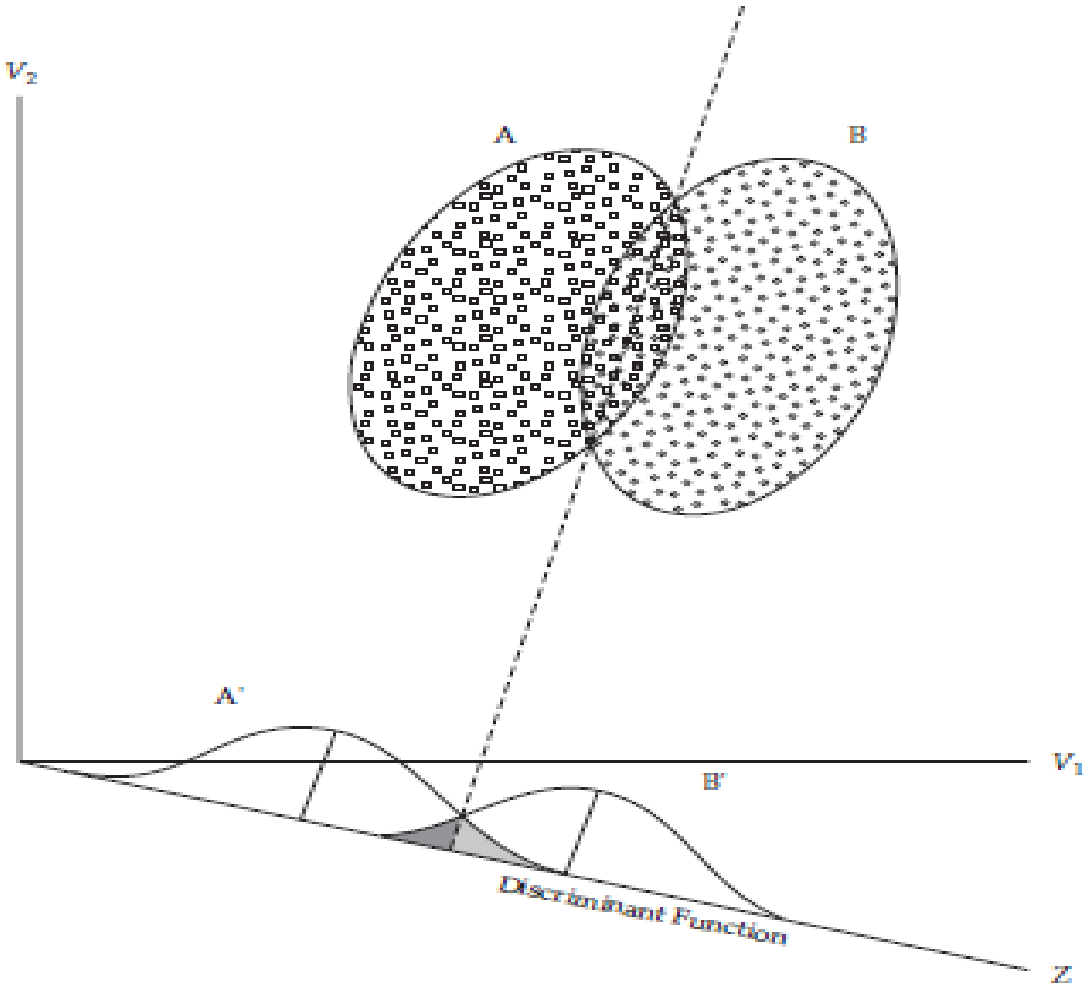
	Predicted Group		Predicted Group		Predicted Group	
Actual Group	1	2	1	2	1	2
1: Would purchase	4	1	5	0	5	0
2: Would not purchase	0	5	0	5	0	5

Discriminant Analysis

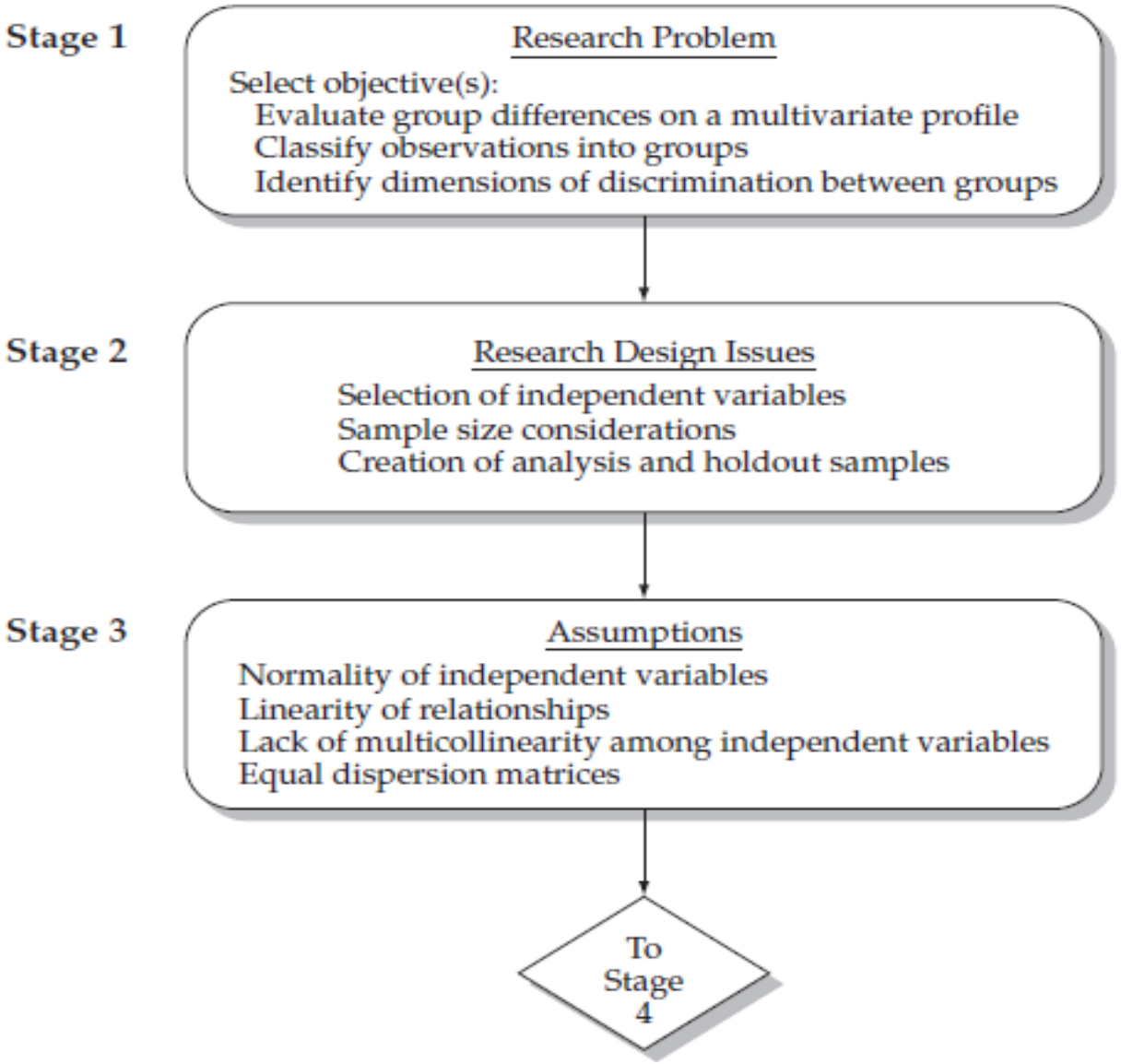
- **HYPOTHETICAL EXAMPLE OF DISCRIMINANT ANALYSIS**
- **CALCULATING A DISCRIMINANT FUNCTION**
- The first discriminant function contains just *X*1, equating the value of *X*1 to the discriminant *Z* score (also implying a weight of 1.0 for *X*1 and weights of zero for all other variables).
- As shown earlier, use of only *X*1, the best discriminator, results in the misclassification of subject 5 as shown in Table 2, where four out of five subjects in group 1 (all but subject 5) and five of five subjects in group 2 are correctly classified (i.e., lie on the diagonal of the classification matrix).
- The percentage correctly classified is thus 90 percent (9 out of 10 subjects).
- • Because *X*2 provides discrimination for subject 5, we can form a second discriminant function by equally combining *X*1 and *X*2 (i.e., implying weights of 1.0 for *X*1 and *X*2 and a weight of 0.0for *X*3) to utilize each variable's unique discriminatory powers. Using a cutting score of 11 with this new discriminant function (see Table 2) achieves a perfect classification of the two groups (100% correctly classified).
- Thus, *X*1 and *X*2 in combination are able to make better predictions of group membership than either variable separately.
- • The third discriminant function in Table 2 represents the actual estimated discriminant function ($Z = -4.53 + .476X1 + .359X2$). Based on a cutting score of 0, this third function also achieves a 100 percent correct classification rate with the maximum separation possible between groups.

