UE projet

Elodie Marchadier

Séances 4 et 5

Séance 4 (aujourd'hui) :

- cours integré sur la représentation de données, quels type de graphique dans quel cas (nuages de points, histogrammes, distribution, boxplots, barplots...)

- choix d'un jeu de données et réalisation de 3 representations graphiques pertinentes et visant à répondre à une question biologique d'intérêt.

Chaque graphe doit permettre de visualiser la réponse à une question et donner une intuition de la réponse à cette question.

Séance 5 :

- fin des graphes et prise en main des jeux de données

UE projet

Représentation de données

Elodie Marchadier

Variables aléatoires

Variable aléatoire : o*n appelle variable aléatoire toute variable dont la valeur dépend du résultat d'une expérience probabiliste*

On notera *X* la variable aléatoire, et x une réalisation de cette variable aléatoire.

La variable aléatoire est caractérisée par

- les valeurs qu'elle peut prendre, que l'on appelle le **support** de la variable aléatoire
- la probabilité d'observer chaque valeur dans la population, c'est sa **loi de probabilité**

Variables quantitatives et qualitatives

Variable qualitative

Les caractéristiques sont non numériques, ce sont des niveaux (=modalités) et les opérations n'ont pas de sens

- **nominales** (sexe, malade ou non, couleur des yeux, groupe sanguin...)
- **ordinales** (peu confortable, assez confortable, très confortable)

Variable quantitative

Les caractéristiques sont numériques, ce sont des nombres réels et les opérations (moyennes, variances…) ont un sens.

- **discrètes** (nombre de descendants d'un individu, comptages,…)
- **continues** (poids, taille, quantités...)

Exemples de variables aléatoires

Couleur d’un papillon

Points au lancé de dé

Taille des garçons de 20 ans

Nombre de graines par plante

Taille des filles de 10 ans

Comportement d’un écureuil

Décrire un échantillon

***n*-échantillon** : sous-ensemble de *n* individus tirés au hasard dans la population de référence

X_i : variable aléatoire associée au tirage de l'individu *i* (*i*= *1...n*)

x_i : valeur observée chez l'individu *i*

Les X_i sont indépendants et de même loi

v.a. quantitative : moyenne et variance

La **moyenne** est un indicateur de position

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

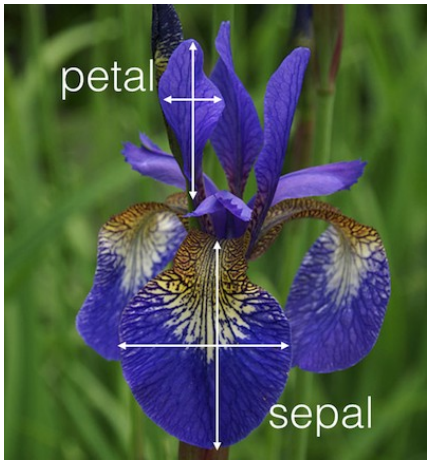
La **variance** donne une indication sur l'ampleur des variations autour de la moyenne

$$s^2_{n_x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

L'**écart-type** est la racine carrée de la variance (même unité que la variable)

Exemple : les Iris de Fisher

Edgar Shannon Anderson, botaniste américain (1897-1969), a collecté des données morphologiques sur 3 variétés d'Iris : *Iris setosa*, *Iris virginica* et *Iris versicolor*



5 variables ont été mesurées sur 150 individus (3x50) :
- largeur de pétale
- longueur de pétale
- largeur de sépale
- longueur de sépale
- espèce

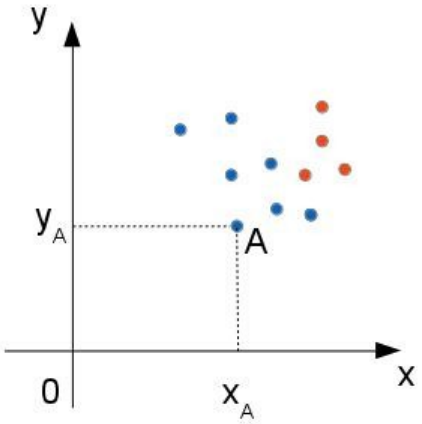
Identifier des caractères morphologiques caractéristiques des 3 espèces.

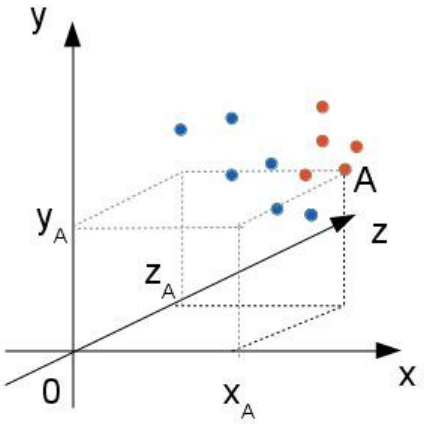
Ronald Fisher (1890-1962) utilise ce jeu de données pour illustrer l'analyse discriminante linéaire.

Introduction

Données quantitatives classiquement rencontrées en biologie :

- individus (points) *les iris*
- traits quantitatifs associés (coordonnées sur les axes) *les longueurs/largeurs*
- traits qualitatifs associés (couleurs des points) *les espèces*





Introduction

Représentation dans un repère : adapté aux jeux de données composés de 2 à 3 variables

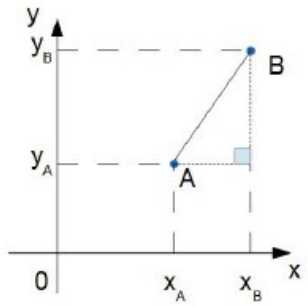
Et si l'on dispose d'un plus grand nombre de variables ?
méthodes d'analyse multivariée !

