EXPERIMENT 6

CO3: To perform data collection and initial data handling by managing data structures and measurement levels.

Aim: To perform Discriminant Analysis using 10 sample values and analyze the classification of data using Origin Pro 2024 software.

Objective:

- To classify a set of observations into predefined groups.
- To examine how well the discriminant function distinguishes between the groups.
- To assess the accuracy of classification based on a set of predictor variables.

Theory:

Discriminant Analysis is a statistical technique used to classify a set of observations into predefined groups based on several continuous or binary predictor variables. It creates a discriminant function, which is a linear combination of the predictor variables that best separates the groups.

Discriminant Analysis relies on constructing one or more linear combinations of predictor variables that can best separate two or more groups. The mathematical formulation of Discriminant Analysis primarily involves **Linear Discriminant Analysis (LDA)** and **Quadratic Discriminant Analysis (QDA)**. Below is an explanation using mathematical concepts.

1. Linear Discriminant Function

The main objective of Discriminant Analysis is to find a function $D(x)$ that classifies a given observation $x = (x_1, x_2, \ldots, x_p)$ into one of the groups (Group A or Group B). For Linear Discriminant Analysis (LDA), the discriminant function $D(x)$ is expressed as:

$$D(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

where:

- w_0 is the intercept or constant.
- $w_1, w_2, \ldots, w_p$ are the weights or coefficients associated with each predictor variable $x_1, x_2, \ldots, x_p$.
- $x_1, x_2, \ldots, x_p$ are the values of the predictor variables for an observation.

The discriminant function $D(x)$ helps assign the observation to one of the predefined groups based on the values of $x_1, x_2, \ldots, x_p$.

2. Discriminant Rule

For two groups (e.g., Group A and Group B), the discriminant function for each group is computed, and the observation is classified into the group for which the function value is largest. The decision rule can be mathematically represented as:

$$\text{Classify } x \text{ into Group A if } D_A(x) > D_B(x)$$

or

$$\text{Classify } x \text{ into Group B if } D_B(x) > D_A(x)$$

Where:

- $D_A(x)$ is the discriminant function for Group A.
- $D_B(x)$ is the discriminant function for Group B.

3. Estimation of Coefficients (Weights)

The coefficients $w_1, w_2, \ldots, w_p$ in the discriminant function are estimated based on the following steps:

- **Pooled Covariance Matrix (S):** The covariance matrix S is computed as a weighted average of the within-group covariance matrices from the training data (i.e., the data used to estimate the discriminant function).

  $$S = \frac{1}{N-g} \sum_{k=1}^{g} (n_k - 1) S_k$$

  Where:

  - g is the number of groups.
  - N is the total number of observations.
  - n_k is the number of observations in group k.
  - S_k is the covariance matrix for group k.
- **Mean Vectors (μ_A, μ_B):** The mean vector μ for each group is calculated as the mean of the predictor variables for all observations in the group. For example, for Group A and Group B:

  $$\mu_A = \frac{1}{n_A} \sum_{i=1}^{n_A} x_i$$

  $$\mu_B = \frac{1}{n_B} \sum_{i=1}^{n_B} x_i$$

  Where n_A and n_B are the number of observations in Group A and Group B, respectively.

- **Discriminant Coefficients (Weights):** The coefficients for the discriminant function are calculated as:

$$w = S^{-1}(\mu_A - \mu_B)$$

This equation shows that the coefficients depend on the difference between the mean vectors of the groups and the inverse of the pooled covariance matrix.

4. Mahalanobis Distance

Discriminant Analysis uses a distance measure called **Mahalanobis Distance** to compute the separation between the groups. The Mahalanobis distance between a sample x and the group mean μ is:

$$d_M(x, \mu) = \sqrt{(x - \mu)^T S^{-1}(x - \mu)}$$

Where:

- x is a row vector of the predictor values for an observation.
- μ is the mean vector for the group.
- S^{-1} is the inverse of the pooled covariance matrix.

A smaller Mahalanobis distance indicates that the observation is closer to the group mean, and hence, it is more likely to belong to that group.

5. Classification Accuracy:

In the classification table, the accuracy of the discriminant function is shown. This table summarizes the predicted group membership versus the actual group membership for the samples.

