Chapter 12
Application of Discriminant Analysis: For Developing a Classification Model

Learning Objectives

After completing this chapter, you should be able to do the following:

- Understand the importance of discriminant analysis in research.
- List down the research situation where discriminant analysis can be used.
- Understand the importance of assumptions used in discriminant analysis.
- Know the different concepts used in discriminant analysis.
- Understand the steps involved in using SPSS for discriminant analysis.
- To interpret the output obtained in discriminant analysis.
- Explain the procedure in developing the decision rule using discriminant model.
- Know to write the results of discriminant analysis in standard format.

Introduction

Often we come across a situation where it is interesting to know as to why the two naturally occurring groups are different. For instance, after passing the school, the students can opt for continuing further studies, or they may opt for some skill-related work. One may be interested to know as to what makes them to choose their course of action. In other words, it may be desired to know on what parameters these two groups may be distinct. Similarly one may like to identify the parameters which distinguish the liking of two brands of soft drink by the customers or which make the engineering and management students different. Thus, to identify the independent parameters responsible for discriminating these two groups, a statistical technique known as discriminant analysis (DA) is used. The discriminant analysis is a multivariate statistical technique used frequently in management, social sciences, and humanities research. There may be varieties of situation where this technique can play a major role in decision-making process. For instance, the government is very keen that more and more students should opt for the science stream in order to have the technological advancement in the country.

J.P. Verma, *Data Analysis in Management with SPSS Software*,
DOI 10.1007/978-81-322-0786-3_12, © Springer India 2013

Therefore, one may investigate the factors that are responsible for class XI students to choose commerce or science stream. After identifying the parameters responsible for discriminating a science and commerce student, the decision makers may focus their attention to divert the mindset of the students to opt for science stream.

Yet another application where discriminant analysis can be used is in the food industry. In launching the new food product, much of its success depends upon its taste, and, therefore, product formulation must be optimized to obtain desirable sensory quality expected by consumers. Thus, the decision maker may be interested to know the parameters that distinguish the existing similar product and new proposed product in terms of the product properties like sensory characteristics, percent of ingredients added, pricing, and contents. In this chapter, the discriminant analysis technique shall be discussed in detail along with its application with SPSS.

What Is Discriminant Analysis?

Discriminant analysis is a multivariate statistical technique used for classifying a set of observations into predefined groups. The purpose is to determine the predictor variables on the basis of which groups can be determined. The discriminant model is built on the basis of a set of observations for which the groups are known. This set of observation is the past data on the basis of which discriminant analysis technique constructs a set of linear functions of the predictors, known as discriminant function, such that

$$Z = c + b_1X_1 + b_2X_2 + \ldots\ldots b_nX_n \tag{12.1}$$

where

c is a constant
b's are the discriminant coefficients
X's are the predictor variables

Only those independent variables are picked up which are found to have significant discriminating power in classifying a subject into any of the two groups. The discriminant function so developed is used for predicting the group of a new observation set.

The discriminant analysis is actually known as discriminant function analysis but in short one may use the term discriminant analysis. In discriminant analysis, the dependent variable is a categorical variable, whereas independent variables are metric. The dependent variable may have more than two classes, but the discriminant analysis is more powerful if it has two classifications. In this text, the discriminant analysis shall be discussed only for two-group problem.

After developing the discriminant model, for a given set of new observation the discriminant function Z is computed, and the subject/object is assigned to first group if the value of Z is less than 0 and to second group if more than 0. This

criterion holds true if an equal number of observations are taken in both the groups for developing a discriminant function. However, in case of unequal sample size, the threshold may vary on either side of zero.

The main purpose of a discriminant analysis is to predict group membership based on a linear combination of the predictive variables. In using this technique, the procedure starts with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known.

A second purpose of the discriminant analysis is to study the relationship between group membership and the variables used to predict group membership. This provides information about the relative importance of independent variables in predicting group membership.

Discriminant function analysis is similar to the ordinary least square (OLS) regression analysis. The only difference is in the nature of dependent variable. In discriminant function analysis, the dependent variable is essentially a categorical (preferably dichotomous) variable, whereas in multiple regression it is a continuous variable. Other differences are in terms of the assumptions being satisfied in using discriminant analysis which shall be discussed later in this chapter.

Terminologies Used in Discriminant Analysis

Discriminant analysis provides discriminant function which is used to classify an individual or cases into two categories on the basis of the observations on the predictor variables. If the discriminant model developed in the analysis is robust for a set of data, the percentage of correct classification of cases in the classification table increases. To understand the application of discriminant analysis using SPSS on any data set, it is essential to know its basics.

Variables in the Analysis

In discriminant analysis, the dependent variable is categorical in nature. It may have two or more categories. The procedure used in discriminant analysis becomes very complicated if the dependent variable has more than two categories. Further the efficiency of the model also decreases in that case. The model becomes very powerful if the dependent variable has only two categories. The dependent variable is also known as criterion variable. In SPSS, dependent variable is known as grouping variable. It is the object of classification on the basis of independent variables.

The independent variables in the discriminant analysis are always metric. In other words, the data obtained on the independent variables must be measured

either on interval or ratio scale. The independent variables in discriminant analysis are also known as predictor variables.

Discriminant Function

A discriminant function is a latent variable which is constructed as a linear combination of independent variables, such that

$$Z = c + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

where
$b_1, b_2, \ldots, b_n$ are discriminant coefficients
$X_1, X_2, \ldots, X_n$ are discriminating variables
c is a constant

The discriminant function is also known as canonical root. This discriminant function is used to classify the subject/cases into one of the two groups on the basis of the observed values on the predictor variables.

Classification Matrix

In discriminant analysis, the classification matrix serves as a yardstick in measuring the accuracy of a model in classifying an individual/case into one of the two groups. The classification matrix is also known as confusion matrix, assignment matrix, or prediction matrix. It tells us as to what percentage of the existing data points are correctly classified by the model developed in discriminant analysis. This percentage is somewhat similar to R^2 (percentage of variation in dependent variable explained by the model).

Stepwise Method of Discriminant Analysis

Discriminant function can be developed either by entering all *independent variables together* or in s*tepwise* depending upon whether the study is confirmatory or exploratory. In confirmatory data analysis, discriminant function is developed on the basis of all the independent variables selected in the study, whereas in exploratory study the independent variables are selected one by one. In stepwise discriminant analysis, a variable is retained in the model if its regression coefficient is significant at 5% level and removed from the model if it is not significant at 10% level.

