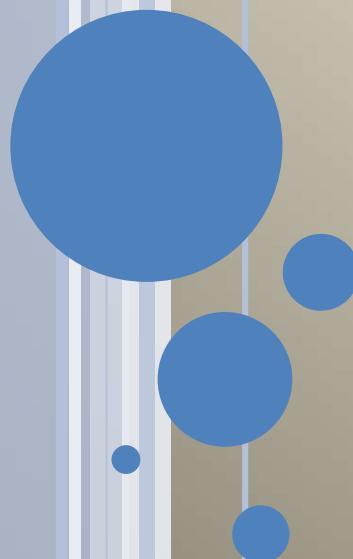


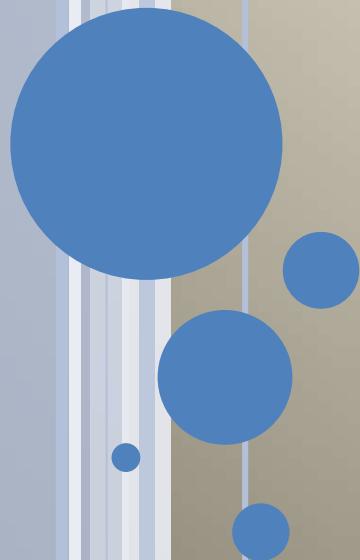
MODULE : ANALYSE DE DONNÉES & MATHÉMATIQUES FINANCIÈRES



IDRISSI NAJLAE
Université Sultan Moulay Slimane
Faculté des Sciences et Techniques
Béni Mellal
Département Informatique
© 2017-2023

PARTIE 1

ANALYSE DE DONNÉES



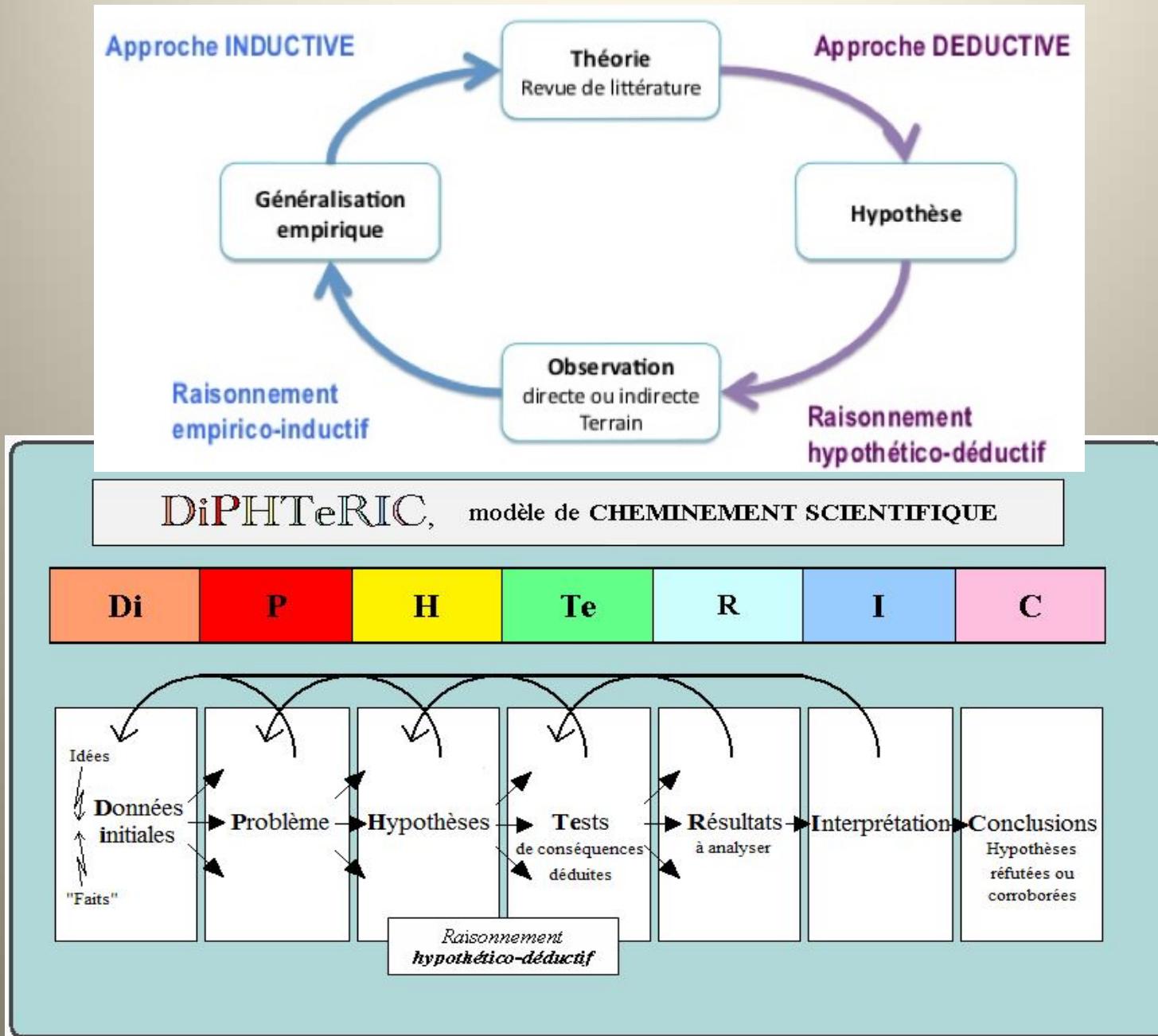
- Cours : 12h (Mme N. Idrissi)
 - principe de l'analyse de données
 - statistiques descriptives, régression,
 - ...
- TD : exercices d'entraînement et d'application des méthodes vues en cours
- TP: manipulation d'un logiciel d'analyse des données