The **misclassification rate** is calculated as:

$$\text{Misclassification Rate} = \frac{\text{Number of Misclassified Observations}}{\text{Total Number of Observations}}$$

For example, from the table in the procedure:

Group	Actual Count	Predicted Group A	Predicted Group B	Misclassification Rate
Group A	5	4	1	0.20
Group B	5	1	4	0.20
Total	10			0.20

This shows that:

- **Group A:** Out of 5 observations, 1 was misclassified as belonging to Group B.
- **Group B:** Out of 5 observations, 1 was misclassified as belonging to Group A.
- Overall, the misclassification rate is $\frac{2}{10} = 0.20$ or 20%.

6. Eigenvalues and Wilks' Lambda

- **Eigenvalues:** In LDA, eigenvalues represent the ratio of between-group variance to within-group variance for each discriminant function. Larger eigenvalues indicate a greater separation between the groups.
- **Wilks' Lambda (Λ):** It is used to test the null hypothesis that the means of the groups on all predictor variables are equal. It is computed as:

$$\Lambda = \frac{|\text{Within-Group Covariance Matrix}|}{|\text{Total Covariance Matrix}|}$$

A smaller Wilks' Lambda (close to 0) indicates that the discriminant function explains a significant proportion of variance between the groups, and the groups are well-separated. Statistical tests (like the chi-square test) are applied to assess the significance of Λ.

The discriminant function constructed during the analysis is used to classify observations into predefined groups. Based on the discriminant scores and the classification table, the function's accuracy is evaluated.

Procedure:

1. **Prepare the Data:**
   - Collect 10 sample values with corresponding group labels (e.g., Group A and Group B).
   - The sample data should consist of predictor variables and a categorical response variable (group).

Example Data (Tabular Form):

Sample No.	Predictor 1	Predictor 2	Group
1	2.5	3.1	A
2	4.2	2.9	A
3	3.1	4.3	B
4	2.7	3.0	B
5	5.1	1.9	A
6	3.6	3.4	B
7	4.9	2.7	A
8	3.3	4.1	B
9	4.5	2.5	A
10	2.9	3.7	B

2. **Open Origin Pro 2024:**

   Launch Origin Pro 2024 software on your computer.

3. **Data Entry:**

   a. Enter the sample data (predictors and groups) into the Origin Pro worksheet.
   b. Ensure that each predictor is entered in a separate column, and the group labels are in a separate column.

4. **Select Analysis Tool:**

   a. Go to the ***Statistics*** menu.
   b. Select **Multivariate Analysis** and then choose **Discriminant Analysis** from the drop-down menu.

5. **Set Parameters:**

   a. In the dialog box, select the predictor variables (e.g., Predictor 1 and Predictor 2) and the group variable (e.g., Group A, Group B).
   b. Choose whether to use Linear or Quadratic Discriminant Analysis (LDA or QDA) based on the data characteristics.
   c. Set any other relevant options like cross-validation or prior probabilities, if necessary.

6. **Run the Analysis:**

   a. Click on **OK** to run the analysis.
   b. Origin Pro will compute the discriminant function(s) and classify the sample values into their respective groups.

7. **View Results:**

   The results will be displayed, including:

   - The discriminant function.
   - Classification table (how well the function classified the sample values).
   - Group centroids.
   - Eigenvalues and Wilks' Lambda for significance testing.

8. **Plot the Discriminant Function:**

   a. Plot the discriminant function by selecting the appropriate graph from the `Graph` menu.
   b. Visualize how well the samples are classified into groups.

Results:

Based on the above data and analysis, the results will typically include:

- **Discriminant Functions:** Linear combination of the predictor variables that best separate Group A and Group B.
- **Classification Table:** Shows how accurately the function classified the samples into the correct groups.
- **Group Centroids:** The mean of the discriminant function values for each group.

- **Significance Testing:** Wilks' Lambda and other statistical tests to assess the discriminative power of the function.

Example Output (Classification Table):

Group	Actual Count	Predicted Group A	Predicted Group B	Misclassification Rate
Group A	5	4	1	0.20
Group B	5	1	4	0.20
Total	10			0.20

Learning Outcomes:

- Understanding the basic principles of Discriminant Analysis.
- Ability to classify observations based on multiple predictor variables.
- Familiarity with using Origin Pro software to perform multivariate statistical analysis.
- Interpretation of discriminant function results, including classification accuracy and group separation.
- Knowledge of how to visualize discriminant analysis outcomes for further insights.