Discriminant Analysis

- **A Geometric Representation of the Two-Group Discriminant Function**
- A graphical illustration of another two-group analysis will help to further explain the nature of discriminant analysis.



THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS



THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 1: OBJECTIVES OF DISCRIMINANT ANALYSIS**
- A review of the objectives for applying discriminant analysis should further clarify its nature.
- Discriminant analysis can address any of the following research objectives:
- **1.** Determining whether statistically significant differences exist between the average score profiles on a set of variables for two (or more) *a priori* defined groups
- **2.** Determining which of the independent variables most account for the differences in the average score profiles of the two or more groups
- **3.** Establishing the number and composition of the dimensions of discrimination between groups formed from the set of independent variables
- **4.** Establishing procedures for classifying objects (individuals, firms, products, etc.) into groups on the basis of their scores on a set of independent variables
- Discriminant analysis, therefore, can be considered either a type of profile analysis or an analytical predictive technique. In either case, the technique is most appropriate in situations with a single categorical dependent variable and several metrically scaled independent variables.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 1: OBJECTIVES OF DISCRIMINANT ANALYSIS**
- As a ***profile analysis,*** discriminant analysis provides an objective assessment of differences between groups on a set of independent variables. In this situation, discriminant analysis is quite similar to multivariate analysis of variance. For understanding group differences, discriminant analysis lends insight into the role of individual variables as well as defining combinations of these variables that represent dimensions of discrimination between groups.
- These dimensions are the collective effects of several variables that work jointly to distinguish between the groups. The use of sequential estimation methods also allows for identifying subsets of variables with the greatest discriminatory power.
- • For ***classification purposes,*** discriminant analysis provides a basis for classifying not only the sample used to estimate the discriminant function but also any other observations that can have values for all the independent variables. In this way, the discriminant analysis can be used to classify other observations into the defined groups.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- Successful application of discriminant analysis requires consideration of several issues. These issues include the selection of both dependent and independent variables, the sample size needed for estimation of the discriminant functions, and the division of the sample for validation purposes.
- **Selecting Dependent and Independent Variables**
- To apply discriminant analysis, the researcher first must specify which variables are to be independent measures and which variable is to be the dependent measure.
- **THE DEPENDENT VARIABLE**
- The researcher should focus on the dependent variable first.
- The number of dependent variable groups (categories) can be two or more, but these groups must be mutually exclusive and exhaustive.
- In other words, each observation can be placed into only one group. In some cases, the dependent variable may involve two groups (dichotomous), such as good versus bad.
- In other cases, the dependent variable may involve several groups (multichotomous), such as the occupations of physician, attorney, or professor.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- Theoretically, discriminant analysis can handle an unlimited number of categories in the dependent variable.
- In addition to being mutually exclusive and exhaustive, the dependent variable categories should be distinct and unique on the set of independent variables chosen.
- Discriminant analysis assumes that each group *should* have a unique profile on the independent variables used and thus develops the discriminant functions to maximally separate the groups based on these variables.
- Discriminant analysis does not, however, have a means of accommodating or combining categories that are not distinct on the independent variables.
- If two or more groups have quite similar profiles, discriminant analysis will not be able to uniquely profile each group, resulting in poorer explanation and classification of the groups as a whole.
- As such, the researcher must select the dependent variables and its categories to reflect differences in the independent variables.
- An example will help illustrate this issue.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- Assume the researcher wishes to identify differences among occupational categories based on a number of demographic characteristics (e.g., income, education, household characteristics).
- If occupations are represented by a small number of categories (e.g., blue-collar, white-collar, clerical/staff, and professional/upper management), then we would expect unique differences between the groups and that discriminant analysis would be best able to develop discriminant functions that would explain the group differences and successfully classify individuals into their correct category.
- The researcher should also strive, all other things equal, for a smaller rather than larger number of categories in the dependent measure. It may seem more logical to expand the number of categories in search of more unique groupings, but expanding the number of categories represents more complexities in the profiling and classification tasks of discriminant analysis.
- If discriminant analysis can estimate up to $NG - 1$ (number of groups minus one) discriminant functions, then increasing the number of groups expands the number of possible discriminant functions, increasing the complexity in identifying the underlying dimensions of discrimination reflected by each discriminant function as well as representing the overall effect of each independent variable.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- **Converting Metric Variables.**
- In some situations, however, discriminant analysis is appropriate even if the dependent variable is not a true nonmetric (categorical) variable. We may have a dependent variable that is an ordinal or interval measurement that we wish to use as a categorical dependent variable. In such cases, we would have to create a categorical variable, and two approaches are the most commonly used:
- The most common approach is to establish categories using the metric scale.
- For example, if we had a variable that measured the average number of cola drinks consumed per day, and the individuals responded on a scale from zero to eight or more per day, we could create an artificial trichotomy (three groups) by simply designating those individuals who consumed none, one, or two cola drinks per day as light users, those who consumed three, four, or five per day as medium users, and those who consumed six, seven, eight, or more as heavy users.
- Such a procedure would create a three-group categorical variable in which the objective would be to discriminate among light, medium, and heavy users of colas.
- Any number of categorical groups can be developed. Most frequently, the approach would involve creating two, three, or four categories.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- **Converting Metric Variables.**
- When three or more categories are created, the possibility arises of examining only the extreme groups in a two-group discriminant analysis. The **polar extremes approach** involves comparing only the extreme two groups and excluding the middle group from the discriminant analysis.
- For example, the researcher could examine the light and heavy users of cola drinks and exclude the medium users. This approach can be used any time the researcher wishes to examine only the extreme groups.
- **THE INDEPENDENT VARIABLES**
- After a decision has been made on the dependent variable, the researcher must decide which independent variables to include in the analysis.
- Independent variables usually are selected in two ways.
- The first approach involves identifying variables either from previous research or from the theoretical model that is the underlying basis of the research question.
- The second approach is intuition—utilizing the researcher's knowledge and intuitively selecting variables for which no previous research or theory exists but that logically might be related to predicting the groups for the dependent variable.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- In both instances, the most appropriate independent variables are those that differ across at least two of the groups of the dependent variable.
- Remember that the purpose of any independent variable is to present a unique profile of at least one group as compared to others.
- Variables that do not differ across the groups are of little use in discriminant analysis.
- **Sample Size**
- Discriminant analysis, like the other multivariate techniques, is affected by the size of the sample being analyzed.
- Very small samples have so much sampling error that identification of all but the largest differences is improbable.
- Moreover, very large sample sizes will make all differences statistically significant, even though these same differences may have little or no managerial relevance.
- In between these extremes, the researcher must consider the impact of sample sizes on discriminant analysis, both at the overall level and on a group-by-group basis.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- **OVERALL SAMPLE SIZE**
- The first consideration involves the overall sample size.
- Discriminant analysis is quite sensitive to the ratio of sample size to the number of predictor variables.
- As a result, many studies suggest a ratio of 20 observations for each predictor variable.
- Although this ratio may be difficult to maintain in practice, the researcher must note that the results become unstable as the sample size decreases relative to the number of independent variables.
- The minimum size recommended is five observations per independent variable.
- **SAMPLE SIZE PER CATEGORY**
- In addition to the overall sample size, the researcher also must consider the sample size of each category.
- At a minimum, the smallest group size of a category must exceed the number of independent variables. As a practical guideline, each category should have at least 20 observations.
- Even when all categories exceed 20 observations, however, the researcher must also consider the relative sizes of the categories.
- Wide variations in the groups' size will impact the estimation of the discriminant function and the classification of observations.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- **Division of the Sample**
- The preferred means of validating a discriminant analysis is to divide the sample into two subsamples, one used for estimation of the discriminant function and another for validation purposes.
- In terms of sample size considerations, it is essential that each subsample be of adequate size to support conclusions from the results.
- As such, all of the considerations discussed in the previous section apply not only to the total sample, but also to each of the two subsamples.
- No hard-and-fast rules have been established, but it seems logical that the researcher would want at least 100 in the total sample to justify dividing it into the two groups.
- **CREATING THE SUBSAMPLES**
- A number of procedures have been suggested for dividing the sample into subsamples.
- The usual procedure is to divide the total sample of respondents randomly into two subsamples.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 2: RESEARCH DESIGN FOR DISCRIMINANT ANALYSIS**
- One of these subsamples, the **analysis sample**, is used to develop the discriminant function.
- The second, the **holdout sample**, is used to test the discriminant function.
- This method of validating the function is referred to as the **split-sample validation** or **cross-validation.**
- No definite guidelines have been established for determining the relative sizes of the analysis and holdout (or validation) subsamples.
- The most popular approach is to divide the total sample so that one-half of the respondents are placed in the analysis sample and the other half are placed in the holdout sample.
- However, no hard-and-fast rule has been established, and some researchers prefer a 60–40 or even 75–25 split between the analysis and the holdout groups, depending on the overall sample size.
- If the original groups are unequal, the sizes of the estimation and holdout samples should be proportionate to the total sample distribution.
- For instance, if a sample consists of 50 males and 50 females, the estimation and holdout samples would have 25 males and 25 females.
- If the sample contained 70 females and 30 males, then the estimation and holdout samples would consist of 35 females and 15 males each.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS**
- As with all multivariate techniques, discriminant analysis is based on a number of assumptions.
- These assumptions relate to both the statistical processes involved in the estimation and classification procedures and issues affecting the interpretation of the results.
- **Impacts on Estimation and Classification**
- The key assumptions for deriving the discriminant function are multivariate normality of the independent variables and unknown (but equal) dispersion and covariance structures (matrices) for the groups as defined by the dependent variable.
- Although the evidence is mixed regarding the sensitivity of discriminant analysis to violations of these assumptions, the researcher must always understand the impacts on the results that can be expected.
- Moreover, if the assumptions are violated and the potential remedies are not acceptable or do not address the severity of the problem, the researcher should consider alternative methods (e.g., logistic regression).