INTRODUCTION

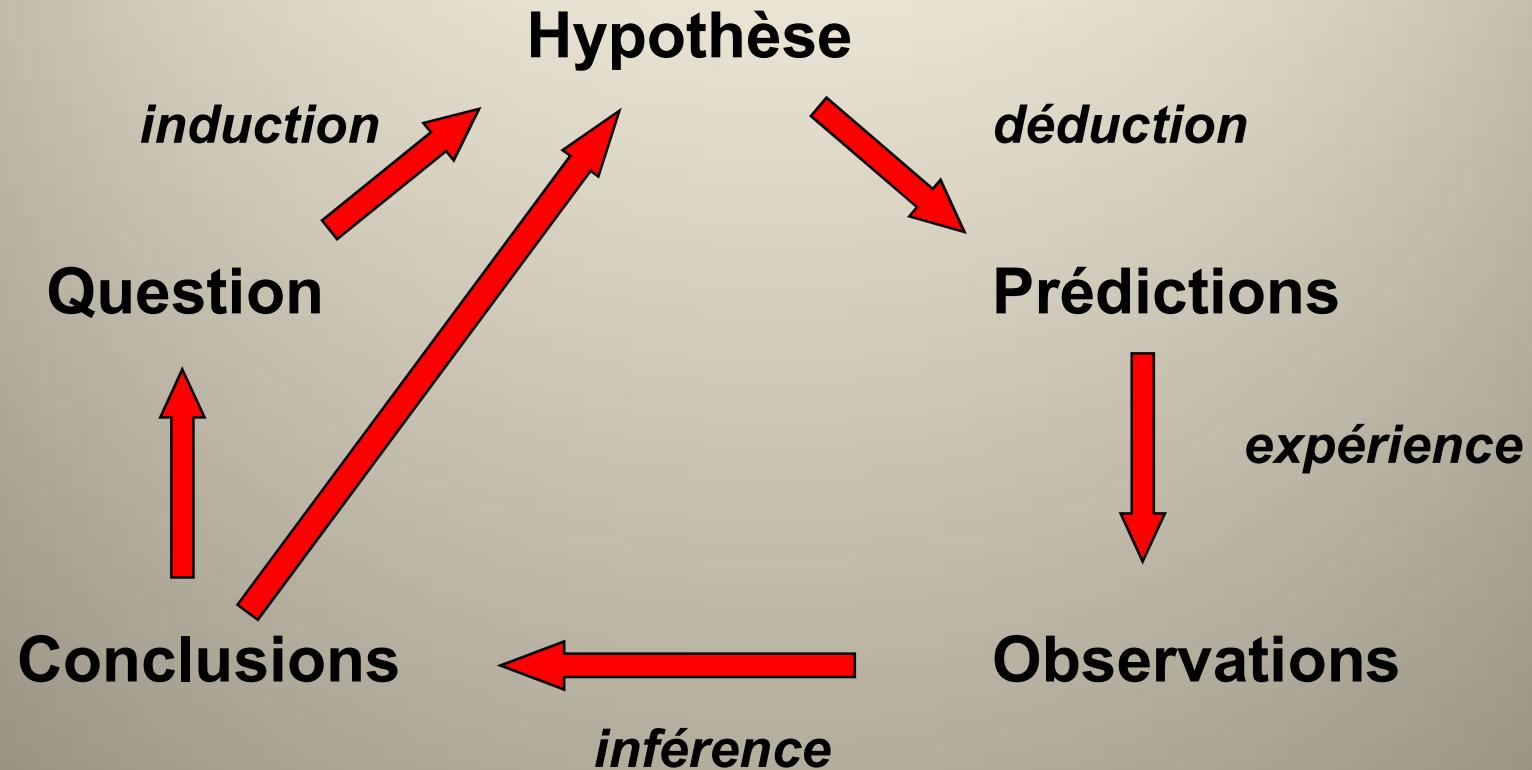
1. Démarches scientifiques
2. Les étapes de l'analyse statistique