Power of Discriminating Variables

After developing the model in discriminant analysis based on selected independent variables, it is important to know the relative importance of the variables so selected. This relative importance of the variable is determined by the coefficient of the discriminating variable in the discriminant function. SPSS provides these coefficients in the output and are named as standardized canonical discriminant function coefficients. The higher the value of coefficient, the better is the discriminating power.

Box's M Test

While applying ANOVA, one of the assumptions was that the variances are equivalent for each group, but in DA the basic assumption is that the variance-covariance matrices are equivalent. By using Box's M tests, we test a null hypothesis that the covariance matrices do not differ between groups formed by the dependent variable. The researcher would not like this test to be significant so that the null hypothesis that the groups do not differ can be retained. Thus, if the Box's M test is insignificant, it indicates that the assumptions required for DA holds true.

However, with large samples, a significant result of Box's M is not regarded as too important. Where three or more groups exist, and Box's M is significant, groups with very small log determinants should be deleted from the analysis.

Eigenvalues

Eigenvalue is the index of overall model fit. It provides information on each of the discriminant functions (equations) produced. In discriminant analysis, the maximum number of discriminant functions produced is the number of groups minus 1. In case dependent variable has two categories, only one discriminant function shall be generated. In DA, one tries to predict the group membership from a set of predictor variables. If the dependent variable has two categories and there are n predictive variables, then a linear discriminant equation, $Z_i = c + b_1X_1 + b_2X_2 + \ldots + b_nX_n$, is constructed such that the two groups differ as much as possible on Z. Here, one tries to choose the weights b_1, b_2,..., b_n in computing a discriminant score (Z_i) for each subject so that if an ANOVA on Z is done, the ratio of the between groups sum of squares to the within groups sum of squares is as large as possible. The value of this ratio is known as eigenvalue.

Thus, eigenvalue is computed with the data on Z and is a quantity maximized by the discriminant function coefficients.

$$\text{Eigenvalue} = \frac{\text{SS}_{\text{Between\_groups}}}{\text{SS}_{\text{Within\_groups}}} \tag{12.2}$$

The larger the eigenvalue, the better is the model in discriminating between the groups.

The Canonical Correlation

The canonical correlation in discriminant analysis is equivalent to eta in an ANOVA and is equal to the point biserial correlation r_b between group and Z. Square of the canonical correlation indicates the percentage of variation explained by the model in the grouping variable and is similar to R^2. The canonical correlation is computed on Z which is as follows:

$$\text{Canonical correlation} = \sqrt{\frac{\text{SS}_{\text{Between\_groups}}}{\text{SS}_{\text{Total}}}} \tag{12.3}$$

Wilks' Lambda

It is used to indicate the significance of discriminant function developed in the discriminant analysis. The value of Wilks' lambda provides the proportion of total variability not explained by the discriminant model. For instance, if the value of Wilks' lambda is 0.28, it indicates that 28% variability is not explained by the model. The value of Wilks' lambda ranges from 0 to 1, and low value of it (closer to 0) indicates better discriminating power of the model. Thus, the Wilks' lambda is the converse of the squared canonical correlation.

What We Do in Discriminant Analysis

Different steps that are involved in discriminant analysis have been discussed in this section. Initially you may not understand all the steps clearly but continue to read this chapter, and once you complete reading the solved example using SPSS discussed in this chapter, your understanding level about this topic shall be enhanced. All the steps discussed below cannot be performed manually but may be achieved by using any statistical package. So go through these steps and try understanding the outputs of your discriminant analysis.

1. The first step in the discriminant analysis is to identify the independent variables having significant discriminant power. This is done by taking all the independent variables together in the model or one by one. The option for these two methods can be seen in SPSS as "*Enter independents together*" and "*Use stepwise method,*" respectively.
   In *stepwise method*, an independent variable is entered in the model if its corresponding regression coefficient is significant at 5% level and excluded at subsequent stages until and unless it is significant at 10% level. Thus, in developing discriminant function, the model will enter only significant independent variables. The model so developed is required to be tested for its robustness.
2. In the second step, a discriminant function model is developed by using the discriminant coefficients of the predictor variables and the value of constant shown in the "*Unstandardized canonical discriminant function coefficients*" table generated in the SPSS output. This is similar to developing of regression equation. This way, the function so generated may be used to classify an individual into any of the two groups. The discriminant function shall look like as follows:

$$Z = c + b_1X_1 + b_2X_2 + \ldots\ldots\ldots + b_nX_n$$

   where

   Z is the discriminant function
   X's are predictor variables in the model
   c is the constant
   b's are the discriminant constants of the predictor variables
3. After developing discriminant model, the Wilks' lambda is computed in the third step for testing the significance of discriminant function developed in the model. This indicates the robustness of discriminant model. The value of Wilks' lambda ranges from 0 to 1, and the lower value of it close to 0 indicates better discriminating power of the model. Further, significant value of chi-square indicates that the discrimination between the two groups is highly significant.
   After selecting independent variables as predictors in the discriminant model, the model is tested for its significance in classifying the subjects/cases correctly into groups. For this, SPSS generates a classification matrix. This is also known as confusion matrix. This matrix shows the number of correct and wrong classification of subjects in both the groups. High percentage of correct classification indicates the validity of the model. The level of accuracy shown in the classification matrix may not hold for all future classification of new subjects/cases.
4. In the fourth step, the relative importance of predictor variables in discriminating the two groups is discussed. The SPSS generates the "*Standardized canonical discriminant function coefficients*" table. The variable with higher coefficient in the table is the most powerful in discriminating the two groups, whereas the variable having least coefficient indicates low discriminating power.

5. Finally, a criterion for classification is developed on the basis of the midpoint of the mean value of the transformed groups obtained in the table titled "*Functions at group centroids*" generated in the SPSS output. If the value of the function Z computed on the basis of the equation developed in step 2 is less than this midpoint, the subject is classified in one group and if it is more than it is classified in second group.