- Simplifier le jeux de données en identifiant les variables les plus informatives qui le composent (i.e. les composantes principales)
- Étudier les corrélations entre les variables
- Etudier les relations entre les individus
- Identifier les variables qui discriminent le mieux les individus

Rappels

Distances entre deux points

$$d(A,B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$



Dans un espace à *p* dimensions

$$d(x_i, x_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{i'j} - x_{ij})^2}$$

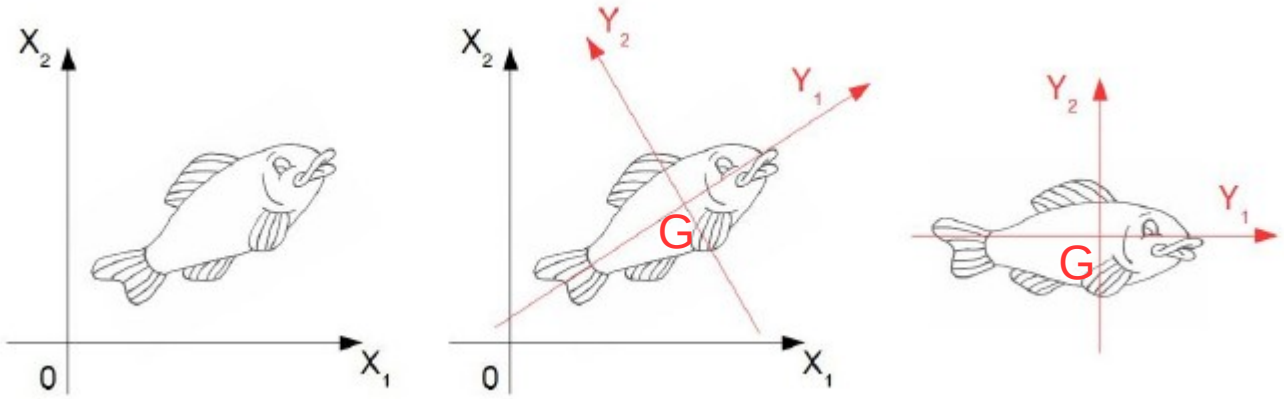
- x_i et $x_{i'}$, les coordonnées de deux points dans le repère X
- x_{ij} , la coordonnée du point x_i sur l'axe X_j
- $x_{i'j}$, la coordonnée du point $x_{i'}$ sur l'axe X_j

Centre de gravité G = [$x_{.1}$, $x_{.2}$, $x_{.3}$, ...$x_{.p}$] avec $x_{.j}$, moyenne des coordonnées des n points selon l'axe X_j

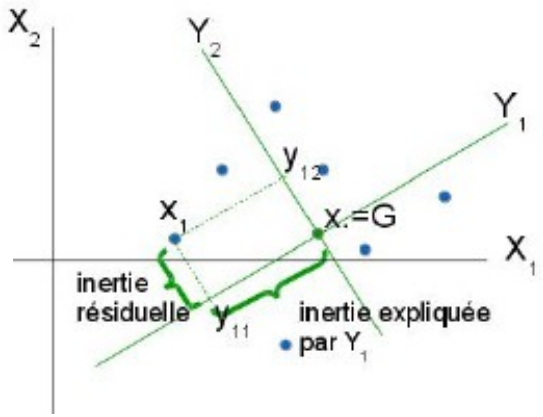
Inertie d'un nuage de points $I_G = \sum_{j=1}^{p} \sigma_j^2$ avec σ^2_j la variance des individus selon l'axe X_j

Principe

A partir d'une représentation de données dans un repère X à *p* dimensions, on réalise un changement de repère, le nouveau repère est noté Y et a pour origine G



L'inertie totale I_G du nuage de points reste la même.
On maximise l'inertie du nuage de points sur le 1er axe tout en minimisant l'inertie résiduelle de cet axe.



$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, x_i) = \frac{1}{n}\sum_{i=1}^{n} d^2(G, y_{i1}) + \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, y_{i1})$$

$$I_G = \text{inertie totale} = \text{inertie expliquée par } Y_1 + \text{inertie résiduelle}$$

Les axes de l'ACP

Axe = variable latente (ou virtuelle), combinaison linéaire des variables initiales

$$Y_{ik} = a_{k1} x_{i1} + a_{k2} x_{i2} + ... + a_{kJ} x_{iJ}$$

Y_{ik}, le projeté du *i*ème point sur le *k*ème axe Y

a_{kj}, le coefficient qui associe la variable initiale *j* à la variable latente *k* du nouveau repère

Les a_{kj} sont appelés **vecteurs propres** notés $\vec{a}_k = (a_{k1}, a_{k2}, a_{k3} ..., a_{kJ})$

Chaque axe Y_k de l'ACP a **une valeur propre** λ_{Yk} qui représente sa variance empirique ($\lambda_{Yk} = I_{Yk}$)

Le maximum d'inertie non-portée par le Y_1 sera portée par Y_2, etc…

$$I_{Y_1} > I_{Y_2} > I_{Y_3} > .... I_{Y_K}$$

Les axes Y sont orthogonaux car indépendants

Dans les cas d'une ACP sur variables réduites $I_G = 1$