**REPUBLIQUE DU BENIN**

🙡🙡🙡

**UNIVERSITE DE PARAKOU (UP)**

🙡🙡🙡🙡🙡

**ECOLE NATIONIONALE DE LA STATISTIQUE DE LA PLANIFICATION ET DE LA DEMOGRAPHIE (ENSPD)**

**Master 1**

**COURS :** Introduction aux logiciels statistiques

R-STATS-SPSS



**Membres du groupe : 27**

1. MAMA Abdou Moukadas
2. TAWAYE DAN DOSSA Issifou

**Chargé du cours :**

Dr Ir. Epiphane SODJINOU

Agroéconomiste, Biostatisticien

(Maitre de conférences)

**THEME : L’analyse discriminante avec le logiciel R**

ANNEE ACADEMIQUE: 2024 -2025

Table des matières

[1](#_Toc199758505)

[Introduction 3](#_Toc199758506)

[1. Cadre théorique de l’analyse discriminante 3](#_Toc199758507)

[1.1. Principaux types d’analyse discriminante 3](#_Toc199758508)

[1.2. Autres Types d’analyse discriminante 4](#_Toc199758509)

[1.3. Hypothèses de l’analyse discriminante 5](#_Toc199758510)

[2. Démarche de mise en œuvre 5](#_Toc199758511)

[1) **Préparation des données** 5](#_Toc199758512)

[2) **Test des hypothèses** 5](#_Toc199758513)

[3) **Choix du type d’analyse** 6](#_Toc199758514)

[4) **Calcul des fonctions discriminantes** 6](#_Toc199758515)

[5) **Validation du modèle** 6](#_Toc199758516)

[6) **Interprétation des coefficients et des scores** 6](#_Toc199758517)

[7) **Visualisation graphique** 6](#_Toc199758518)

[3. Avantages et limites de l’analyse discriminante 7](#_Toc199758519)

[a. Avantages : 7](#_Toc199758520)

[b. Limites : 7](#_Toc199758521)

[Conclusion 7](#_Toc199758522)

[Cas Pratique 8](#_Toc199758523)

## Introduction

L’analyse discriminante est une méthode statistique multivariée permettant de classer des individus dans des groupes préalablement définis, en se basant sur des variables explicatives quantitatives. Elle est généralement utilisée en sciences sociales, médicales, biologiques et marketing pour étudier les différences entre groupes et prédire l'appartenance d'un nouvel individu. L’objectif de ce travail est de présenter de manière détaillée l’analyse discriminante, ses variantes, son cadre théorique, ses applications pratiques, et de l’illustrer avec un exemple concret sous le logiciel R.

## 1. Cadre théorique de l’analyse discriminante

L’analyse discriminante repose sur l’hypothèse que les individus peuvent être regroupés dans des classes connues, sous le contrôle des variables quantitatives pour expliquer cette classification. Elle cherche à déterminer les fonctions discriminantes, qui sont des combinaisons linéaires des variables explicatives, maximisant la séparation entre les groupes. A l’aide cette analyse, on peut faire une classification des groupes des individus ou objet. On distingue deux principaux types d’analyse discriminante.

## 1.1. Principaux types d’analyse discriminante

**A. LDA (Linear Discriminant Analysis)** :

**LDA** est une méthode statistique utilisée pour classifier des observations statistiques utilisée pour classifier des observations en groupes prédéfinis en fonction de variables de variables explicatives. Elle est particulièrement utile pour la réduction de dimension et la classification supervisée.

**Principes de LDA :**

* **Maximisation de la séparation des classes** : LDA cherche à projeter les les données dans un espace où les classes sont le mieux séparées.

**B. QDA (Quadratic Discriminant Analysis)** :

**QDA** est une méthode de classification statistique qui, contrairement à la **Linear Discriminant Analysis (LDA)**, permet à chaque classe d’avoir sa propre matrice de covariance. Cela signifie que QDA peut capturer des frontières de décision non linéaires, ce qui la rend plus flexible que LDA.

**Principes de QDA :**

* **Modélisation probabiliste** : QDA suppose que les données de chaque classe suivent une distribution normale multivariée avec une matrice de covariance propre.
* **Frontière quadratique** : Contrairement à LDA qui produit des frontières linéaires, QDA génère des frontières quadratiques, ce qui permet une meilleure séparation des classes lorsque les distributions diffèrent.
* **Application** : Utilisée en reconnaissance de formes, classification d’images et analyse biomédicale.

**Différence entre QDA et LDA :**

* **LDA** : Suppose une matrice de covariance identique pour toutes les classes → frontière linéaire.
* **QDA** : Permet des matrices de covariance différentes → frontière quadratique, plus flexible.

Les fonctions discriminantes obtenues peuvent être utilisées à des fins descriptives (comprendre les différences entre groupes) et prédictives (prédire l’appartenance d’un nouvel individu).

## 1.2. Autres Types d’analyse discriminante

* **Analyse discriminante canonique (ADC)** : Utilisée pour maximiser la séparation entre les groupes en trouvant des combinaisons linéaires optimales des variables explicatives. Elle est souvent employée en analyse multivariée.
* **Analyse discriminante pas à pas** : Sélectionne progressivement les variables les plus pertinentes pour la classification en fonction de critères statistiques. Elle est utile lorsque l’on veut réduire le nombre de variables explicatives.
* **Analyse discriminante sur variables qualitatives ou mixtes** : Adaptée aux données catégorielles ou mixtes (quantitatives et qualitatives). Elle utilise des méthodes comme l’**analyse factorielle des correspondances** ou des techniques spécifiques comme **DISQUAL**.