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS**
- **IDENTIFYING ASSUMPTION VIOLATIONS**
- Achieving univariate normality of individual variables will many times suffice to achieve multivariate normality.
- A number of tests for normality are available to the researcher, along with the appropriate remedies, those most often being transformations of the variables.
- The issue of equal dispersion of the independent variables (i.e., equivalent covariance matrices) is similar to homoscedasticity between individual variables.
- The most common test is the **Box's M** test assessing the significance of differences in the matrices between the groups.
- Here the researcher is looking for a *non-significant* probability level which would indicate that there were not differences between the group covariance matrices.
- We should use very conservative levels of significant differences (e.g., .01 rather than .05) when assessing whether differences are present.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS**
- **IMPACT ON ESTIMATION**
- Data not meeting the multivariate normality assumption can cause problems in the estimation of the discriminant function.
- Remedies may be possible through transformations of the data to reduce the disparities among the covariance matrices.
- However, in many instances these remedies are ineffectual. In these situations, the models should be thoroughly validated.
- If the dependent measure is binary, logistic regression should be used if at all possible.
- **IMPACT ON CLASSIFICATION**
- Unequal covariance matrices also negatively affect the classification process.
- If the sample sizes are small and the covariance matrices are unequal, then the statistical significance of the estimation process is adversely affected.
- The more likely case is that of unequal covariances among groups of adequate sample size, whereby observations are overclassified into the groups with larger covariance matrices.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

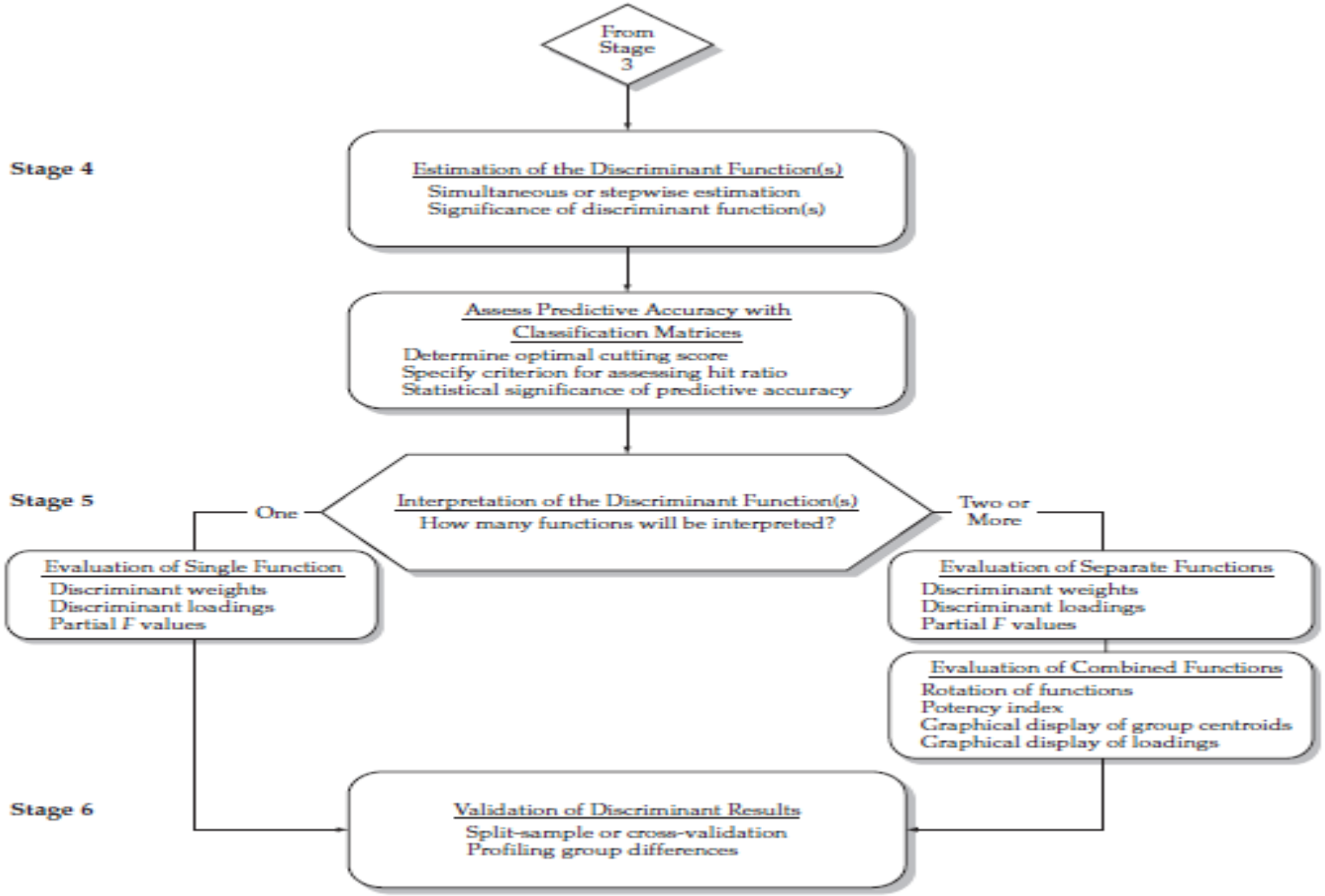
- **STAGE 3: ASSUMPTIONS OF DISCRIMINANT ANALYSIS**
- This effect can be minimized by increasing the sample size and also by using the group-specific covariance matrices for classification purposes, but this approach mandates cross-validation of the discriminant results.
- Finally, quadratic classification techniques are available in many of the statistical programs if large differences exist between the covariance matrices of the groups and the remedies do not minimize the effect.
- **Impacts on Interpretation**
- Another characteristic of the data that affects the results is multicollinearity among the independent variables.
- Multicollinearity, measured in terms of **tolerance**, denotes that two or more independent variables are highly correlated, so that one variable can be highly explained or predicted by the other variable(s) and thus it adds little to the explanatory power of the entire set.
- As with any of the multivariate techniques employing a variate, an implicit assumption is that all relationships are linear.
- Nonlinear relationships are not reflected in the discriminant function unless specific variable transformations are made to represent nonlinear effects.
- Finally, outliers can have a substantial impact on the classification accuracy of any discriminant analysis results.
- The researcher is encouraged to examine all results for the presence of outliers and to eliminate true outliers if needed.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- To derive the discriminant function, the researcher must decide on the method of estimation and then determine the number of functions to be retained.
- With the functions estimated, overall model fit can be assessed in several ways.
- First, **discriminant *Z* scores**, also known as the ***Z* scores**, can be calculated for each object.
- Comparison of the group means (centroids) on the *Z* scores provides one measure of discrimination between groups.
- Predictive accuracy can be measured as the number of observations classified into the correct groups, with a number of criteria available to assess whether the classification process achieves practical or statistical significance.
- Finally, casewise diagnostics can identify the classification accuracy of each case and its relative impact on the overall model estimation.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**



THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Selecting an Estimation Method**
- The first task in deriving the discriminant function(s) is to choose the estimation method.
- The two methods available are the simultaneous (direct) method and the stepwise method, each discussed next.
- **SIMULTANEOUS ESTIMATION**
- **Simultaneous estimation** involves computing the discriminant function so that all of the independent variables are considered concurrently.
- Thus, the discriminant function is computed based upon the entire set of independent variables, regardless of the discriminating power of each independent variable.
- The simultaneous method is appropriate when, for theoretical reasons, the researcher wants to include all the independent variables in the analysis and is not interested in seeing intermediate results based only on the most discriminating variables.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **STEPWISE ESTIMATION**
- **Stepwise estimation** is an alternative to the simultaneous approach.
- It involves entering the independent variables into the discriminant function one at a time on the basis of their discriminating power.
- The stepwise approach follows a sequential process of adding or deleting variables in the following manner:
- **1.** Choose the single best discriminating variable.
- **2.** Pair the initial variable with each of the other independent variables, one at a time, and select the variable that is best able to improve the discriminating power of the function in combination with the first variable.
- **3.** Select additional variables in a like manner. Note that as additional variables are included, some previously selected variables may be removed if the information they contain about group differences is available in some combination of the other variables included at later stages.
- **4.** Consider the process completed when either all independent variables are included in the function or the excluded variables are judged as not contributing significantly to further discrimination.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Statistical Significance**
- After estimation of the discriminant function(s), the researcher must assess the level of significance for the collective discriminatory power of the discriminant function(s) as well as the significance of each separate discriminant function.
- Evaluating the overall significance provides the researcher with the information necessary to decide whether to continue on to the interpretation of the analysis or if respecification is necessary.
- If the overall model is significant, then evaluating the individual functions identifies the function(s) that should be retained and interpreted.
- **OVERALL SIGNIFICANCE**
- In assessing the statistical significance of the overall model, different statistical criteria are applicable for simultaneous versus stepwise estimation procedures.
- In both situations, the statistical tests relate to the ability of the discriminant function(s) to derive discriminant *Z* scores that are significantly different between the groups.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Simultaneous Estimation.**
- When a simultaneous approach is used, the measures of Wilks' lambda, Hotelling's trace, and Pillai's criterion all evaluate the statistical significance of the discriminatory power of the discriminant function(s).
- Roy's greatest characteristic root evaluates only the first discriminant function.
- **Stepwise Estimation.**
- If a stepwise method is used to estimate the discriminant function, the Mahalanobis *D*2 and Rao's *V* measures are most appropriate. Both are measures of generalized distance.
- The Mahalanobis *D*2 procedure is based on generalized squared Euclidean distance that adjusts for unequal variances.
- The major advantage of this procedure is that it is computed in the original space of the predictor variables rather than as a collapsed version used in other measures.
- The Mahalanobis *D*2 procedure becomes particularly critical as the number of predictor variables increases, because it does not result in any reduction in dimensionality.
- A loss in dimensionality would cause a loss of information because it decreases variability of the independent variables.
- In general, Mahalanobis *D*2 is the preferred procedure when the researcher is interested in the maximal use of available information in a stepwise process.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **SIGNIFICANCE OF INDIVIDUAL DISCRIMINANT FUNCTIONS**
- If the number of groups is three or more, then the researcher must decide not only whether the discrimination between groups overall is statistically significant but also whether each of the estimated discriminant functions is statistically significant.
- As discussed earlier, discriminant analysis estimates one less discriminant function than there are groups.
- If three groups are analyzed, then two discriminant functions will be estimated; for four groups, three functions will be estimated; and so on.
- The conventional significance criterion of .05 or beyond is often used, yet some researchers extend the required significance level (e.g., .10 or more) based on the trade-off of cost versus the value of the information.
- If the higher levels of risk for including nonsignificant results (e.g., significance levels > .05) are acceptable, discriminant functions may be retained that are significant at the .2 or even the .3 level.
- If one or more functions are deemed not statistically significant, the discriminant model should be reestimated with the number of functions to be derived limited to the number of significant functions.
- In this manner, the assessment of predictive accuracy and the interpretation of the discriminant functions will be based only on significant functions.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Assessing Overall Model Fit**
- Once the significant discriminant functions have been identified, attention shifts to ascertaining the overall fit of the retained discriminant function(s).
- This assessment involves three tasks:
- **1.** Calculating discriminant *Z* scores for each observation
- **2.** Evaluating group differences on the discriminant *Z* scores
- **3.** Assessing group membership prediction accuracy
- The discriminant *Z* score is calculated for each discriminant function for every observation in the sample.
- The discriminant score acts as a concise and simple representation of each discriminant function, simplifying the interpretation process and the assessment of the contribution of independent variables.
- Groups can be distinguished by their discriminant scores and, as we will see, the discriminant scores can play an instrumental role in predicting group membership.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Assessing Overall Model Fit**
- **CALCULATING DISCRIMINANT *Z* SCORES**
- With the retained discriminant functions defined, the basis for calculating the discriminant *Z* scores has been established.
- As discussed earlier, the discriminant *Z* score of any discriminant function can be calculated for each observation by the following formula:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

where

Z_{jk} = discriminant Z score of discriminant function j for object k
a = intercept
W_i = discriminant weight for independent variable i
X_{ik} = independent variable i for object k

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- The discriminant *Z* score, a metric variable, provides a direct means of comparing observations on each function.
- Observations with similar *Z* scores are assumed more alike on the variables constituting this function than those with disparate scores.
- The discriminant function can be expressed with either standardized or unstandardized weights and values.
- The standardized version is more useful for interpretation purposes, but the unstandardized version is easier to use in calculating the discriminant *Z* score.
- **EVALUATING GROUP DIFFERENCES**
- Once the discriminant *Z* scores are calculated, the first assessment of overall model fit is to determine the magnitude of differences between the members of each group in terms of the discriminant *Z* scores.
- A summary measure of the group differences is a comparison of the group centroids, the average discriminant *Z* score for all group members.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- A measure of success of discriminant analysis is its ability to define discriminant function(s) that result in significantly different group centroids.
- The differences between centroids are measured in terms of Mahalanobis *D*2 measure, for which tests are available to determine whether the differences are statistically significant.
- The researcher should ensure that even with significant discriminant functions, significant differences occur between each of the groups.
- Group centroids on each discriminant function can also be plotted to demonstrate the results from a graphical perspective.
- Plots are usually prepared for the first two or three discriminant functions (assuming they are statistically significant functions).
- The values for each group show its position in reduced discriminant space. The researcher can see the differences between the groups on each function;

