

Classification supervisée

Présentation générale et étude préliminaire

V. Vandewalle (vincent.vandewalle@univ-lille.fr)

C. Preda (cristian.preda@univ-lille.fr)

Polytech'Lille GIS4

Année universitaire 2019-2020

Organisation du module

Chapitres

- Présentation générale
- Analyse discriminante
- Evaluation d'une règle de classement
- Régression logistique

Volume 8h CM, 2h TD, 10h TP

Logiciel utilisé R

Evaluation Devoir sur table de 2h

- 1 Panorama
- 2 Principe général en classification supervisée
- 3 Déroutement d'une étude de data mining
- 4 Exploration et préparation des données
- 5 MANOVA

Data mining

La classification supervisée peut être replacée dans le domaine plus général du data-mining.

Data mining = exploration de données, extraction de connaissances à partir de données

- exploration et *analyse de bases de données* souvent grandes, stockées informatiquement
- détection dans ces données de *règles*, d'*associations*, de *tendances* inconnues ou cachées, de *structures* particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données.
- besoin d'outils en statistique et en informatique

Domaines d'applications variés

- **Banque** : modèles score d'octroi de crédit pour des montants unitaires modérés et un grand nombre de dossiers portant un caractère standard
- **Assurances** : adaptation de la tarification au risque
- **Grande distribution** : cartes fidélité, bases de données initialement développées pour des besoins de gestion ensuite utilisées à des fins décisionnelles
- **e-commerce, marketing direct** : publicités personnalisées, score d'appétence
- **Contrôle qualité, gestion des risques industriels** : recherche de l'origine de défauts
- **Pilotage de la production** : prévision de la demande

Domaines d'applications variés

- Prévention du terrorisme, détection de fraude dans l'utilisation de téléphonie mobile ou de carte bancaire
- Enquêtes en sciences humaines
- Etudes biologiques, médicales, pharmaceutiques : recherche de facteurs de risque ou de décès dans certaines pathologies
- Etudes agro-alimentaires : recherche d'ingrédients qui permettent d'améliorer le goût d'un produit...
- ...

Aide à la décision : finalité du data mining et de la statistique

La statistique ne doit pas seulement (uniquement) aider à *comprendre le réel* en le modélisant, mais doit aussi aider à *prédire*.

Autrement dit, l'essentiel n'est pas forcément de savoir comment cela fonctionne (science) mais plutôt que cela fonctionne (ingénierie).

Exemple : En médecine certains traitements ont été mis en place sur la base d'analyses statistiques sans comprendre en détail le mécanisme biologique.

Le data mining permet de *limiter la subjectivité humaine* dans le processus de *décision*, en combinant statistique et informatique pour traiter de plus en *plus rapidement* un grand nombre de dossiers.

Historique

- Fin du XIX^e jusqu'aux années 1950 : statistique classique
 - Quelques centaines d'individus
 - Quelques variables
 - Fortes hypothèses (linéarité, normalité, homoscedasticité, ...)
 - Utilisation en laboratoires
- Années 1960-1980 : apparition de l'informatique
 - Quelques milliers d'individus
 - Quelques dizaines de variables
 - Constructions de tableaux individus \times variables
 - Importance du calcul et de la représentation visuelle

Historique

- Années 1990 : avènement du concept de data mining
 - plusieurs millions ou dizaines de millions d'individus
 - plusieurs millions ou dizaines de millions de variables
 - nombreuses variables non numériques, parfois textuelles
 - faibles hypothèses sur les lois statistiques suivies
 - données recueillies avant l'étude, et souvent à d'autres fins
 - population constamment en évolution
 - présence d'individus « aberrants »
 - données imparfaites
 - utilisation en entreprise

« *Prédire n'est pas expliquer* »

Protection des données personnelles

L'évolution du data-mining soulève des questions réglementaires

Données à caractère personnel : données qui peuvent être directement ou indirectement rattachées à une personne physique particulière :

- état-civil
- autre identifiant : numéro de client ou d'assuré
- tout élément propre : voix, image, empreintes génétiques ou biométriques, adresse, ...