Démarches scientifiques



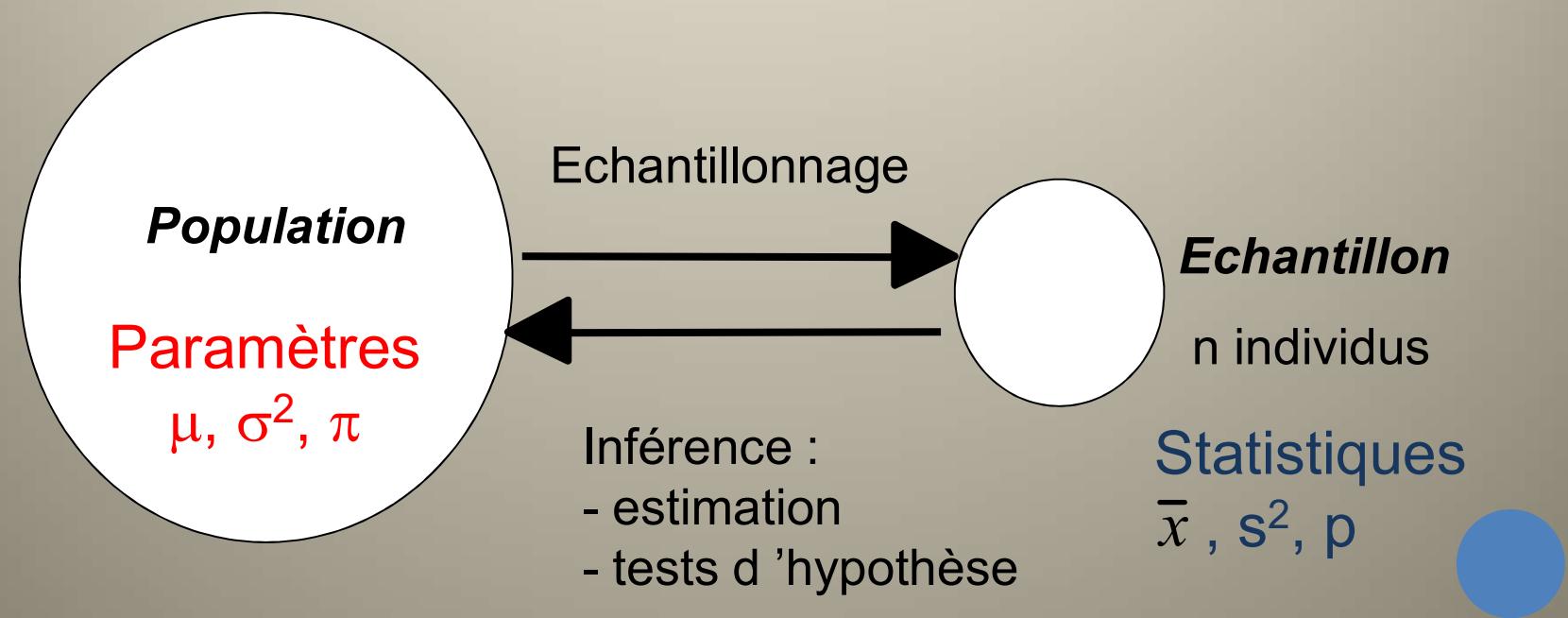
La démarche hypothético-déductive: la falsification d'hypothèses



2. Les étapes de la démarche statistique

population : ensemble d'individus (=personnes, villes...).

échantillon : ensemble d'éléments extraits d'une population.



1. Echantillonnage/collecte de données

Ensemble des opérations qui visent à prélever un échantillon dans une population.

- * **objectif** : obtenir un échantillon informatif
 - un **échantillon représentatif** de la population.
 - et/ou : un échantillon permettant d'obtenir des informations précises.
- * **méthode** : échantillonnage stratifié, en grappes...
 - le plus simple : un **tirage aléatoire simple**.

2. Statistiques descriptives

décrire les données, les présenter (choisir les tests appropriés).



3. Estimation

Estimer des paramètres de la population à partir de l'échantillon: \bar{x} **n'est pas égale à μ** , mais est « proche » de μ et nous donne des informations sur sa valeur.



4. Tests d'hypothèses

* **ajustement** : la distribution de la population est-elle conforme à une distribution de référence?

La glycémie est-elle une variable normale ?

* **conformité** : le paramètre de la population est-il conforme à une valeur de référence?

La glycémie des patients atteints de bizarre est-elle identique à celle de patients sains ?

2. Les étapes de la démarche statistique

Tests d'hypothèses

* **égalité** ou d'homogénéité : comparent plusieurs populations, à l'aide d'un nombre correspondant d'échantillons.

La glycémie des patients traités avec le traitement A est-elle identique à celle des patients traités par B ?

* **indépendance** entre deux caractères.

L'intensité du diabète est-il indépendant du régime alimentaire ?