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **ASSESSING GROUP MEMBERSHIP PREDICTION ACCURACY**
- Given that the dependent variable is nonmetric, it is not possible to use a measure such as *R*2, as is done in multiple regression, to assess predictive accuracy.
- Rather, each observation must be assessed as to whether it was correctly classified. In doing so, several major considerations must be addressed:
- • The statistical and practical rationale for developing classification matrices
- • Classifying individual cases
- • Construction of the classification matrices
- • Standards for assessing classification accuracy

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Why Classification Matrices Are Developed.**
- The statistical tests for assessing the significance of the discriminant function(s) only assess the degree of difference between the groups based on the discriminant *Z* scores, but do not indicate how well the function(s) predicts.
- These statistical tests suffer the same drawbacks as the classical tests of hypotheses.
- For example, suppose the two groups are deemed significantly different beyond the .01 level. Yet with sufficiently large sample sizes, the group means (centroids) could be virtually identical and still have statistical significance.
- To determine the predictive ability of a discriminant function, the researcher must construct classification matrices.
- The **classification matrix** procedure provides a perspective on practical significance rather than statistical significance.
- With multiple discriminant analysis, the **percentage correctly classified**, also termed the **hit ratio**, reveals how well the discriminant function classified the objects.
- With a sufficiently large sample size in discriminant analysis, we could have a statistically significant difference between the two (or more) groups and yet correctly classify only 53 percent (when chance is 50%, with equal group sizes) .
- In such instances, the statistical test would indicate statistical significance, yet the hit ratio would allow for a separate judgment to be made in terms of practical significance.
- Thus, we must use the classification matrix procedure to assess predictive accuracy beyond just statistical significance.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Classifying Individual Observations.**
- The development of classification matrices requires that each observation be classified into one of the groups of the dependent variable based on the discriminant function(s).
- The objective is to characterize each observation on the discriminant function(s) and then determine the extent to which observation in each group can be consistently described by the discriminant functions.
- There are two approaches to classifying observations, one employing the discriminant scores directly and another developing a specific function for classification.
- ***Cutting Score Calculation***
- Using the discriminant functions deemed significant, we can develop classification matrices by calculating the **cutting score** (also called the *critical Z value*) for each discriminant function.
- The cutting score is the criterion against which each object's discriminant score is compared to determine into which group the object should be classified.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- The cutting score represents the dividing point used to classify observations into groups based on their discriminant function score.
- The calculation of a cutting score between any two groups is based on the two group centroids (group mean of the discriminant scores) and the relative size of the two groups.
- The group centroids are easily calculated and provided at each stage of the stepwise process.
- ***Developing a Classification Function***
- As noted earlier, using the discriminant function is only one of two possible approaches to classification.
- The second approach employs a **classification function**, also known as **Fisher's linear discriminant function**.
- The classification functions, one for each group, are used strictly for classifying observations.
- In this method of classification, an observation's values for the independent variables are inserted in the classification functions and a classification score for each group is calculated for that observation.
- The observation is then classified into the group with the highest classification score.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Defining Prior Probabilities.**
- The impact and importance of each group's sample size in the classification process is many times overlooked, yet is critical in making the appropriate assumptions in the classification process.
- Here we are concerned about the representativeness of the sample as it relates to representation of the relative sizes of the groups in the actual in the actual population, which can be stated as prior probabilities (i.e., the relative proportion of each group to the total sample).
- If, however, the sample was conducted randomly and the researcher feels that the group sizes are representative of the population, then the researcher can specify prior probabilities to be based on the estimation sample.
- Thus, the actual group sizes are assumed representative and used directly in the calculation of the cutting score.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- For example, consider a holdout sample consisting of 200 observations, with group sizes of 60 and 140 that relate to prior probabilities of 30 percent and 70 percent, respectively.
- If the sample is assumed representative, then the sample sizes of 60 and 140 are used in calculating the cutting score. If, however, the sample is deemed not representative, the researcher must specify the prior probabilities.
- If they are specified as equal (50% and 50%), sample sizes of 100 and 100 would be used in the cutting score calculation rather than the actual sample sizes.
- Specifying other values for the prior probabilities would result in differing sample sizes for the two groups.
- ***Calculating the Optimal Cutting Score***
- The importance of the prior probabilities can be illustrated in the calculation of the "optimal" cutting score, which takes into account the prior probabilities through the use of group sizes.
- The basic formula for computing the **optimal cutting score** between any two groups is:

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**

$$Z_{CS} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$$

where

Z_{CS} = optimal cutting score between groups A and B
N_A = number of observations in group A
N_B = number of observations in group B
Z_A = centroid for group A
Z_B = centroid for group B

- With unequal group sizes, the optimal cutting score for a discriminant function is now the weighted average of the group centroids.
- The cutting score is weighted toward the smaller group, hopefully making for a better classification of the larger group.
- If the groups are specified to be of equal size (prior probabilities defined as equal), then the optimum cutting score will be halfway between the two group centroids and becomes simply the average of the two centroids:

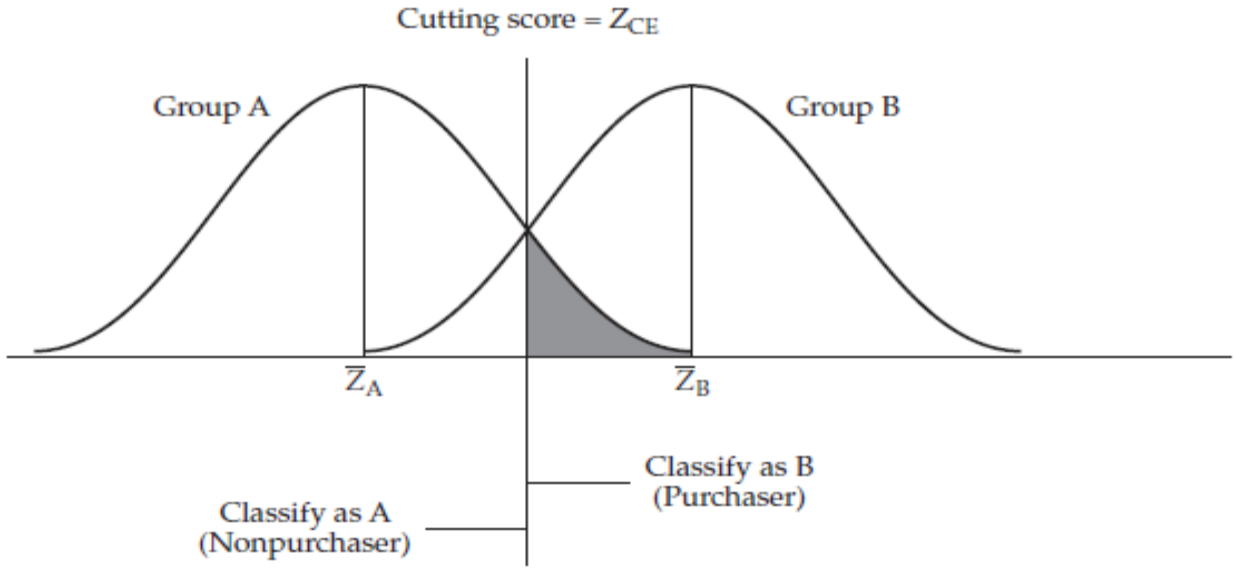
THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**