La plupart des pays ont adopté des législations pour encadrer la collecte, le stockage, le traitement et l'utilisation des données à caractère personnel et tout particulièrement les données sensibles.

Protection des données personnelles

Exemple de données sensibles :

- santé
- orientations sexuelles
- condamnations pénales
- origine raciale
- opinions politiques
- convictions religieuses

En France, nécessité de demander l'autorisation à la CNIL avant de lancer certaines études ou certains projets.

L'interconnexion de fichiers est particulièrement surveillée en raison d'un risque d'accumulation de données personnelles sur un individu insensibles isolément.

Principaux facteurs de succès d'un projet

- ❶ Des objectifs précis, importants et réalistes.
- ❷ La richesse, et surtout la qualité, des données collectées.
- ❸ La collaboration des compétences métiers et statistiques de l'établissement.
- ❹ Une bonne restitution des informations générées et une bonne intégration le cas échéant dans le système d'information.
- ❺ L'analyse des résultats et des retours de chaque utilisation du data mining pour l'utilisation suivante.

Mise en œuvre en entreprise

Une entreprise peut soit :

- externaliser totalement l'activité de data mining,
- sous-traiter l'essentiel de l'activité à un prestataire,
- développer elle-même ses modèles de data mining.

- 1 Panorama
- 2 Principe général en classification supervisée
- 3 Déroulement d'une étude de data mining
- 4 Exploration et préparation des données
- 5 MANOVA

Notations

Variables

- $Y \in \{1, 2, \dots, K\}$: variable qualitative à expliquer (souvent binaire : $K = 2$)
- $\mathbf{X} = (X_1, X_2, \dots, X_p)$: p variables quantitatives explicatives

Extension possible

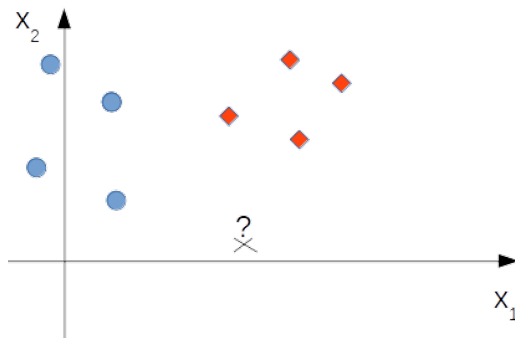
dans le cas où \mathbf{X} comporte des variables qualitatives.

Données d'apprentissage

$$\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

Ces données sont généralement supposées être issues de n réalisations indépendantes et identiquement distribuées (i.i.d.) du couple (\mathbf{X}, Y) .

Illustration



Objectif :

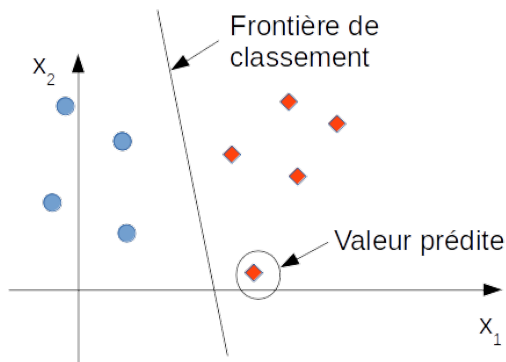
Pour un nouvel individu pour lequel on n'a observé que x , prédire la valeur de y associée avec la plus faible erreur possible.

Méthodologie générale

Objectif :

Apprendre une règle de classement r qui à tout $\mathbf{x} \in \mathbb{R}^p$ associe un $\hat{y} \in \{1, 2, \dots, K\}$:

$$r : \mathbf{x} \in \mathbb{R}^p \mapsto \hat{y} = r(\mathbf{x}) \in \{1, 2, \dots, K\}$$



De très nombreuses méthodes statistiques pour trouver r

Exemples de méthodes

- Analyse factorielle discriminante
- Analyse discriminante (linéaire et quadratique)
- Régression logistique
- k -plus proches voisins
- Arbre de régression
- ...