Assumptions in Using Discriminant Analysis

While applying discriminant analysis, one should test the assumptions used in this analysis. Following are the assumptions which are required to be fulfilled while using this analysis:

1. Each of the independent variables is normally distributed. This assumption can be examined by the histograms of frequency distributions. In fact, violations of the normality assumption are usually not serious because in that case the resultant significance tests are still reliable. One may use specific tests like skewness and kurtosis for testing the normality in addition to graphs.
2. All variables have linear and homoscedastic relationships. It is assumed that the variance/covariance matrices of variables are homogeneous in both the groups. Box M test is used for testing the homogeneity of variances/covariances in both the groups. However, it is sensitive to deviations from multivariate normality and should not be taken too seriously.
3. Dependent variable is a true dichotomy. The continuous variable should never be dichotomized for the purpose of applying discriminant analysis.
4. The groups must be mutually exclusive, with every subject or case belonging to only one group.
5. All cases must be independent. One should not use correlated data like before-after and matched pair data.
6. Sample sizes of both the groups should not differ to a great extent. If the sample sizes are in the ratio 80:20, logistic regression may be preferred.
7. Sample size must be sufficient. As a guideline, there should be at least five to six times as many cases as independent variables.
8. No independent variables should have a zero variability in either of the groups formed by the dependent variable.
9. Outliers should not be present in the data. To solve this problem, inspect descriptive statistics.

Research Situations for Discriminant Analysis

The discriminant analysis is used to develop a model for discriminating the future cases/objects into one of the two groups on the basis of predictor variables. Hence, it is widely used in the studies related to management, social sciences, humanities,

and other applied sciences. Some of the research situations where this analysis can be used are discussed below:

1. In a hospitality firm, the data can be collected on employees in two different job classifications: (1) customer support personnel and (2) back office management. The human resources manager may like to know if these two job classifications require different personality types. Each employee may be tested by a battery of psychological test which consists of a measure of socialization trait, extrovertness, frustration level, and orthodox approach.
   The model can be used to priorities the predictor variable which can be used to identify the employees in different category during selection process. Further, the model may be helpful in developing the training program for future employees recruited in different categories.
2. A college authority might divide a group of past graduate students into two groups: students who finished the economics honors program in 3 years and those who did not. The discriminant analysis could be used to predict successful completion of the honors program based on the independent variables like SAT score, XII maths score, and age of the candidates. Investigating the prediction model might provide insight as to how each predictor individually and in combination predicted completion or noncompletion of the economics honors program at the undergraduate level.
3. A marketing manager may like to develop a model on buying two different kinds of toothpaste on the basis of the product and customer profiles. The independent variables may consist of age and sex of the customer and contained quantity, taste, price of the products, etc. The insight from the developed model may provide the decision makers in the company to develop and market their products with success.
4. A social scientist may like to know the predictor variable which is responsible for smoking. The data on variables like the age at which the first cigarette was smoked and other reasons of smoking like self-image, peer pressure, and frustration level can be studied to develop a model for classifying an individual into smoker and nonsmoker. The knowledge so accrued from the developed model may be used to start the ad campaign against smoking.
5. In medical research, one may like to predict whether patient would survive from burn injury based on the combinations of demographic and treatment variables. The predictor variables might include burn percentage, body parts involved, age, sex, and time between incident and arrival at hospital. In such situations, the discriminant model so developed would allow a doctor to assess the chances of recovery based on predictor variables. The discriminant model might also give insight into how the variables interact in predicting recovery.

Solved Example of Discriminant Analysis Using SPSS

Example 12.1 The marketing division of a bank wants to develop a policy for issuing visa gold card to its customers through which one can shop and withdraw up

to Rs. 100,000 at a time for 30 days without any interest. Out of several customers, only a handful number of customers are required to be chosen for such facility. Thus, a model is required to be made on the basis of the existing practices for issuing similar card to the customers on the basis of the following data. The data was collected on 28 customers in the bank who were either issued or denied similar card earlier. Apply discriminant analysis to develop a discriminant function for issuing or denying the golden visa card to the customers on the basis of their profile. Also test the significance of the model so obtained. Discuss the efficiency of classification and relative importance of the predictor variables retained in the model (Table 12.1).

Solution

Here it is required to do the following:

1. To develop a discriminant function for deciding whether a customer be issued a golden credit card

Table 12.1 Account details of the customers

S. N.	Credit card	Average daily balance last 1 year	Number of days balance <50,000 last 1 year	Annual income in lakh	Family size	Average number of transaction/ month
1	Issued	68,098	2	36.52	4	8
2	Denied	43,233	12	26.45	3	13
3	Issued	50,987	0	25.6	5	11
4	Denied	39,870	31	26.85	5	12
5	Denied	37,653	51	25.65	6	11
6	Denied	35,347	48	28.45	5	14
7	Issued	65,030	1	22.45	2	4
8	Issued	72,345	0	42.34	5	6
9	Denied	34,534	32	31.9	4	8
10	Issued	87,690	1	30.45	6	15
11	Denied	43,563	4	28.45	5	10
12	Denied	50,879	6	24.8	6	9
13	Denied	58,034	1	24.45	5	12
14	Issued	76,345	0	29.45	6	3
15	Issued	69,067	3	34.24	4	11
16	Denied	43,008	5	54.45	4	8
17	Issued	75,437	2	28.76	8	20
18	Denied	34,009	8	34.25	4	14
19	Issued	52,409	4	31.45	4	7
20	Denied	51,654	4	31.8	3	13
21	Issued	64,065	2	25.67	5	10
22	Denied	49,003	4	33.45	2	7
23	Issued	65,030	1	25.63	4	15
24	Issued	59,024	2	32.52	5	12
25	Issued	75,007	0	28.45	3	8
26	Denied	46,342	12	34.54	5	15
27	Denied	56,803	1	32.76	4	17
28	Issued	59,034	3	26.87	3	8

2. To identify the predictor variable in developing the model and find their relative importance
3. To test the significance of the model
4. To explain the efficiency of classification

These issues shall be discussed with the output generated by the SPSS in this example. Thus, the procedure of using SPSS for discriminant analysis in the given example shall be explained first, and thereafter the output shall be discussed in the light of the objectives of the study.

SPSS Commands for Discriminant Analysis

In order to perform discriminant analysis with SPSS, a data file needs to be prepared first. Since the initial steps in preparing the data file has been explained in earlier chapters, it will not be repeated here again. In case of difficulty, you may go through the procedure discussed in Chap. 1 in this regard. Take the following steps for generating the outputs in discriminant analysis:

(i) *Data file:* Here, five independent variables and one dependent variable need to be defined. The dependent variable *Card_decision* is defined as a nominal variable, whereas all five independent variables as scale variables in SPSS. After preparing the data file by defining variable names and their labels, the screen will look like as shown in Fig. 12.1.