Plan du **COURS**

Prérequis : notions de probabilités, notions sur les variables aléatoires

1 - Statistiques descriptives, lois de probabilité, estimation

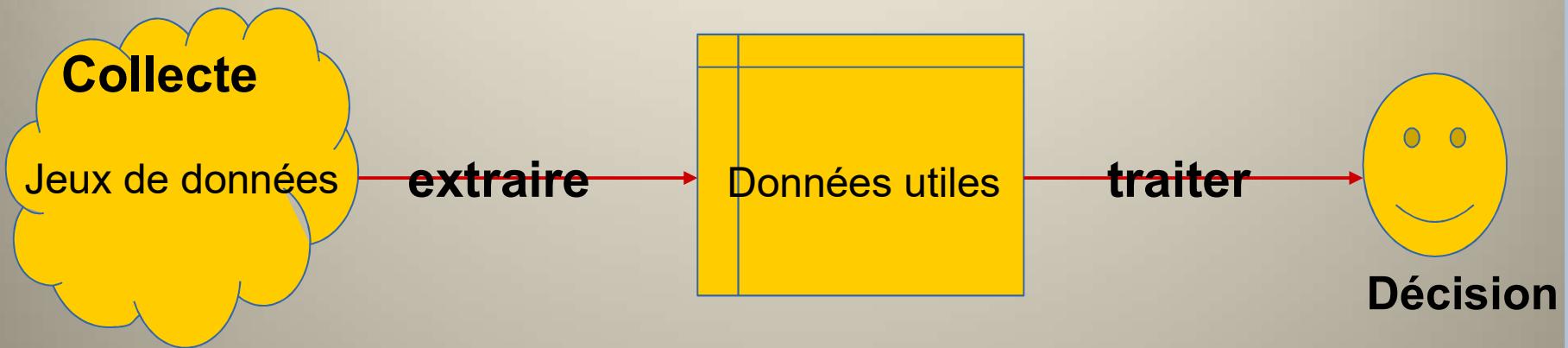
2 - tests d'hypothèse, tests de base

3 - Tests non paramétriques

4 - ANOVA, Régression



C'est quoi l'AD ?



1. Objectif(s) visé(s)
2. Collecte
3. Analyse
4. Décision

Exemples

- ➔ Recensement de la population marocaine en 2014 (HCP)(taux de scolarité, genre, milieu, niveau salariale, ...)
- ➔ Entreprises de production (attitudes et préférences des consommateurs, avis, prix, ...)
- ➔ Banques (ménages à crédit, ...)
- ➔ Étude de pathologies (genre, âge, milieu, ...)//
- ➔/// phénomènes sociales, économiques, religieux,
- ➔



Enquête, sondage, questionnaire, ...

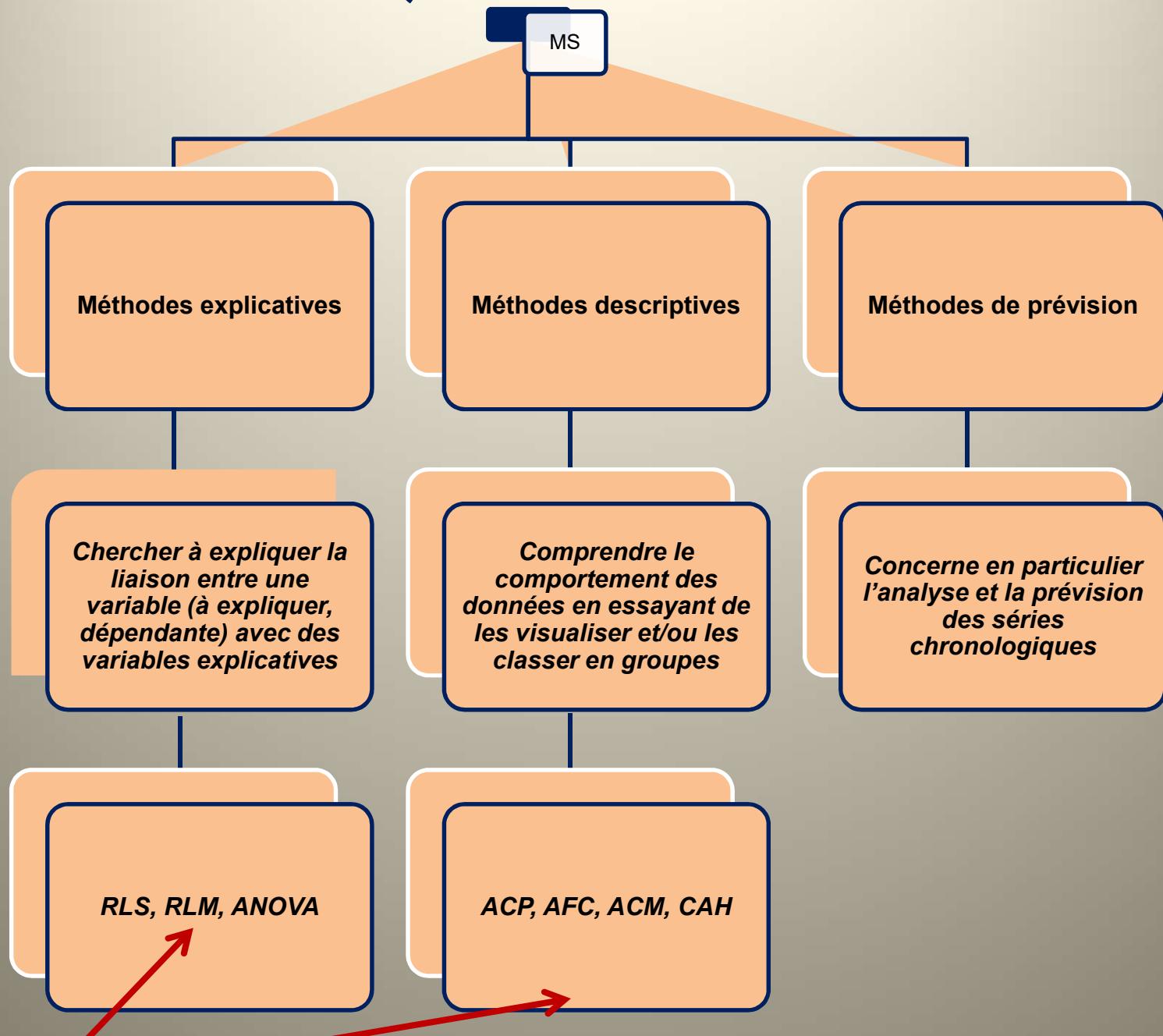
POURQUOI LA STATISTIQUE?