- 1 Panorama
- 2 Principe général en classification supervisée
- 3 Déroutement d'une étude de data mining
- 4 Exploration et préparation des données
- 5 MANOVA

Principales phases d'un projet de data mining

- 1 Définition des objectifs
- 2 Inventaire des données existantes
- 3 Collecte des données
- 4 Exploration et préparation des données
- 5 Segmentation de la population
- 6 Élaboration et validation des modèles prédictifs
- 7 Déploiement des modèles
- 8 Formation des utilisateurs des modèles
- 9 Suivi des modèles
- 10 Enrichissement des modèles

Définition des objectifs

- Choisir un sujet
- Définir la population cible : tous les clients, seulement les clients fidèles, tous les malades, seulement les malades curables par le traitement testé, ...
- Définir l'entité statistique étudiée : la personne, le foyer réduit aux conjoints, le foyer étendu aux enfants à charge, l'entreprise avec ou sans les filiales, ...
- Définir le phénomène à prédire
- Planifier le projet
- Prévoir l'utilisation opérationnelle des informations extraites et des modèles produits
- Spécifier les résultats attendus

Inventaire des données existantes

- Recenser les données utiles, accessibles (internes ou externes), légalement et techniquement exploitables, fiables et suffisamment à jour sur les caractéristiques et le comportement des individus étudiés : clients, patients, usagers, ...
- Récupérer les informations provenant du SI de l'entreprise, hors du SI ou calculées à partir des données précédentes (indicateurs, ratio, évolution au cours du temps, ...)
- Pour construire un modèle prédictif, il faut de plus rechercher un second type de données qui sont les historiques sur le phénomène à prédire (résultats d'expériences médicales, campagnes commerciales, ...)

Collecte des données

- Constituer la *base d'analyse* qui servira à la construction des modèles
- Créer une table (DB2, Orable, SAS, ...) ou un fichier (texte, CSV, ...) avec une ligne par individu et une colonne par variable
- Base d'analyse obtenue à partir de clichés pris à intervalles réguliers (mensuels, journaliers, ...) sur les données pendant une certaine durée
- Données souvent définies au niveau non pas de l'individu, mais du produit, du compte, de la facture, de l'examen médical, ...
- Il faut procéder à des synthèses selon plusieurs axes :
 - Axe *individu* en synthétisant au niveau de l'individu des données définies à un niveau plus fin
 - Axe *temps* en remplaçant des indicateurs définis à des instants différents, par un seul indicateur défini sur l'ensemble de la période

Collecte des données

Pour l'élaboration d'un modèle prédictif la base d'analyse comporte quatre types de variables :

- ❶ l'identifiant de l'individu
- ❷ la variable cible Y
- ❸ les variables explicatives X_1, X_2, \dots, X_p
- ❹ une variable "échantillon" indiquant s'il participe à l'élaboration (apprentissage) du modèle ou à son test

Exploration et préparation des données

- ❶ Fiabiliser, remplacer ou supprimer les données incorrectes
- ❷ Créer des indicateurs pertinents à partir des données brutes : ratio, évolution temporelle, recodage, montants moyens, récurrence, fréquence, ...
- ❸ Réduire le nombre d'individus : suppression des données aberrantes, échantillonnage, ...
- ❹ Réduire le nombre de variables : ignorer certaines variables trop corrélées entre-elles, ignorer des variables non discriminantes, rassembler plusieurs variables en une seule, réduire la dimension à l'aide d'ACP
- ❺ Réduire le nombre de modalités : regrouper les modalités qui sont trop nombreuses ou dont les effectifs sont trop petits, regrouper des modalités qui ont le même sens, discrétiser certaines variables continues

Segmentation de la population

Segmentation préalable de la population possible à partir de classification non supervisée (CAH, k-means, ...), puis conception d'un modèle de prédiction sur chacune des classes.