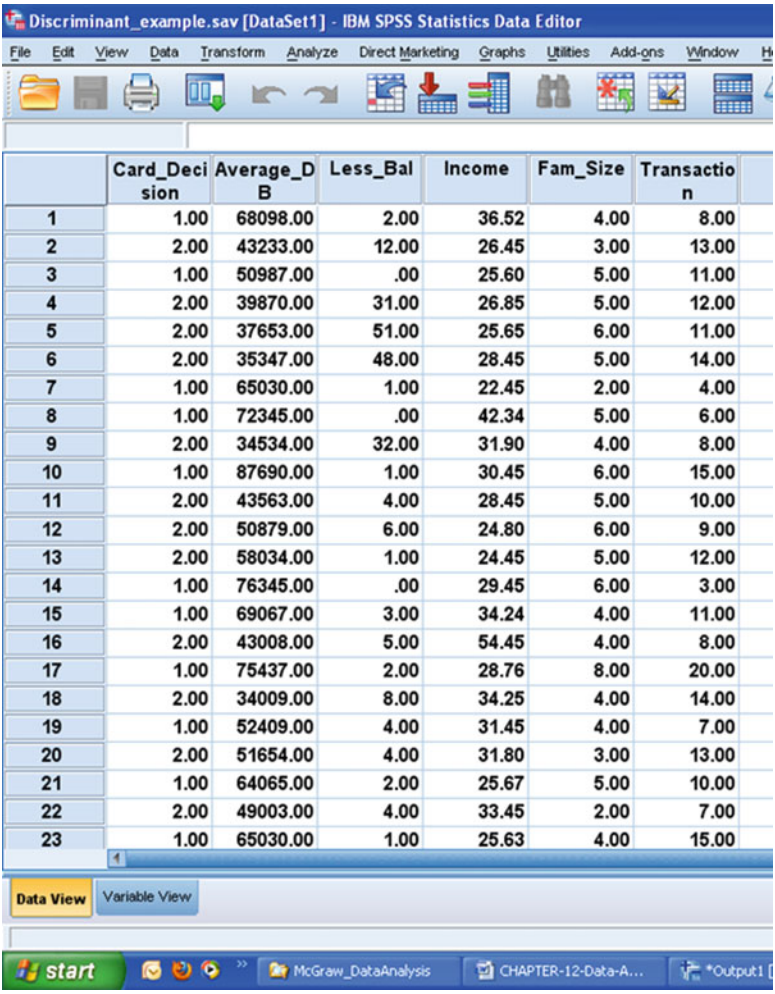
(ii) *Initiating command for discriminant analysis:* After preparing the data file, click the following command sequence in the Data View:

Analyze → Classify → Discriminant

The screen shall look like Fig. 12.2.

(iii) *Selecting variables for discriminant analysis:* After clicking the **Discriminant** option, the SPSS will take you to the window where variables are selected.

- Select the dependent variable *Card_Decision* from left panel to the "Grouping Variable" section of the right panel. Define minimum and maximum range of the grouping variable as "1" and "2" and click continue.
- Select all independent variables from left panel and bring them to the "Independents" section of the right panel.
- Check the option "Use stepwise method" if you have many independent variables and the effort is to identify the relevant predictive variables. Such studies are known as explorative studies. Whereas if you want to go for confirmatory analysis, check the option "Enter independents together." Here, the model is built on all the independent variables; hence, the option "Enter independents together" is checked. In this case, the effort is to test



	Card_Decision	Average_DB	Less_Bal	Income	Fam_Size	Transaction
1	1.00	68098.00	2.00	36.52	4.00	8.00
2	2.00	43233.00	12.00	26.45	3.00	13.00
3	1.00	50987.00	.00	25.60	5.00	11.00
4	2.00	39870.00	31.00	26.85	5.00	12.00
5	2.00	37653.00	51.00	25.65	6.00	11.00
6	2.00	35347.00	48.00	28.45	5.00	14.00
7	1.00	65030.00	1.00	22.45	2.00	4.00
8	1.00	72345.00	.00	42.34	5.00	6.00
9	2.00	34534.00	32.00	31.90	4.00	8.00
10	1.00	87690.00	1.00	30.45	6.00	15.00
11	2.00	43563.00	4.00	28.45	5.00	10.00
12	2.00	50879.00	6.00	24.80	6.00	9.00
13	2.00	58034.00	1.00	24.45	5.00	12.00
14	1.00	76345.00	.00	29.45	6.00	3.00
15	1.00	69067.00	3.00	34.24	4.00	11.00
16	2.00	43008.00	5.00	54.45	4.00	8.00
17	1.00	75437.00	2.00	28.76	8.00	20.00
18	2.00	34009.00	8.00	34.25	4.00	14.00
19	1.00	52409.00	4.00	31.45	4.00	7.00
20	2.00	51654.00	4.00	31.80	3.00	13.00
21	1.00	64065.00	2.00	25.67	5.00	10.00
22	2.00	49003.00	4.00	33.45	2.00	7.00
23	1.00	65030.00	1.00	25.63	4.00	15.00

Fig. 12.1 Screen showing partial data file for the discriminant analysis in SPSS

the model. Such studies are known as confirmatory studies. In this example, all the variables have been selected to build the model. The screen will look like Fig. 12.3.

(iv) *Selecting the option for computation:* After selecting variables, different option needs to be defined for generating the output in discriminant analysis. Take the following steps:

- Click the tag **Statistics** in the screen shown in Fig. 12.3. and
- Check the option of "Means" and "Box's M" in the "Descriptives" section.