$$Z_{CE} = \frac{Z_A + Z_B}{2}$$

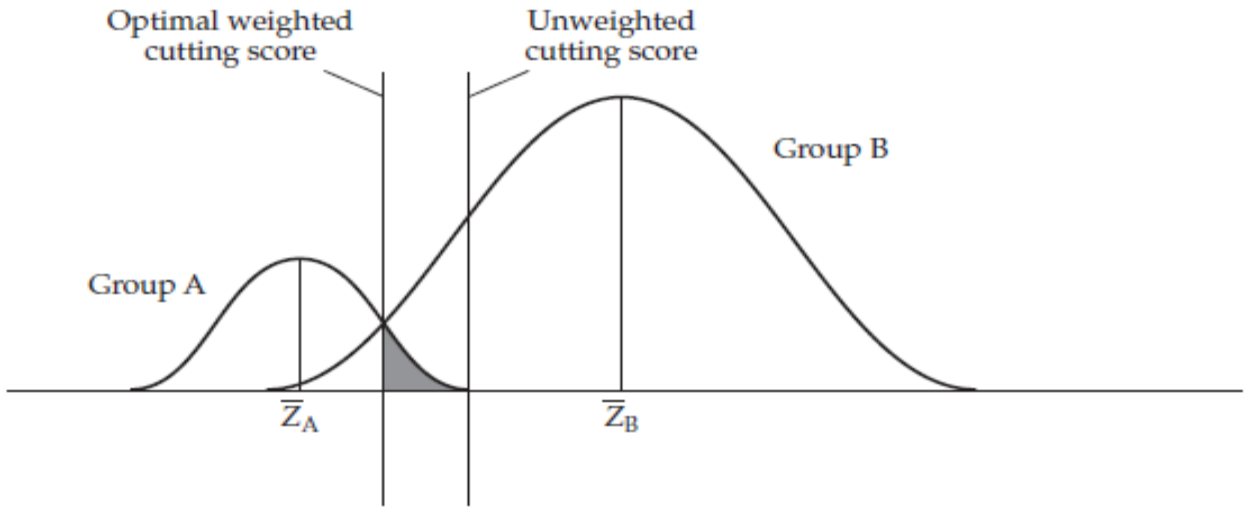
where

Z_{CE} = critical cutting score value for equal group sizes
Z_A = centroid for group A
Z_B = centroid for group B



THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**



- Both of the formulas for calculating the optimal cutting score assume that the distributions are normal and the group dispersion structures are known.
- Both the weighted and unweighted cutting scores are shown.
- It is apparent that if group A is much smaller than group B, the optimal cutting score will be closer to the centroid of group A than to the centroid of group B.
- Also, if the unweighted cutting score was used, none of the objects in group A would be misclassified, but a substantial portion of those in group B would be misclassified.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Costs of Misclassification**
- The optimal cutting score also must consider the cost of misclassifying an object into the wrong group.
- If the costs of misclassifying are approximately equal for all groups, the optimal cutting score will be the one that will misclassify the fewest number of objects across all groups.
- If the misclassification costs are unequal, the optimum cutting score will be the one that minimizes the costs of misclassification.
- **Constructing Classification Matrices**
- To validate the discriminant function through the use of classification matrices, the sample should be randomly divided into two groups.
- One of the groups (the analysis sample) is used to compute the discriminant function.
- The other group (the holdout or validation sample) is retained for use in developing the classification matrix.
- The classification of each observation can be accomplished through either of the classification approaches discussed earlier.
- For the Fisher's approach, an observation is classified into the group with the largest classification function score.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- When using the discriminant scores and the optimal cutting score, the procedure is as follows:
  - Classify an individual into group A if $Z_n < Z_{ct}$
  - or
  - Classify an individual into group B if $Z_n > Z_{ct}$
- Where
  - Zn = discriminant Z score for the nth individual
  - Zct = critical cutting score value
- The results of the classification procedure are presented in matrix form, as shown in Table 4.
- The entries on the diagonal of the matrix represent the number of individuals correctly classified.
- The numbers off the diagonal represent the incorrect classifications.
- The entries under the column labelled “Actual Group Size” represent the number of individuals actually in each of the two groups.
- The entries at the bottom of the columns represent the number of individuals assigned to the groups by the discriminant function.
- The percentage correctly classified for each group is shown at the right side of the matrix, and the overall percentage correctly classified, also known as the hit ratio, is shown at the bottom.
- In our example, the number of individuals correctly assigned to group 1 is 22, whereas 3 members of group 1 are incorrectly assigned to group 2. Similarly, the number of correct classifications to group 2 is 20, and the number of incorrect assignments to group 1 is 5.
- Thus, the classification accuracy percentages of the discriminant function for the actual groups 1 and 2 are 88 and 80 percent, respectively. The overall classification accuracy (hit ratio) is 84 percent.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**

TABLE 4 Classification Matrix for Two-Group Discriminant Analysis

	Predicted Group			
Actual Group	1	2	Actual Group Size	Percentage Correctly Classified
1	22	3	25	88
2	5	20	25	80
Predicted group size	27	23	50	84[a]

[a]Percent correctly classified = (Number correctly classified/Total number of observations) × 100
= [(22 + 20)/50] × 100
= 84%

- One final topic regarding classification procedures is the *t* test available to determine the level of significance for the classification accuracy.
- The formula for a two-group analysis (equal sample size) is

$$t = \frac{p - .5}{\sqrt{\frac{.5(1.0 - .5)}{N}}}$$

where

p = proportion correctly classified
N = sample size

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Establishing Standards of Comparison for the Hit Ratio.**
- As noted earlier, the predictive accuracy of the discriminant function is measured by the hit ratio, which is obtained from the classification matrix.
- ***Standards of Comparison for the Hit Ratio for Equal Group Sizes***
- When the sample sizes of the groups are equal, the determination of the chance classification is rather simple; it is obtained by dividing 1 by the number of groups.
- The formula is: CEQUAL = 1 / Number of groups
- For instance, for a two-group function the chance probability would be .50; for a three-group function the chance probability would be .33; and so forth.
- ***Standards of Comparison for the Hit Ratio for Unequal Group Sizes***
- The determination of the chance classification for situations in which the group sizes are unequal is somewhat more involved.
- Let us assume that we have a total sample of 200 observations divided into holdout and analysis samples of 100 observations each. In the holdout sample, 75 subjects belong to one group and 25 to the other.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- Referred to as the **maximum chance criterion**, we could arbitrarily assign all the subjects to the largest group.
- The maximum chance criterion should be used when the sole objective of the discriminant analysis is to maximize the percentage correctly classified.
- It is also the most conservative standard because it will generate the highest standard of comparison.
- In our simple example of a sample with two groups (75 and 25 people each), using this method would set a 75 percent classification accuracy, what would be achieved by classifying everyone into the largest group without the aid of any discriminant function.
- It could be concluded that unless the discriminant function achieves a classification accuracy higher than 75 percent, it should be disregarded because it has not helped us improve the prediction accuracy we could achieve without using any discriminant analysis at all.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- When group sizes are unequal and the researcher wishes to correctly identify members of all of the groups, not just the largest group, the **proportional chance criterion** is deemed by many to be the most appropriate. The formula for this criterion is

$$CPRO = p2 + (1 - p)2$$

- Where
  - p = proportion of individuals in group 1
  - 1 - p = proportion of individuals in group 2
- Using the group sizes from our earlier example (75 and 25), we see that the proportional chance criterion would be 62.5 percent [.752 + (1.0 – .75)2 = .625] compared with 75 percent.
- Therefore, in this instance, the actual prediction accuracy of 75 percent may be acceptable because it is above the 62.5 percent proportional chance criterion.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Overall Versus Group-Specific Hit Ratios.**
- To this point, we focused on evaluating the overall hit ratio across all groups in assessing the predictive accuracy of a discriminant analysis.
- The researcher also must be concerned with the hit ratio (percent correctly classified) for each separate group.
- If you focus solely on the overall hit ratio, it is possible that one or more groups, particularly smaller groups, may have unacceptable hit ratios while the overall hit ratio is acceptable.
- The researcher should evaluate each group's hit ratio and assess whether the discriminant analysis provides adequate levels of predictive accuracy both at the overall level as well as for each group.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Statistically Based Measures of Classification Accuracy Relative to Chance.**
- A statistical test for the discriminatory power of the classification matrix when compared with a chance model is **Press's *Q* statistic**.
- This simple measure compares the number of correct classifications with the total sample size and the number of groups.
- The calculated value is then compared with a critical value (the chi-square value for 1 degree of freedom at the desired confidence level).
- If it exceeds this critical value, then the classification matrix can be deemed statistically better than chance.
- The *Q* statistic is calculated by the following formula:
  - Press's Q = [N - (nK)]2/N(K - 1)
- Where
  - K = number of groups
  - n = number of observations correctly classified
  - N = total sample size
- Press's Q = [50 - (42 * 2)]/50(2 - 1) = 23.12