Élaboration et validation des modèles prédictifs

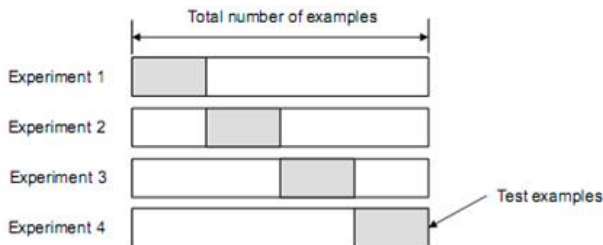
Si la taille de l'échantillon le permet on le découpe en 3 parties :

- **Echantillon d'apprentissage** ($\sim 70\%$) : on élabore différents modèles à partir de cet échantillon (analyse discriminante linéaire, regression logistique, ...)
- **Echantillon de validation** ($\sim 20\%$) : on retient le modèle qui produit les meilleurs performances sur l'échantillon de validation
- **Echantillon de test** ($\sim 10\%$) : cet échantillon permet de déterminer les performances réelles du modèle retenu

Élaboration et validation des modèles prédictifs

Si la taille est réduite on n'utilise pas d'échantillon test :

- Modèles élaborés à partir de tout l'échantillon
- Choix du meilleur modèle par validation croisée v -fold :
l'échantillon est partagé en v sous échantillons servant tour à tour d'échantillon de validation tandis que le reste est utilisé pour l'apprentissage, enfin on moyenne les performances sur les v sous échantillons



Dernières étapes du projet

- Déploiement des modèles
- Formation des utilisateurs des modèles
- Suivi des modèles
- Enrichissement des modèles

- 1 Panorama
- 2 Principe général en classification supervisée
- 3 Déroulement d'une étude de data mining
- 4 Exploration et préparation des données
- 5 MANOVA

Possibilité de conversion des données

Type de départ	Type d'arrivée	Opération	Principe
continu	discret	discrétisation	découpage de l'ensemble des valeurs en tranches
discret ou qualitatif	continu	ACM	on réalise l'ACM sur les variables qualitatives et on retient les premiers axes factoriels
qualitatif	binaire	binarisation	codage disjonctif complet des variables

Examen de la distribution des variables : 1D

Objectifs

- détecter d'éventuelles anomalies dans la distribution (en particulier valeurs extrêmes ou manquantes)
- se mettre quelques ordres de grandeur en tête (âge, revenu moyen de la population, ...), utiles dans la suite des analyses
- voir comment discrétiser les variables continues le cas échéant

Variables qualitatives

- fréquences de chaque modalité

Variables quantitatives

- boîtes à moustaches
- quantiles : 1%, 10%, 25%, 50%, 75%, 90% et 99%

Examen de la distribution des variables : 2D

Objectif

- détecter des incohérences entre variables
 - études des liaisons entre la variable cible et les variables explicatives : suppression des variables sans influence
 - suppression des variables explicatives redondantes
-
- Tableaux de contingence, profils ligne, profils colonne / Nuage de points
 - Test de liaison / Coefficient de corrélation linéaire

Examen de la distribution des variables : détection des valeurs rares ou manquantes

Pour les valeurs rares

- Supprimer les observations avec des valeurs rares
- Substituer une valeur rare par une valeur plus fréquente si cette substitution a un sens

Pour les valeurs manquantes

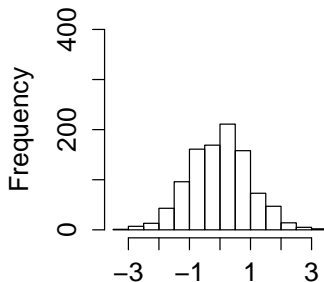
- Ne pas utiliser la variable si sa contribution au problème n'est pas essentielle, ou la remplacer par une variable proche mais sans valeurs manquantes
- Remplacer la valeur manquante par une valeur déterminée (méthodes d'imputation)
- Pour des variables qualitatives la "valeur" variable manquante peut être considérée comme une modalité à part entière

Examen de la distribution des variables : détection des valeurs rares ou manquantes

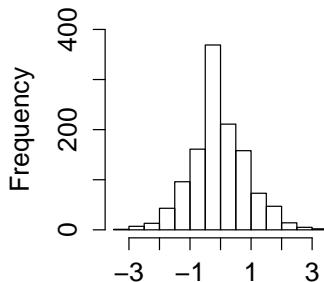
Remarque Les méthodes d'imputation simple peuvent conduire à une sous estimation de la variabilité, on peut leur préférer des méthodes d'imputation multiple qui remplacent les valeurs manquantes par plusieurs valeurs probables.