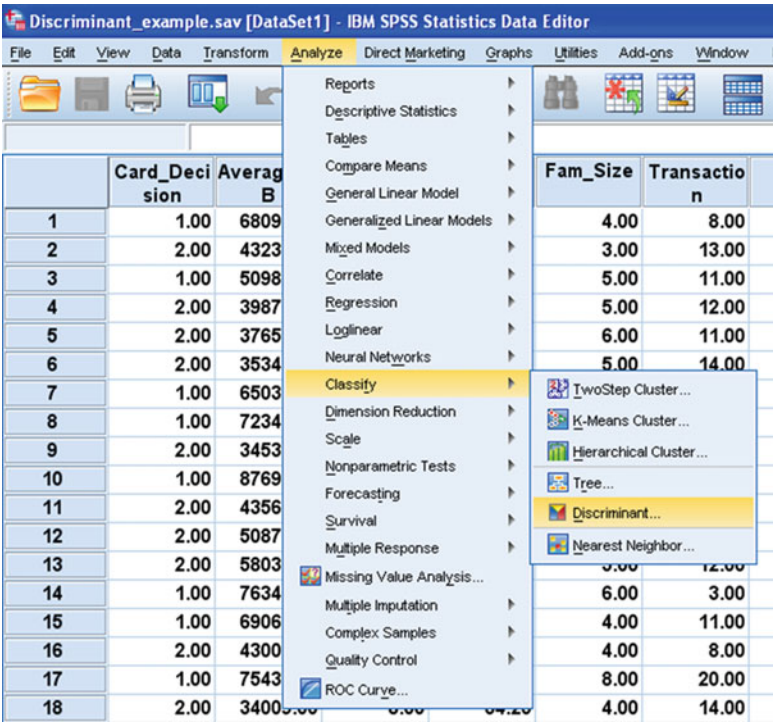


Fig. 12.2 Screen showing SPSS commands for discriminant analysis

- Check the options "Fisher's" and "Unstandardized" in the "Function Coefficients" section. The screen showing these options shall look like as shown in Fig. 12.4.
- Press ***Continue***. This will take you back to the screen shown in Fig. 12.3.
- Click the tag **Classify** in the screen as shown in screen 12.3. and
- Check the option "Summary table" in the Display section.
- Check the option "Casewise results" if you want to know wrongly classified cases by the model.

The screen for these options shall look like Fig. 12.5.

- Click ***Continue***.
- Click **OK** for output.

(v) *Getting the output:* After clicking the **OK** option in Fig. 12.3, the output in the discriminant analysis shall be generated in the output window. Selected outputs can be copied in the word file by using the right click of the mouse over identified area of the output. Out of many outputs generated by the SPSS, the following relevant outputs have been picked up for discussion:

1. Group statistics including mean and standard deviation

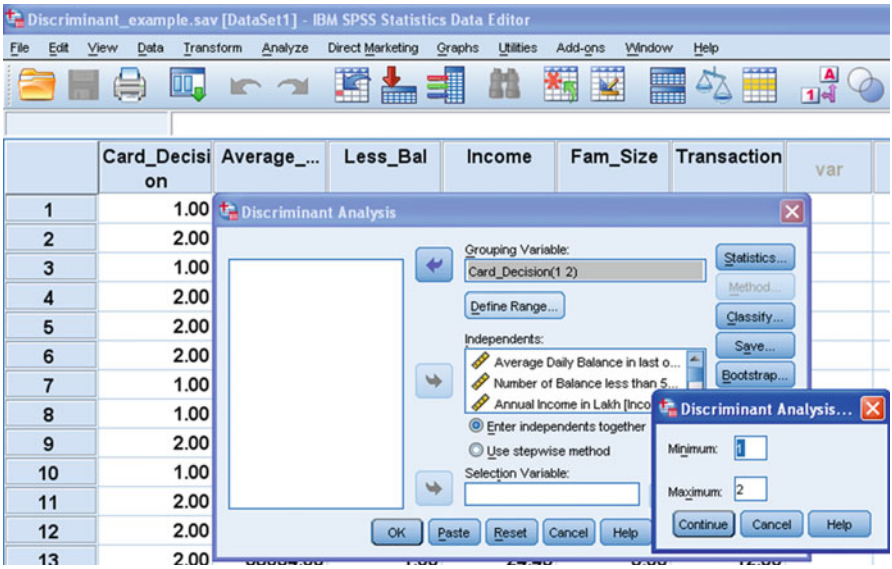


Fig. 12.3 Screen showing selection of variables for discriminant analysis

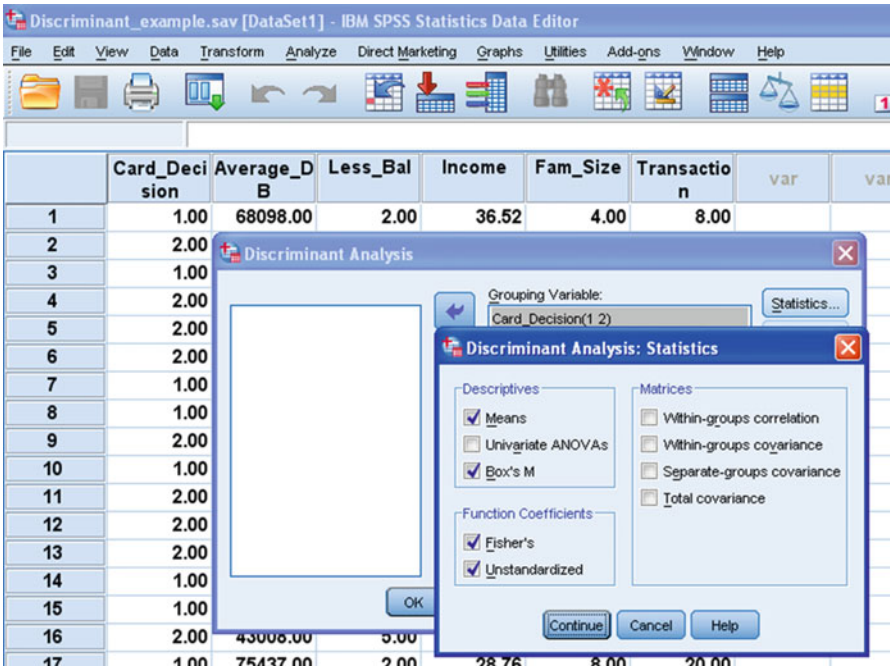


Fig. 12.4 Screen showing the options for descriptive statistics and discriminant coefficients