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Statistically Based Measures of Classification Accuracy Relative to Chance.**
- The critical value at a significance level of .01 is 6.63. Thus, we would conclude that in the example the predictions were significantly better than chance, which would have a correct classification rate of 50 percent.
- This simple test is sensitive to sample size; large samples are more likely to show significance than small sample sizes of the same classification rate.
- For example, if the sample size is increased to 100 in the example and the classification rate remains at 84 percent, the *Q* statistic increases to 46.24.
- Thus, examine the *Q* statistic in light of the sample size because increases in sample size will increase the *Q* statistic even for the same overall classification rate.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **Casewise Diagnostics**
- The final means of assessing model fit is to examine the predictive results on a case-by-case basis.
- Similar to the analysis of residuals in multiple regression, the objective is to understand which observations (1) have been misclassified and (2) are not representative of the remaining group members.
- Although the classification matrix provides overall classification accuracy, it does not detail the individual case results.
- Also, even if we can denote which cases are correctly or incorrectly classified, we still need a measure of an observation's similarity to the remainder of the group.
- **MISCLASSIFICATION OF INDIVIDUAL CASES**
- When analyzing residuals from a multiple regression analysis, an important decision involves setting the level of residual considered substantive and worthy of attention.
- In discriminant analysis, this issue is somewhat simpler because an observation is either correctly or incorrectly classified.
- All computer programs provide information that identifies which cases are misclassified and to which group they were misclassified.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- **ANALYZING MISCLASSIFIED CASES**
- The purpose of identifying and analyzing the misclassified observations is to identify any characteristics of these observations that could be incorporated into the discriminant analysis for improving predictive accuracy.
- This analysis may take the form of profiling the misclassified cases on either the independent variables or other variables not included in the model.
- **Profiling on the Independent Variables.**
- Examining these cases on the independent variables may identify nonlinear trends or other relationships or attributes that led to the misclassification.
- Several techniques are particularly appropriate in discriminant analysis:
- A graphical representation of the observations is perhaps the simplest yet effective approach for examining the characteristics of observations, especially the misclassified observations.
- The most common approach is to plot the observations based on their discriminant *Z* scores and portray the overlap among groups and the misclassified cases. If two or more functions are retained, the optimal cutting points can also be portrayed to give what is known as a **territorial map** depicting the regions corresponding to each group.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 4: ESTIMATION OF THE DISCRIMINANT MODEL AND ASSESSING OVERALL FIT**
- Plotting the individual observations along with the group centroids, as discussed earlier, shows not only the general group characteristics depicted in the centroids, but also the variation in the group members.
- A direct empirical assessment of the similarity of an observation to the other group members can be made by evaluating the Mahalanobis *D*2 distance of the observation to the group centroid.
- Based on the set of independent variables, observations closer to the centroid have a smaller Mahalanobis *D*2 and are assumed more representative of the group than those farther away.
- The empirical measure should be combined with a graphical analysis, however, because although a large Mahalanobis *D*2 value does indicate observations that are quite different from the group centroids, it does not always indicate misclassification.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- If the discriminant function is statistically significant and the classification accuracy is acceptable, the researcher should focus on making substantive interpretations of the findings.
- This process involves examining the discriminant functions to determine the relative importance of each independent variable in discriminating between the groups.
- Three methods of determining the relative importance have been proposed:
- **1.** Standardized discriminant weights
- **2.** Discriminant loadings (structure correlations)
- **3.** Partial *F* values
- **Discriminant Weights**
- The traditional approach to interpreting discriminant functions examines the sign and magnitude of the standardized **discriminant weight** (also referred to as a **discriminant coefficient**) assigned to each variable in computing the discriminant functions.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- When the sign is ignored, each weight represents the relative contribution of its associated variable to that function.
- Independent variables with relatively larger weights contribute more to the discriminating power of the function than do variables with smaller weights. The sign denotes only that the variable makes either a positive or a negative contribution.
- **Discriminant Loadings**
- **Discriminant loadings**, referred to sometimes as **structure correlations**, are increasingly used as a basis for interpretation because of the deficiencies in utilizing weights.
- Measuring the simple linear correlation between each independent variable and the discriminant function, the discriminant loadings reflect the variance that the independent variables share with the discriminant function.
- In that regard they can be interpreted like factor loadings in assessing the relative contribution of each independent variable to the discriminant function.
- One unique characteristic of loadings is that loadings can be calculated for all variables, whether they were used in the estimation of the discriminant function or not.
- This aspect is particularly useful when a stepwise estimation procedure is employed and some variables are not included in the discriminant function.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- In either simultaneous or stepwise discriminant analysis, variables that exhibit a loading of +/-.40 or higher are considered substantive.
- With stepwise procedures, this determination is supplemented because the technique prevents nonsignificant variables from entering the function.
- However, multicollinearity and other factors may preclude a variable from entering the equation, which does not necessarily mean that it does not have a substantial effect.
- **Partial *F* Values**
- As discussed earlier, two computational approaches—simultaneous and stepwise—can be utilized in deriving discriminant functions.
- When the stepwise method is selected, an additional means of interpreting the relative discriminating power of the independent variables is available through the use of partial *F* values.
- It is accomplished by examining the absolute sizes of the significant *F* values and ranking them.
- Large *F* values indicate greater discriminatory power. In practice, rankings using the *F* values approach are the same as the ranking derived from using discriminant weights, but the *F* values indicate the associated level of significance for each variable.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **Interpretation of Two or More Functions**
- In cases of two or more significant discriminant functions, we are faced with additional problems of interpretation.
- These problems are found both in measuring the total discriminating effects across functions and in assessing the role of each variable in profiling each function separately.
- We address these two questions by introducing the concepts of rotation of the functions, the potency index, and stretched vectors representations.
- **ROTATION OF THE DISCRIMINANT FUNCTIONS**
- After the discriminant functions are developed, they can be rotated to redistribute the variance.
- Basically, rotation preserves the original structure and the reliability of the discriminant solution while making the functions easier to interpret substantively.
- In most instances, the VARIMAX rotation is employed as the basis for rotation.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **POTENCY INDEX**
- Previously, we discussed using the standardized weights or discriminant loadings as measures of a variable's contribution to a discriminant function.
- When two or more functions are derived, however, a composite or summary measure is useful in describing the contributions of a variable across all significant functions.
- The **potency index** is a relative measure among all variables and is indicative of each variable's discriminating power.
- It includes both the contribution of a variable to a discriminant function (its discriminant loading) and the relative contribution of the function to the overall solution (a relative measure among the functions based on eigenvalues).
- The composite is simply the sum of the individual potency indices across all significant discriminant functions.
- Interpretation of the composite measure is limited, however, by the fact that it is useful only in depicting the relative position (such as the rank order) of each variable, and the absolute value has no real meaning.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **POTENCY INDEX**
- The potency index is calculated by a two-step process:
- **Step 1:** *Calculate a potency value of each variable for each significant function.* In the first step, the discriminating power of a variable, represented by the squared value of the unrotated discriminant loading, is "weighted" by the relative contribution of the discriminant function to the overall solution. First, the relative eigenvalue measure for each significant discriminant function is calculated simply as:

$$\text{Relative eigenvalue of discriminant function } j = \frac{\text{Eigenvalue of discriminant function } j}{\text{Sum of eigenvalues across all significant functions}}$$

The potency value of each variable on a discriminant function is then:

$$\text{Potency value of variable } i \text{ on function } j = (\text{Discriminant loading}_{ij})^2 \times \text{Relative eigenvalue of function } j$$