Examen de la distribution des variables : détection des valeurs rares ou manquantes

Avant imputation

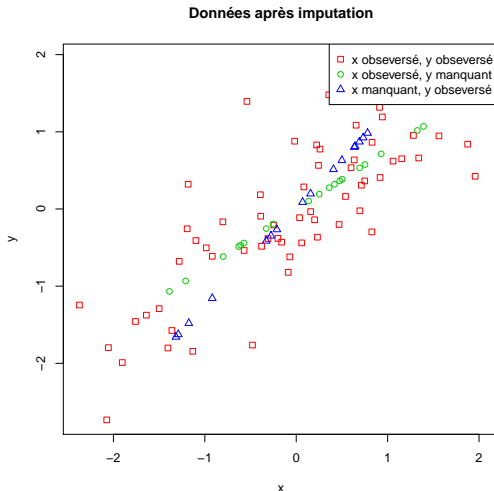


Après imputation



Risque d'une imputation par la moyenne, possibilité de faire une imputation par régression.

Examen de la distribution des variables : détection des valeurs rares ou manquantes



Détection des valeurs aberrantes

Types de valeurs aberrantes

- dates incohérentes, dates de naissances inconnues remplacées par des dates rondes (1900, 2000,...), date de souscription antérieure à la date de naissance
- clients déclarés comme particuliers alors qu'ils sont professionnels, montants saisis en centimes au lieu d'euros, ...

Traitement des valeurs aberrantes

- supprimer les observations si leur nombre n'est pas trop grand
- conserver l'observation mais écarter la variable
- conserver les observations et la variable, mais remplacer la valeur aberrante par une valeur manquante ou une autre valeur
- utiliser telle quelle la variable mais avoir cela en tête pour l'analyse

Détection des valeurs extrêmes

Les **valeurs extrêmes** sont des valeurs éloignées du centre de la distribution (typiquement à plus de 3 écarts-types de la moyenne). Ces valeurs peuvent affecter des méthodes telles que la régression logistique ou l'analyse discriminante linéaire.

Possibilités face à des valeurs extrêmes

- Les écarter de l'échantillon d'apprentissage (on peut les garder pour le test), éviter d'enlever plus que 1 à 2 % des données
- Découper la variable continue en classes de manière à neutraliser les valeurs extrêmes
- winsoriser la variable : remplacer les valeurs de la variable au-delà du 99^e centile par ce centile, tandis que les valeurs en deçà du premier centile sont remplacées par ce premier centile

Détection des valeurs les plus discriminantes

La plupart des méthodes de data mining fournissent de bien meilleurs résultats si on les applique à une liste de variables limitées aux plus pertinentes.

Pour tester ces liaisons on effectue les tests entre la variable cible et chacune des variables explicatives :

- Si la variable explicative est discrète ou discrétisée on s'appuie sur le χ^2
- Si la variable explicative est continue on s'appuie sur l'ANOVA

Détection des valeurs les plus discriminantes : variables explicatives qualitatives

Calcul du χ^2 , de la probabilité critique, du V de Cramer.

$$V = \frac{\chi^2}{\chi_{max}^2},$$

avec χ_{max}^2 la valeur maximale du χ^2 atteignable sur le tableau de I et J colonnes considéré.

$$\chi_{max}^2 = n \times \min(I - 1, J - 1),$$

avec n le nombre de données, I le nombre de modalités de la première variable, J le nombre de modalités de la seconde variable. $V \in [0; 1]$ donne une mesure absolue de l'intensité de la liaison entre 2 variables qualitatives, à préférer à la probabilité critique pour la sélection des variables pertinentes.