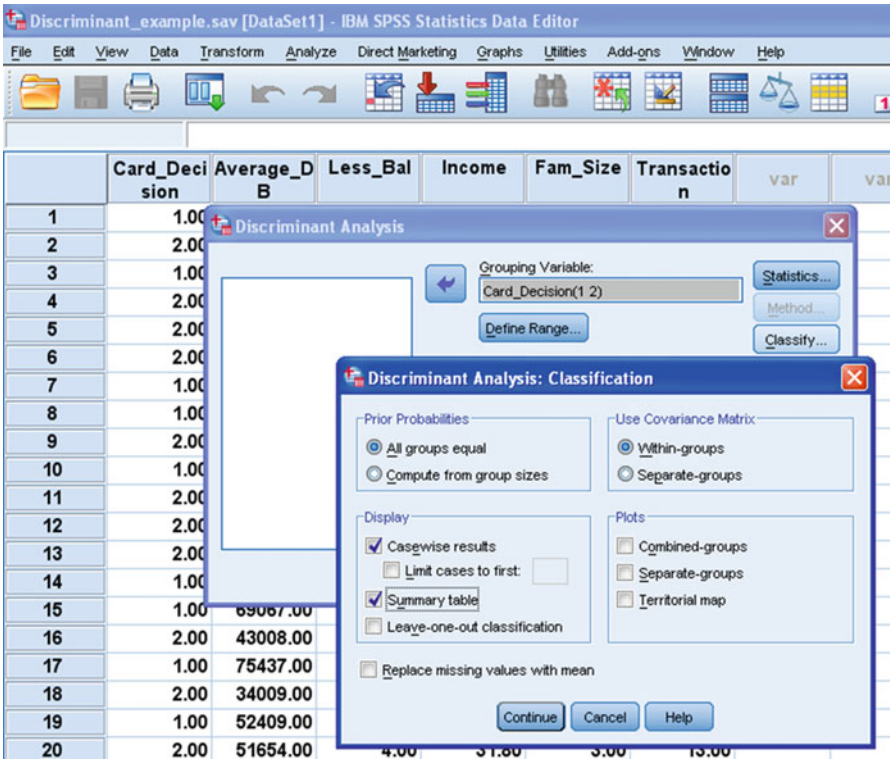


Fig. 12.5 Screen showing the options for classification matrix

2. Unstandardized canonical discriminant function coefficients table
3. Eigen values and canonical correlation
4. Wilks' lambda and chi-square test
5. Classification matrix
6. Standardized canonical discriminant function coefficients
7. Functions at group centroids

These outputs so generated by the SPSS are shown in Tables 12.2–12.8 and Fig. 12.6.

Interpretation of Various Outputs Generated in Discriminant Analysis

The above-mentioned output so generated by the SPSS will now be discussed to answer the issues raised in the example.

1. Table 12.2 shows descriptive statistics containing mean and standard deviation for all the variables in both the groups, that is, card issued group and card denied

Table 12.2 Group statistics: mean and standard deviation of all independent variables in different groups

Issue decision		Mean	SD
Card issued	Average daily balance in last 1 year	67,112.00	9989.74
	Number of balance less than 5,000 in last 1 year	1.509	1.29
	Annual income in lakh	30.03	5.19
	Family size	4.57	1.50
	Average transaction per month	9.86	4.62
Card denied	Average daily balance in last 1 year	44,566.57	7923.67
	Number of balance less than 5,000 in last 1 year	15.64	17.38
	Annual income in lakh	31.30	7.56
	Family size	4.36	1.15
	Average transaction per month	11.64	2.95
Total	Average daily balance in last 1 year	55,839.29	14,493.43
	Number of balance less than 5,000 in last 1 year	8.5714	14.08
	Annual income in lakh	30.67	6.39
	Family size	4.46	1.32
	Average transaction per month	10.75	3.91

Table 12.3 Unstandardized canonical discriminant function coefficients

Variables selected	Function 1
Average daily balance in last 1 year (X_1)	.000
Number of balance less than 5,000 in last 1 year (X_2)	−.002
Annual income in lakh (X_3)	−.028
Family size (X_4)	.017
Average transaction per month (X_5)	−.099
Constant	−4.253

group. The readers may draw relevant conclusions as per their objectives from this table.

2. Table 12.3 reveals the value of unstandardized discriminant coefficients which are used in constructing discriminant function. Since all independent variables were included to develop the model, the discriminant coefficients of all the five independent variables are shown in Table 12.3.
   Thus, discriminant function can be constructed by using the values of constant and coefficients of these five independent variables as shown in Table 12.3.

$$Z = -4.253 - .002 \times X_2 - .028 \times X_3 + .017 \times X_4 - .099 \times X_5$$

where

X_2 is number of balance less than 5,000 in last 1 year
X_3 is annual income in lakh
X_4 is family size
X_5 is average transaction per month

Table 12.4 Eigenvalues

Function	Eigenvalue	% of variance	Cumulative %	Canonical correlation
1	1.975[a]	100.0	100.0	.815

[a]First 1 canonical discriminant functions were used in the analysis

Table 12.5 Wilks' lambda and chi-square test

Test of function(s)	Wilks' lambda	Chi-square	df	Sig.
1	.336	25.618	5	.000

Table 12.6 Classification results[a]

		Predicted group membership		
	Issue decision	Card issued	Card denied	Total
Original count	Card issued	12	2	14
	Card denied	1	13	14
%	Card issued	85.7	14.3	100.0
	Card denied	7.1	92.9	100.0

[a]89.3% of original grouped cases correctly classified

3. The canonical correlation is 0.815 as shown in Table 12.4. This indicates that approximately 66% of the variation in the two groups is explained by the discriminant model.
   Since the Wilks' lambda provides the proportion of unexplained variance by the model, the lesser its value, the better is the discriminant model. The value of Wilks' lambda lies in between 0 and 1. Its value here is 0.336 as shown in Table 12.5; hence, the model can be considered good because only 33.6% variability is not explained by the model. To test the significance of Wilks' lambda, the value of chi-square is calculated which is shown in Table 12.5. Since the *p* value associated with it is .000 which is less than .05, it may be inferred that the model is good.
4. Table 12.6 is a classification matrix which shows the summary of correct and wrong classification of cases in both the groups on the basis of the developed discriminant model. This table shows that out of 14 customers whom credit card was issued, 12 were correctly classified by the developed model and 2 were wrongly classified in the card denied group. On the other hand, out of 14 customers whom card was denied, 13 were classified by the model correctly in the card denied group and only 1 customer was wrongly classified in the card issued group. Thus, out of 28 cases, 25 (89.3%) cases were correctly classified by the model which is quite high; hence, the model can be considered as valid. Since this model is developed on the basis of a small sample, the level of accuracy shown in the classification matrix may not hold for all future classification of new cases.
5. Table 12.7 shows the standardized discriminant coefficients of the independent variables in the model. The magnitude of these coefficients indicates the discriminating power of the variables in the model. The variable having higher

Table 12.7 Standardized canonical discriminant function coefficients

Variables selected	Function 1
Average daily balance in last 1 year	.988
Number of balance less than 5,000 in last 1 year	−.019
Annual income in lakh	−.184
Family size	.023
Average transaction per month	−.382

Table 12.8 Functions at group centroids

Issue decision	Function 1
Card issued	1.354
Card denied	−1.354

Unstandardized canonical discriminant functions evaluated at group means

magnitude of the absolute function value is more powerful in discriminating the two groups. Since absolute function value of the variable *average daily balance in last one year* is .988 which is higher than that of the variable *average transaction per month (.382),* average daily balance is having more discriminating power than the average transaction per month. By careful examination, you may notice that the coefficient of average daily balance is the maximum, and in fact it is very large in comparison to other variables; hence, it may be concluded that this is the most important variable in taking a decision to issue or not to issue the golden visa card.