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **POTENCY INDEX**
- **Step 2:** *Calculate a composite potency index across all significant functions.* Once a potency value has been calculated for each function, the composite potency index for each variable is calculated as:
- **Composite potency of variable i = Sum of potency values of variable i across all significant discriminant functions**
- The potency index now represents the total discriminating effect of the variable across all of the significant discriminant functions. It is only a relative measure, however, and its absolute value has no substantive meaning.
- **GRAPHICAL DISPLAY OF DISCRIMINANT SCORES AND LOADINGS**
- To depict group differences on the predictor variables, the researcher can use two different approaches to graphical display.
- The territorial map plots the individual cases on the significant discriminant functions to enable the researcher to assess the relative position of each observation based on the discriminant function scores.
- The second approach is to plot the discriminant loadings to understand the relative grouping and magnitude of each loading on each function.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **Territorial Map.**
- The most common graphical method is the territorial map, where each observation is plotted in a graphical display based on the discriminant function *Z* scores of the observations.
- For example, assume that a three-group discriminant analysis had two significant discriminant functions.
- A territorial map is created by plotting each observation's discriminant *Z* scores for the first discriminant function on the X axis and the scores for the second discriminant function on the Y axis.
- As such, it provides several perspectives on the analysis:
- • Plotting each group's members with differing symbols allows for an easy portrayal of the distinctiveness of each group as well as its overlap with each other group.
- • Plotting each group's centroids provides a means for assessing each group member relative to its group centroid. This procedure is particularly useful when assessing whether large Mahalanobis *D*2 measures lead to misclassification.
- • Lines representing the cutting scores can also be plotted, denoting boundaries depicting the ranges of discriminant scores predicted into each group. Any group's members falling outside these boundaries are misclassified.
- Denoting the misclassified cases allows for assessing which discriminant function was most responsible for the misclassification as well as the degree to which a case is misclassified.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **Vector Plot of Discriminant Loadings.**
- The simplest graphical approach to depicting discriminant loadings is to plot the actual rotated or unrotated loadings on a graph.
- The preferred approach would be to plot the rotated loadings. Similar to the graphical portrayal of factor loadings, this method depicts the degree to which each variable is associated with each discriminant function.
- An even more accurate approach, however, involves plotting the loadings as well as depicting vectors for each loading and group centroid. A **vector** is merely a straight line drawn from the origin (center) of a graph to the coordinates of a particular variable's discriminant loadings or a group centroid.
- With a **stretched vector** representation, the length of each vector becomes indicative of the relative importance of each variable in discriminating among the groups.
- The plotting procedure proceeds in three steps:
- **1.** *Selecting variables:* All variables, whether included in the model as significant or not, may be plotted as vectors. In this way, the importance of collinear variables that are not included, such as in a stepwise solution, can still be portrayed.
- **2.** *Stretching the vectors:* Each variable's discriminant loadings are stretched by multiplying the discriminant loading (preferably after rotation) by its respective univariate *F* value. We note that vectors point toward the groups having the highest mean on the respective predictor and away from the groups having the lowest mean scores.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 5: INTERPRETATION OF THE RESULTS**
- **3.** *Plotting the group centroids:* The group centroids are also stretched in this procedure by multiplying them by the approximate *F* value associated with each discriminant function. If the loadings are stretched, the centroids must be stretched as well to plot them accurately on the same graph.
- The approximate *F* values for each discriminant function are obtained by the following formula:

$$F \text{ value}_{\text{Function}_i} = \text{Eigenvalue}_{\text{Function}_i}\left(\frac{N_{\text{Estimation Sample}} - NG}{NG - 1}\right)$$

where

$$N_{\text{Estimation Sample}} = \text{sample size of estimation sample}$$

- As an example, assume that the sample of 50 observations was divided into three groups. The multiplier of each eigenvalue would be (50 - 3) ÷ (3 - 1) = 23.5.
- When completed, the researcher has a portrayal of the grouping of variables on each discriminant function, the magnitude of the importance of each variable (represented by the length of each vector), and the profile of each group centroid (shown by the proximity to each vector).
- Although this procedure must be done manually in most instances, it provides a complete portrayal of both discriminant loadings and group centroids.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 6: VALIDATION OF THE RESULTS**
- The final stage of a discriminant analysis involves validating the discriminant results to provide assurances that the results have external as well as internal validity.
- *With the propensity of discriminant analysis to inflate the hit ratio if evaluated only on the analysis sample, validation is an essential step.*
- In addition to validating the hit ratios, the researcher should use group profiling to ensure that the group means are valid indicators of the conceptual model used in selecting the independent variables.
- **Validation Procedures**
- Validation is a critical step in any discriminant analysis because many times, especially with smaller samples, the results can lack generalizability (external validity).
- The most common approach for establishing external validity is the assessment of hit ratios.
- Validation can occur either with a separate sample (holdout sample) or utilizing a procedure that repeatedly processes the estimation sample.
- External validity is supported when the hit ratio of the selected approach exceeds the comparison standards that represent the predictive accuracy expected by chance.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 6: VALIDATION OF THE RESULTS**
- **UTILIZING A HOLDOUT SAMPLE**
- Most often the validation of the hit ratios is performed by creating a holdout sample, also referred to as the **validation sample**.
- The purpose of utilizing a holdout sample for validation purposes is to see how well the discriminant function works on a sample of observations not used to derive the discriminant function.
- This process involves developing a discriminant function with the analysis sample and then applying it to the holdout sample.
- The justification for dividing the total sample into two groups is that an upward bias will occur in the prediction accuracy of the discriminant function if the individuals used in developing the classification matrix are the same as those used in computing the function; that is, the classification accuracy will be higher than is valid when applied to the estimation sample.
- Other researchers have suggested that even greater confidence could be placed in the validity of the discriminant function by following this procedure several times.
- Instead of randomly dividing the total sample into analysis and holdout groups once, the researcher would randomly divide the total sample into analysis and holdout samples several times, each time testing the validity of the discriminant function through the development of a classification matrix and a hit ratio.
- Then the several hit ratios would be averaged to obtain a single measure.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 6: VALIDATION OF THE RESULTS**
- **CROSS-VALIDATION**
- The cross-validation approach to assessing external validity is performed with multiple subsets of the total sample.
- The most widely used approach is the jackknife method.
- Cross-validation is based on the “leave-one-out” principle.
- The most prevalent use of this method is to estimate $k - 1$ subsamples, eliminating one observation at a time from a sample of k cases.
- A discriminant function is calculated for each subsample and then the predicted group membership of the eliminated observation is made with the discriminant function estimated on the remaining cases.
- After all of the group membership predictions have been made, one at a time, a classification matrix is constructed and the hit ratio calculated.
- Cross-validation is quite sensitive to small sample sizes. Guidelines suggest that it be used only when the smallest group size is at least three times the number of predictor variables, and most researchers suggest a ratio of 5:1.
- However, cross-validation may represent the only possible validation approach in instances where the original sample is too small to divide into analysis and holdout samples but still exceeds the guidelines already discussed.
- Cross-validation is also becoming more widely used as major computer programs provide it as a program option.

THE DECISION PROCESS FOR DISCRIMINANT ANALYSIS

- **STAGE 6: VALIDATION OF THE RESULTS**
- **Profiling Group Differences**
- Another validation technique is to profile the groups on the independent variables to ensure their correspondence with the conceptual bases used in the original model formulation.
- After the researcher identifies the independent variables that make the greatest contribution in discriminating between the groups, the next step is to profile the characteristics of the groups based on the group means.
- This profile enables the researcher to understand the character of each group according to the predictor variables.
- For example, referring to the KitchenAid survey data presented in Table 1, we see that the mean rating on “durability” for the “would purchase” group is 7.4, whereas the comparable mean rating on “durability” for the “would not purchase” group is 3.2. Thus, a profile of these two groups shows that the “would purchase” group rates the perceived durability of the new product substantially higher than the “would not purchase” group.
- Another approach is to profile the groups on a separate set of variables that should mirror the observed group differences.
- This separate profile provides an assessment of external validity in that the groups vary on both the independent variable(s) and the set of associated variables.