Détection des valeurs les plus discriminantes : variables explicatives qualitatives

Exercice 1 : calculer le V de Cramer du tableau suivant :

V2 \ V1	X	Y	TOTAL
A	50	0	50
B	0	25	25
C	0	25	25
TOTAL	50	50	100

Rappel :

$$k = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} = \sum_{i,j} \frac{O_{ij}^2}{T_{ij}} - n$$

Détection des valeurs les plus discriminantes : variables explicatives qualitatives

Exercice 2 : calculer le χ^2 et le V de Cramer des tableaux suivants :

V3 \ V1	X	Y	TOTAL
A	30	20	50
B	30	20	50
C	10	15	25
D	10	15	25
TOTAL	80	70	150

V4 \ V1	X	Y	TOTAL
A+B	60	40	100
C+D	20	30	50
TOTAL	80	70	150

Commenter. Est-ce que le même constat pourrait être fait pour le calcul de la probabilité critique ?

Refaire les calculs pour V4 en multipliant les effectifs par 100.

Détection des valeurs les plus discriminantes : variables explicatives quantitatives

Possibilité de réaliser une ANOVA à 1 facteur ou un test non-paramétrique.

Solution la plus simple : classer les variables explicatives de celle qui a la plus grande valeur de R^2 à celle qui a la valeur de R^2 la plus faible.

Exercice 3 : Pour $Y = 0$, on observe 5 valeurs de X : 4, 5, 7, 8, 9. Pour $Y = 1$, on observe 6 valeurs de X : 2, 3, 4, 6, 7, 8. Calculer le R^2 .

Détection des valeurs les plus discriminantes : variables explicatives quantitatives

Rappel ANOVA :

$$\underbrace{\sum_{i=1}^K \sum_{h=1}^{n_i} (x_{hi} - \bar{x})^2}_{SCT} = \underbrace{\sum_{i=1}^K \sum_{h=1}^{n_i} (x_{hi} - \bar{x}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^K \sum_{h=1}^{n_i} (\bar{x}_i - \bar{x})^2}_{SCF}$$

$$F = \frac{\frac{SCF}{K-1}}{\frac{SCR}{n-K}} \quad R^2 = \frac{SCF}{SCT}$$

- 1 Panorama
- 2 Principe général en classification supervisée
- 3 Déroulement d'une étude de data mining
- 4 Exploration et préparation des données
- 5 MANOVA**

Etude préliminaire : MANOVA

Remarque : Pour répondre à la question "la distribution de \mathbf{X} varie-t-elle selon le groupe?", on peut, en se limitant au test de l'égalité des espérances, utiliser une extension de l'anova :

MANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K = \mu$$

$$H_1 : \exists i \neq j \text{ tel que } \mu_i \neq \mu_j$$

où $\mu_i = E(\mathbf{X}/Y = i)$ et $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip})$, où $\mu_{ij} = E(X_j/Y = i)$, $i = 1, \dots, K$, $j = 1, \dots, p$.

Etude préliminaire : MANOVA

Conditions d'application :

- 1 $X = (X_1, \dots, X_p)$ est distribué selon une loi gaussienne dans chaque groupe $Y = i$:

$$X/Y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$$

- 2 Homogénéité des matrices de variances-covariances :

$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_K$$

MANOVA : ANOVA multivariée

$$\underbrace{V}_{\text{variance totale}} = \underbrace{B}_{\text{variance inter-groupes}} + \underbrace{W}_{\text{variance intra-groupes}}$$

Estimation : données

n unités statistiques (individus) tels que :

- Groupe 1 : n_1 ($Y = 1$)
- Groupe 2 : n_2 ($Y = 2$)
- \vdots
- Groupe K : n_K ($Y = K$)

avec $n_1 + n_2 + \dots + n_K = n$.

Soit X_{ijh} ($i = 1, \dots, K, j = 1, \dots, p, h = 1, \dots, n_i$), l'observation de la variable X_j sur l'individu h dans le groupe i .