- ✓ ***Le fait d'étudier des phénomènes simples ou complexes à partir des faits constatés (une enquête, un sondage, des expériences, ...)***
- ✓ ***constater l'augmentation et l'amélioration de la récolte en introduisant des engrais spécifiques***
- ✓ ***l'augmentation du stress avec l'usage fréquent des appareils intelligentes***
- ✓ ***...***

LA STATISTIQUE?

« C ’est un ensemble de méthodes permettant de décrire, d’analyser et d’interpréter d’une manière quantifiable des phénomènes observés

MÉTHODES STATISTIQUES



NOTIONS STATISTIQUES

Population :

Une population est l'ensemble sur lequel on effectue des observations, des expériences.

Échantillon:

Une partie de la population tirée au hasard ou représente un ensemble de données bien choisi -→ taille

Individu (unité statistique) :

Les individus sont les éléments de la population/échantillon étudié(e).

Variable statistique (les statistiques):

C'est une caractéristique précise observée sur les individus en question.

Série statistique :

C'est l'ensemble de valeurs numériques ou autres observées d'un caractère statistique (sur l'échantillon)

Exemple:

1. L'entreprise MyLadyInc voudrait lancer une nouvelle gamme de son produit Gold. Pour cela, elle mène une étude sur 1692 personnes qui a montré que 75% de personnes sont satisfaits pour le prix de 700 dhs dont 63% sont des jeunes, 55% sont des femmes, parmi elles 25% sont des chefs d'entreprises.

1. Le ministère d'agriculture voudrait étudier la répartition des terres agricoles de la région de BM. Pour cela, il procède à un inventaire des exploitations agricoles de la région et noter pour chacune d'elle sa taille.

VARIABLES

Types de variables

Quantitatives

Qualitatives

L'ensemble des valeurs numériques ou nombres qu'elle peut prendre ou mesurer

L'ensemble des valeurs exprimées sous forme littérale ou par un codage numérique. On parle de modalités

continu

discret

Taille, salaire, nombre d'enfants, nombre d'étudiants, ...

nominal

ordinal

Sexe, couleur, taille vêtements, catégorie d'âge, ...

REPRÉSENTATION DES DONNÉES (PHASE 1)

Tableau (ind x vars)

Chaque caractère est observé sur un individu

ind\vars	x_1	x_2	...	x_m
1				
2				
3				
...				
...				
n				

Tableau effectif / fréquence

Chaque caractère en commun est observé sur un ensemble d'individus

	sexe
F	620
H	150

	Niveau scolaire
PM	1208
SC	25678
HG	23886

Exemples

N° Exploitation	Taille (ha)	Age du chef d'exploitation (années)	Culture dominante	Nombre de personnes employées
1	50	50	blé	2
2	50.5	45	vigne	4
3	35	38	orge	3
4	62.1	25	blé	6
5	20	65	vigne	1
6	10	57	vigne	1
.
.
630	56	45	blé	2

Dans le tableau présenté ci-dessus, il y a :

combien d'individus ?

combien de variables ?