Chaque type répond à des besoins spécifiques selon la structure des données et les hypothèses statistiques

## 1.3. Hypothèses de l’analyse discriminante

* **Indépendance des observations** Chaque observation doit être indépendante des autres pour éviter les biais dans l'estimation des paramètres.
* **Normalité multivariée des variables explicatives** Les variables doivent suivre une distribution normale dans chaque groupe pour garantir la fiabilité des scores discriminants.
* **Homogénéité des matrices de covariance**
* **LDA** : Suppose que les matrices de covariance sont identiques entre les groupes, ce qui permet de générer des frontières **linéaires**.
* **QDA** : Autorise des matrices de covariance différentes, ce qui permet des **frontières quadratiques** plus flexibles.
* **Absence de multi colinéarité excessive** Les variables explicatives ne doivent pas être trop corrélées entre elles pour éviter les problèmes de stabilité du modèle et garantir des coefficients interprétables.

## 2. Démarche de mise en œuvre

1. **Préparation des données**

* Nettoyage des valeurs aberrantes et des données manquantes.
* Codage des variables catégorielles si nécessaire.
* Normalisation des variables quantitatives pour assurer une distribution cohérente.

1. **Test des hypothèses**

* Vérification de la **normalité multivariée** des variables explicatives (test de Shapiro-Wilk, histogrammes, Q-Q plots).
* Contrôle de l’**homogénéité des matrices de covariance** via les tests de Box ou Levene.

1. **Choix du type d’analyse**

* **LDA** si les matrices de covariance sont homogènes.
* **QDA** si les matrices de covariance sont différentes entre groupes.

1. **Calcul des fonctions discriminantes**

* Estimation des coefficients discriminants pour chaque variable.
* Détermination des scores discriminants pour classer les observations.

1. **Validation du modèle**

* Utilisation de la **cross-validation** pour mesurer la robustesse du modèle.
* Analyse de la **matrice de confusion** pour évaluer la qualité de la classification.

1. **Interprétation des coefficients et des scores**

* Identification des variables les plus discriminantes.
* Analyse de la signification statistique des coefficients.

1. **Visualisation graphique**

* Représentation des individus dans l’espace discriminant via des **biplots** ou des **graphes de dispersion**.
* Utilisation de **cercles de corrélation** pour visualiser les contributions des variables.

# 3. Avantages et limites de l’analyse discriminante

## a. Avantages :

* Interpretable et mathématiquement rigoureuse
* Bonne performance en classification supervisée avec peu de variables
* Adaptée aux petits échantillons

## b. Limites :

* Sensible aux violations des hypothèses (normalité, homoscédasticité)
* Moins performante avec des données fortement non linéaires
* Moins flexible que les méthodes modernes (arbres, forêts, SVM)

## Conclusion

L’analyse discriminante reste une méthode puissante pour comprendre et prédire l’appartenance à un groupe à partir de variables explicatives. En dépit de certaines limitations liées aux hypothèses statistiques, elle demeure pertinente dans de nombreux domaines. L’exemple sur le jeu de données iris montre son efficacité et sa simplicité d’implémentation sous R. Son usage combiné avec des outils de visualisation permet de mieux comprendre la structure des données et d’aider à la prise de décision basée sur les données.

## Cas Pratique

#Cas pratique : Classification d’espèces de fleurs avec LDA Nous allons utiliser le célèbre jeu de données Iris, qui contient des mesures de fleurs de trois espèces différentes (Setosa, Versicolor, Virginica). L’objectif est de classifier les fleurs en fonction de leurs caractéristiques.

## Chargement du jeu de données Iris

data(iris)

head(iris) # Aperçu des premières lignes

#Interprétation : On observe cinq colonnes : quatre variables quantitatives (longueur et largeur des sépales et pétales) et une variable qualitative (Species).

## Préparation des données

# Vérification des valeurs manquantes

sum(is.na(iris)) # Doit être 0

# Séparation des variables explicatives et de la variable cible

X <- iris[, 1:4] # Variables explicatives

Y <- iris$Species # Variable cible

#Interprétation : Il n’y a pas de valeurs manquantes, et nous avons bien séparé les variables explicatives et la variable à prédire.

## Vérification des hypothèses

# Test de normalité des variables explicatives

shapiro.test(X[,1]) # Exemple sur la première variable

#Interprétation : Si la p-valeur est supérieure à 0.05, on peut considérer que la variable suit une distribution normale.

# Vérification de l’homogénéité des matrices de covariance

library(MASS)

boxM(X, Y)

#Interprétation : Si la p-valeur est inférieure à 0.05, cela signifie que les matrices de covariance diffèrent entre les groupes, ce qui pourrait favoriser QDA plutôt que LDA.

##Application de l’analyse discriminante linéaire (LDA)

# Modèle LDA

lda\_model <- lda(Species ~ ., data = iris)

print(lda\_model)

#Interprétation : On obtient les coefficients discriminants, qui indiquent l’importance de chaque variable dans la classification.

##Prédiction et validation du modèle

# Prédiction sur les données

predictions <- predict(lda\_model)

table(predictions$class, iris$Species) # Matrice de confusion

#Interprétation : La matrice de confusion montre combien de fleurs ont été correctement classées. Un taux de classification élevé indique un bon modèle.

##Visualisation des résultats

# Graphique des scores discriminants

library(ggplot2)

df <- data.frame(predictions$x, Species = iris$Species)

ggplot(df, aes(x = LD1, y = LD2, color = Species)) + geom\_point()

#Interprétation : Ce graphique permet de visualiser la séparation des groupes dans l’espace discriminant.

##Conclusion

#L’analyse discriminante linéaire permet de classifier efficacement les espèces de fleurs en fonction de leurs caractéristiques. Si les hypothèses de normalité et d’homogénéité des variances ne sont pasrespectées, QDA pourrait être une meilleure alternative.