Estimation : données

	X_1	\dots	X_j	\dots	X_p	Y
\vdots						1
\vdots						\vdots
\vdots						1
1						i
\vdots						\vdots
h						i
\vdots						\vdots
\vdots						K
\vdots						\vdots
\vdots						K

X matrice $n \times p$, Y vecteur $n \times 1$.

Estimation : espérances

$$\bar{X}_{ij} = \frac{1}{n_i} \sum_{h=1}^{n_i} X_{ijh}$$

est un estimateur de μ_{ij} et donc

$$\bar{X}_i = (\bar{X}_{i1}, \dots, \bar{X}_{ij}, \dots, \bar{X}_{ip})$$

est un estimateur de μ_i .

Moyennes regroupées dans un tableau $K \times p$, noté G

	X_1	\dots	X_j	\dots	X_p
$Y = 1$	\bar{X}_{11}		\bar{X}_{1j}		\bar{X}_{1p}
\vdots	\vdots		\vdots		\vdots
$Y = i$	\bar{X}_{i1}		\bar{X}_{ij}		\bar{X}_{ip}
\vdots	\vdots		\vdots		\vdots
$Y = K$	\bar{X}_{K1}		\bar{X}_{Kj}		\bar{X}_{Kp}

Estimation : espérances

$$\bar{X} = \sum_{i=1}^K \frac{n_i}{n} \bar{X}_i$$

la moyenne globale de X , estimateur de

$$\mu = E(X_1, \dots, X_p).$$

$$\bar{X} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$$

où

$$\bar{X}_j = \sum_{i=1}^K \frac{n_i}{n} \bar{X}_{ij} = \frac{1}{n} \sum_{i=1}^K \sum_{h=1}^{n_i} X_{ijh}$$

estimateur de la moyenne de la variable X_j sans la connaissance du groupe.

Estimation : variances

- W_i : matrice de variance-covariance dans le groupe $Y = i$

$$W_i[j, j'] = \text{cov}(X_j, X_{j'})_{/Y=i} = \frac{1}{n_i} \sum_{h=1}^{n_i} (X_{ijh} - \bar{X}_{ij})(X_{ij'h} - \bar{X}_{ij'})$$

- W : matrice de variance-covariance **intra-groupes**

$$W = \sum_{i=1}^K \frac{n_i}{n} W_i.$$

- B : la matrice de variance-covariance **inter-groupes**

$$B[j, j'] = \sum_{i=1}^K \frac{n_i}{n} (\bar{X}_{ij} - \bar{X}_j)(\bar{X}_{ij'} - \bar{X}_{j'})$$

- V : matrice de variance-covariance du tableau X :
variance-covariance **totale**

$$V[j, j'] = \text{cov}(X_j, X_{j'}) = \frac{1}{n} \sum_{i=1}^K \sum_{h=1}^{n_i} (X_{ijh} - \bar{X}_j)(X_{ij'h} - \bar{X}_{j'})$$

Test : homogénéité des variances

Test F ("M de Box") : extension du test de Bartlett

$$M = (n - K) \ln(|W|) - \sum_{i=1}^K (n_i - 1) \ln(|W_i|).$$

Test : égalité des espérances

Test sont basées sur les valeurs propres de la matrice $W^{-1}B$: $\lambda_1, \lambda_2, \dots$:

- Wilks's :

$$T = \Lambda = \frac{|W|}{|V|} = \frac{|W|}{|W + B|} = \frac{1}{|I + W^{-1}B|} = \prod_i \frac{1}{1 + \lambda_i}$$

- Pillai-Bartlett :

$$T = \text{trace}((I + W^{-1}B)^{-1}) = \sum_i \frac{1}{1 + \lambda_i}$$

- Hotelling :

$$T = \text{trace}(W^{-1}B) = \sum_i \lambda_i$$

- Roy's :

$$T = \text{la plus grande valeur propre de } W^{-1}B = \lambda_1$$