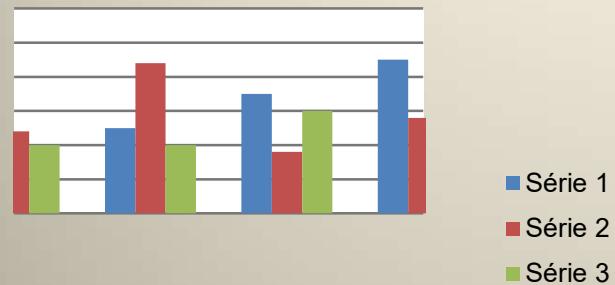
Phase 2 –

Statistique descriptive



1- REPRÉSENTATION GRAPHIQUE

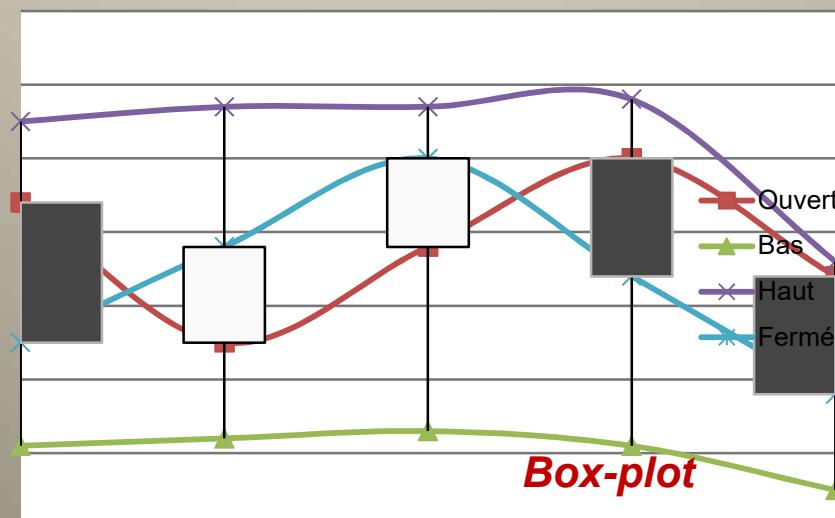
Permet une première analyse visuelle de la distribution des données/variables



Histogramme



Nuage de points



2- TABLEAUX DE FRÉQUENCE

Valeurs de la variable	Effectifs	Fréquences	%	Effectifs cumulés croissants N_i	Effectifs cumulés décroissants N'_i
x_1	n_1	$f_1 = n_1/n$	$f_1 \times 100$	$N_1 = n_1$	$N'_1 = n_k + \dots + n_1 = n$
...		$N_2 = n_1 + n_2$	$N'_2 = n_k + \dots + n_2$
x_i	n_i	$f_i = n_i/n$	$f_i \times 100$	$N_3 = n_1 + n_2 + n_3$	$N'_3 = n_k + \dots + n_3$
...
x_k	n_k	$f_k = n_k/n$	$f_k \times 100$	$N_{k-1} = n_1 + \dots + n_{k-1}$	$N'_{k-1} = n_k + n_{k-1}$
Total :	$\sum n_i = n$	$\sum f_i = 1$	100		$N'_k = n_k$

VARIABLES QUALITATIVES

Modalités	Effectifs	Fréquences	%
Bleu	60	0.200	20,0
Noir	160	0,533	53,3
Noisette	40	0,133	13,3
Vert	40	0,133	13,3

300 personnes sur lesquelles on a observé la couleur des yeux

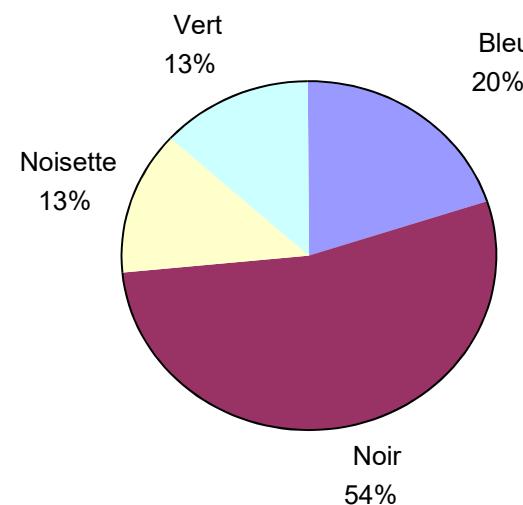


Diagramme circulaire (camembert)

Modalités	Effectifs
Pas satisfait (A)	10
Un peu (B)	25
satisfait (C)	40
Passionnément (D)	32

107 personnes ont été interrogées sur leur satisfaction du nouveau produit laitier