Remark: You may run the discriminant analysis on the same data by using the option "Use stepwise method" in order to ascertain the fact, whether the variable *average daily balance* gets selected in the model. You may also check as to how much accuracy is reduced in the model if some of the independent variables are dropped from the model.

6. One of the purposes of running discriminant analysis in this example was to develop a decision model for classifying a customer into two categories, that is, card issued and card denied. Table 12.8 shows the means for the transformed group centroids. Thus, the new mean for group 1 (card denied) is −1.354 and for group 2 (card issued) is +1.354. This indicates that the midpoint of these two is 0. These two means can be plotted on a straight line by locating the midpoint as shown in Fig. 12.6.

Figure 12.6 defines the decision rule for classifying any new customer into any of the two categories. If the discriminant score of any customer falls to the right of the midpoint ($Z > 0$), he/she is classified into the card issue category, and if it falls to the left of the midpoint ($Z < 0$), he/she is classified into card denied category.

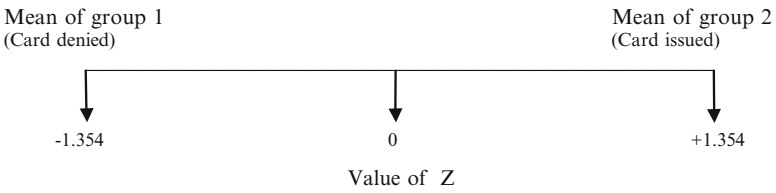


Fig. 12.6 Means of the transformed group centroids

Summary of the SPSS Commands for Discriminant Analysis

(i) After preparing the data file, follow the below-mentioned command sequence for discriminant analysis:

Analyze ⟶ **Classify** ⟶ **Discriminant**

(ii) Select the dependent variable *Card_Decision* from left panel to the "Grouping Variables" section of the right panel and define its minimum and maximum range as "1" and "2." Further, select all independent variables from the left panel to the "Independents" section of the right panel. Check the option "Enter independents together."
(iii) Click the tag **Statistics** and check options for "Means," "Fisher's," and "Unstandardized" in it. Click ***Continue***.
(iv) Click the tag **Classify** and check option for "Summary table." Press ***Continue***.
(v) Press **OK** for output.

Exercise

Short Answer Questions

Note: Write answer to each of the following questions in not more than 200 words.

Q.1. Explain a research situation where discriminant analysis can be used and discuss its utility.
Q.2. In discriminant analysis, what dependent variable refers to? What is the data type of dependent variable in SPSS?
Q.3. Discuss situations in which the discriminant analysis uses the two different methods like "*Enter independents together*" and "*Use stepwise method*" for developing the discriminant model.
Q.4. What do you mean by discriminating variables? What is its significance in discriminant analysis?
Q.5. What is the significance of Box's M test in discriminant analysis? What does the magnitude of Box's M signify?
Q.6. What do you mean by eigenvalues? Explain its importance.
Q.7. Explain the significance of canonical correlation. What does it convey?

Q.8. Explain the role of Wilks' lambda in discriminant analysis. Comment on the models if its values are 0, 0.5, and 1 in three different situations.
Q.9. Explain the purpose of classification matrix in discriminant analysis. How the percentage of correct classification is similar to R^2?
Q.10. What is discriminant function and how it is developed? How this function is used in decision-making?
Q.11. One of the conditions in discriminant analysis is that "All variables have linear and homoscedastic relationships." Explain the meaning of this statement.
Q.12. What do you mean by the discriminating power of the variables? How will you asses it?

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers. Tick mark the one that you consider the closest to the correct answer.

1. In discriminant analysis, independent variables are treated as

   (a) Scale
   (b) Nominal
   (c) Ordinal
   (d) Ratio

2. In discriminant analysis, dependent variable is measured on the scale known as

   (a) Grouping
   (b) Ordinal
   (c) Nominal
   (d) Criterion

3. Discriminant function is also known as

   (a) Eigenvalue
   (b) Regression coefficient
   (c) Canonical root
   (d) Discriminant coefficient

4. Confusion matrix is used to denote

   (a) Correctly classified cases
   (b) Discriminant coefficients
   (c) F-values
   (d) Robustness of different models

5. The decision criteria in discriminant analysis are as follows:
   Classify in first group if $Z < 0$
   Classify in second group if $Z > 0$
   The above criteria hold true

   (a) If size of the samples in both the groups are equal
   (b) If size of the samples in both the groups are nearly equal
   (c) If size of the samples in both the groups are in the proportion of 4:1
   (d) In all the situations

6. In stepwise method of discriminant analysis, a variable is included in the model if it is found significant at

   (a) 2% level
   (b) 1% level
   (c) 10% level
   (d) 5% level

7. The Wilks' lambda indicates

   (a) Percentage variability in the two groups explained by the model
   (b) Robustness of the model
   (c) Proportion of total variability not explained by the discriminant model
   (d) Significance of discriminant coefficients

8. One of the assumptions in discriminant analysis is that

   (a) All variables have curvilinear and homoscedastic relationships.
   (b) All variables have linear and non-homoscedastic relationships.
   (c) All variables have curvilinear and non-homoscedastic relationships.
   (d) All variables have linear and homoscedastic relationships.