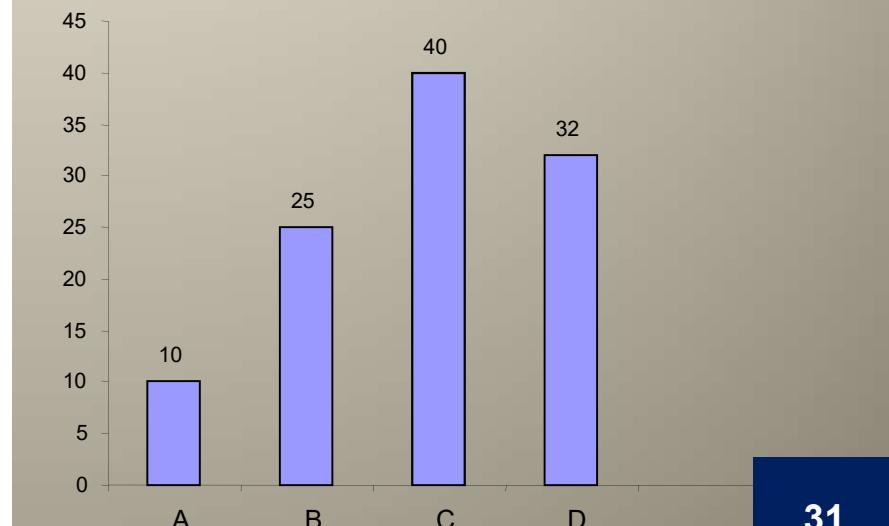
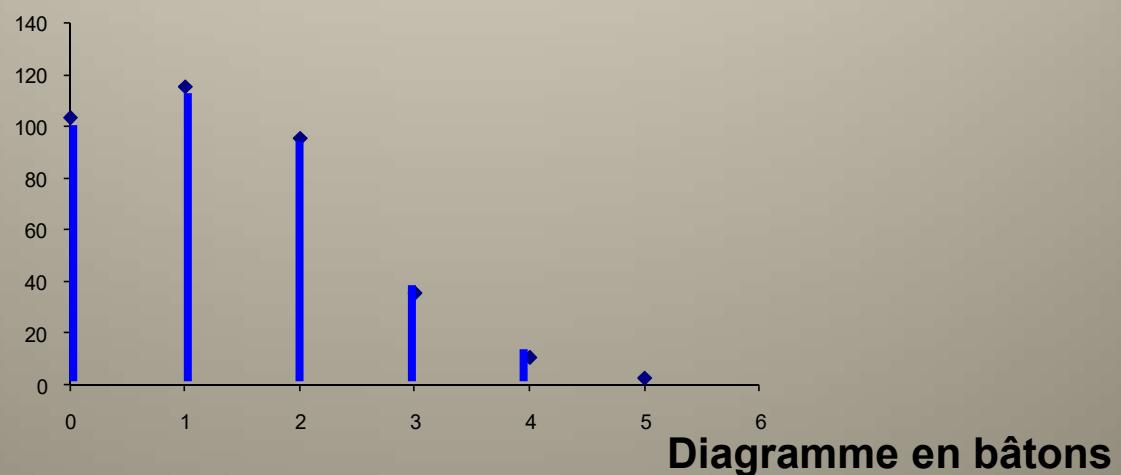


Diagramme en barres ordonné

VARIABLES QUANTITATIVES

-- cas discret --

Niveau scolaire x_i	Effectif n_i	Fréquence f_i
0 (maternelle)	103	0,286
1 (CP)	115	0,319
2 (CE1)	95	0,264
3 (CE2)	35	0,097
4 (CM1)	10	0,028
5 (CM2)	2	0,006



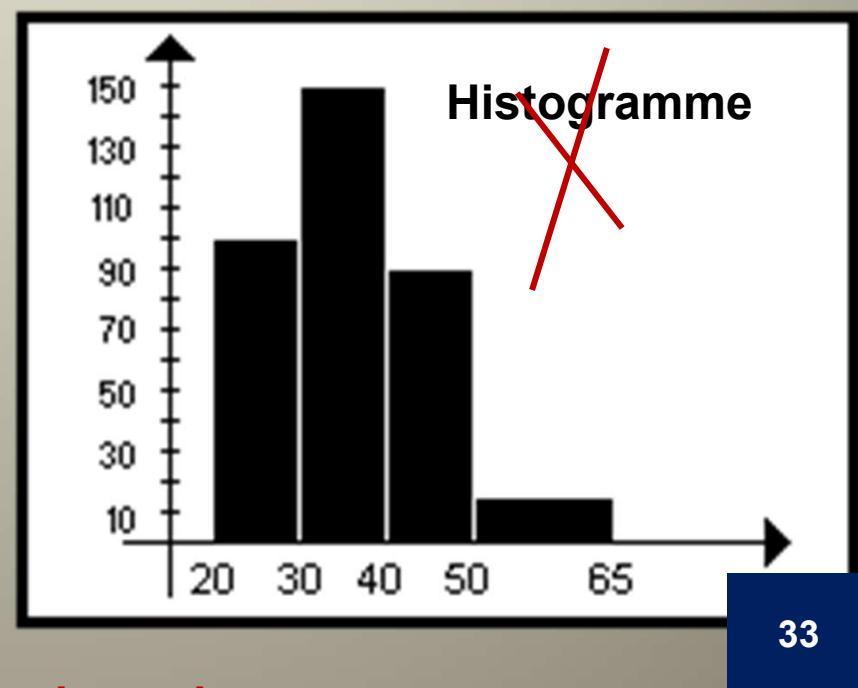
VARIABLES QUANTITATIVES

-- cas continu--

Classes	Effectifs
$[e_1 - e_2[$	n_1
$[e_2 - e_3[$	n_2
....
$[e_k - e_{k+1}[$	n_k

Représentation par intervalles ou classes

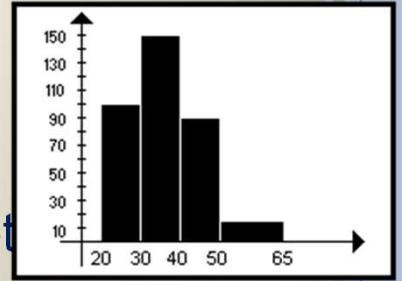
Age (ans)	Nombre de personnes
20 à 30	100
30 à 40	150
40 à 50	90
50 à 65	20