9. Correct sequence of commands in SPSS for discriminant analysis is

   (a) Analyze → Discriminant → Classify
   (b) Analyze → Classify → Discriminant
   (c) Discriminant → Analyze → Classify
   (d) Discriminant → Classify → Analyze

10. Value of Wilks' lambda ranges from

    (a) −1 to +1
    (b) 0 to 1
    (c) −1 to 0
    (d) −2 to 2

11. Discriminant function is developed on the basis of

    (a) Standardized coefficients
    (b) Unstandardized coefficients
    (c) Classification matrix
    (d) Functions at group centroids

12. The power of discrimination of an independent variable is determined by

    (a) Unstandardized canonical coefficients
    (b) Wilks' lambda
    (c) Standardized canonical coefficients
    (d) Eigenvalues

13. In explorative discriminant analysis,

    (a) All the independent variables are taken in the study.
    (b) Only those variables which are known to have sufficient discriminating power are taken in the study.
    (c) Maximum number of relevant independent variables are taken in the study.
    (d) It is up to researcher to identify the independent variables in the study.

14. Choose the correct statement in discriminant analysis.

    (a) Dependent variable is an ordinal variable.
    (b) The groups should not be mutually exclusive.
    (c) Sample sizes should differ to a great extent.
    (d) No independent variables should have a zero variability in either of the groups formed by the dependent variable.

15. In discriminant analysis, the square of the canonical correlation is an indicator of

    (a) Relationship among the independent variables
    (b) Efficiency of the predictor variables
    (c) Discriminating power of the independent variables
    (d) The percentage of variability explained by the predictor variables in the two groups

Assignments

1. A study was conducted to know the variables responsible for selection in the bank probationary officers examination. Thirty candidates who appeared in the examination were identified for the study, and following data were obtained on them.

Results of the examination and subject's profile

S.N.	Bank examination result	IQ	English	Numerical aptitude	Reasoning
1	Successful	78	56	65	78
2	Successful	76	76	76	89
3	Not successful	74	52	63	93
4	Not successful	65	49	62	90
5	Successful	83	71	82	85
6	Successful	79	80	86	84
7	Not successful	91	54	52	89
8	Not successful	64	65	53	84
9	Not successful	53	69	54	85
10	Successful	60	78	75	92
11	Not successful	65	69	63	83
12	Successful	86	73	83	83
13	Not successful	53	65	67	83
14	Successful	74	69	80	78
15	Successful	60	68	81	74
16	Successful	75	75	78	85

(continued)

(continued)

Results of the examination and subject's profile

S.N.	Bank examination result	IQ	English	Numerical aptitude	Reasoning
17	Not successful	56	73	75	83
18	Not successful	65	64	56	84
19	Not successful	56	58	64	86
20	Successful	95	68	78	82
21	Successful	92	80	74	83
22	Not successful	45	73	71	91
23	Successful	85	56	89	74
24	Successful	68	45	83	85
25	Not successful	64	73	64	84
26	Not successful	70	71	56	86
27	Successful	78	74	84	94
28	Not successful	64	70	55	86
29	Not successful	42	67	51	76
30	Successful	82	67	90	83

Develop a discriminant model. Test the significance of the developed model and find the relative importance of the independent variables in the model. Compare the efficiency of the two discriminant function models obtained by taking all the variables at once and stepwise methods.

2. A branded apparel company wanted to reward its loyal customers by means of incentives in the form of 60% discount in the first week of New Year. The company had a loose policy of identifying a customer into loyal or disloyal on the basis of certain criterion which was more subjective. However, the management was interested to develop a more scientific approach to build up a model of classifying a customer into loyal and disloyal group. A sample of 30 customers were chosen from the database, and their purchase details were recorded which are shown in the following table:
   Apply discriminant analysis to build up a classification model which can be used for the existing and future customers to reward as per the company policy. Test

Purchase data of the customers of the apparel company

S. N.	Customer classification	No. of purchases/ year in a year	Purchase amount in a year	No. of kids' wear apparel/year	No. of ladies apparel/year	No. of gents
1	Loyal	6	109,870	23	12	2
2	Disloyal	8	27,000	4	8	18
3	Loyal	11	135,000	22	23	11
4	Loyal	15	12,340	12	5	4
5	Disloyal	9	54,000	20	23	8
6	Disloyal	4	34,000	12	8	20
7	Loyal	8	98,000	16	9	22
8	Loyal	8	80,002	23	25	3
9	Disloyal	4	71,000	25	15	19
10	Loyal	8	180,000	35	24	12

(continued)

(continued)

S. N.	Customer classification	No. of purchases/ year in a year	Purchase amount in a year	No. of kids' wear apparel/year	No. of ladies apparel/year	No. of gents
11	Disloyal	6	34,012	3	2	15
12	Loyal	12	67,000	12	8	5
13	Loyal	5	92,008	20	12	9
14	Disloyal	4	12,000	6	2	8
15	Loyal	10	71,540	6	15	8
16	Disloyal	4	13,450	1	2	15
17	Loyal	14	125,000	24	15	8
18	Loyal	20	80,000	5	20	7
19	Disloyal	5	56,021	15	10	15
20	Loyal	9	170,670	21	25	12
21	Disloyal	6	1,012	1	1	1
22	Disloyal	7	54,276	13	8	15
23	Loyal	15	100,675	25	25	5
24	Loyal	12	106,750	30	15	4
25	Disloyal	11	3,500	2	2	3
26	Disloyal	5	2,500	2	1	3
27	Loyal	10	89,065	14	21	8
28	Loyal	9	80,540	15	19	16
29	Disloyal	7	12,000	4	4	6
30	Disloyal	3	5,056	4	2	3

the significance of discriminant function, explain the percentage of correct classification by the model, and discuss the relative importance of independent variable. Find out the percentage of variability explained by the discriminant model in both the situations when all the variables are included in the model and when the variables are identified using stepwise procedure.

Hint: In Assignment 2, since the number of scores in loyal and disloyal customer groups are not same, you may not get mean of Z as 0. In this example, you will get the new mean for group 1 (Disloyal group) as -1.603 and new mean for group 2 (Loyal group) as 1.403. Thus, midpoint of these groups would be $-.1$ instead of 0. A customer would be classified as disloyal or loyal depending upon $Z < -.1$ or $Z > -.1$.

Answers to Multiple-Choice Questions

Q.1 a Q.2 c
Q.3 c Q.4 a
Q.5 a Q.6 d
Q.7 c Q.8 d
Q.9 b Q.10 b
Q.11 b Q.12 c
Q.13 c Q.14 d
Q.15 d