Rectification en cas de dynamique différente

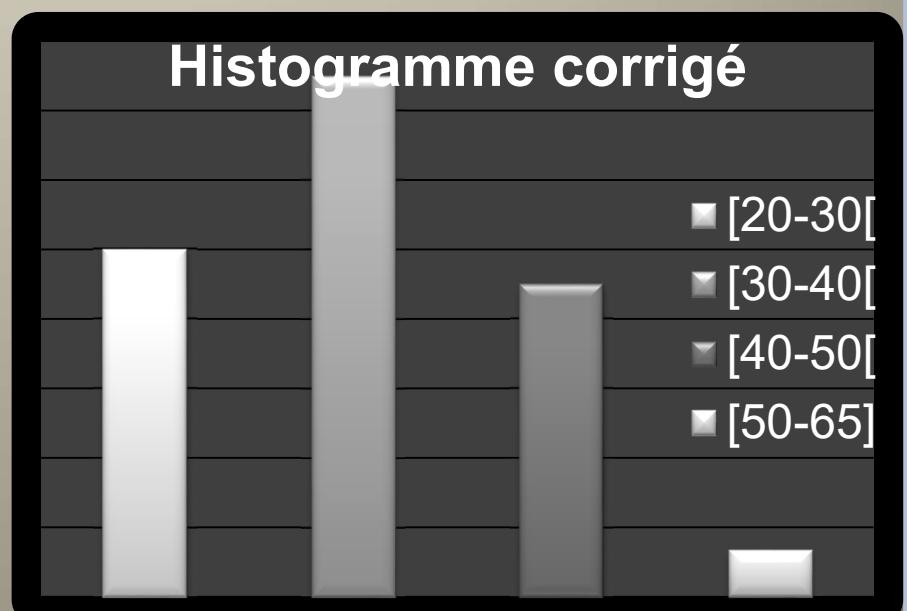
Rectification

La correction des effectifs ou des fréquences se fait en trois étapes :



- 1-** Calcul des amplitudes des classes a_i ;
- 2-** Choix d'une amplitude de base a (généralement l'amplitude la plus petite) et calcul du rapport amplitude de la classe sur l'amplitude de base (a_i/a)
- 3-** Calcul des effectifs corrigés : $n'_i = n_i/(a_i/a)$ ou $f'_i = f_i/(a_i/a)$

Age x_i	Effectif n_i	a_i	a_i/a	n'_i
[20-30[100	10	1	100
[30-40[150	10	1	150
[40-50[90	10	1	90
[50-65]	20	15	15/10	13,33



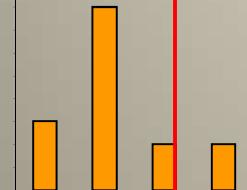
En résumé

VARIABLE QUALITATIVE

Nominale

Diagramme en barres

Effectifs ou Fréquences



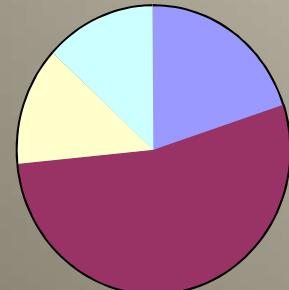
Ordinal

Diagramme en barres
Modalités dans l'ordre

Effectifs ou Fréquences



Diagramme circulaire

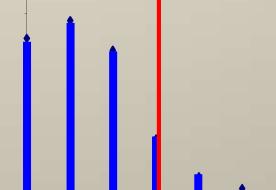


VARIABLE QUANTITATIVE

Discrète

Diagramme en bâtons

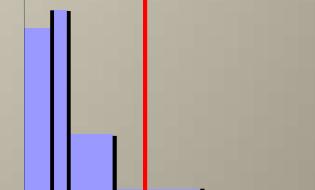
Effectifs ou Fréquences



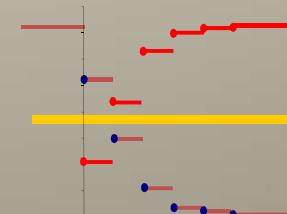
Continue

Histogramme

Effectifs ou Fréquences



Courbes cumulatives des effectifs ou des fréquences



35

3- PARAMÈTRES STATISTIQUES

- Les paramètres (indicateurs, mesures) statistiques sont des calculs (une seule quantité numérique) qui ont pour but de :
- Résumer d'une manière claire et précise l'essentiel de l'information relative au caractère statistique observé
 - Permettre d'avoir une idée sur la distribution statistique du caractère observé;



Les paramètres statistiques ne concernent que les variables quantitatives



PARAMETRES STATISTIQUES

Tendance centrales

- mode
- moyenne
- médiane

Dispersion

- variance
- écart-type
- étendue
- coefficient de variation

Position

- quartile
- centile

Indicateurs centrales

- **Moyenne:** valeur numérique autour de laquelle les observations sont réparties et notée \bar{X}

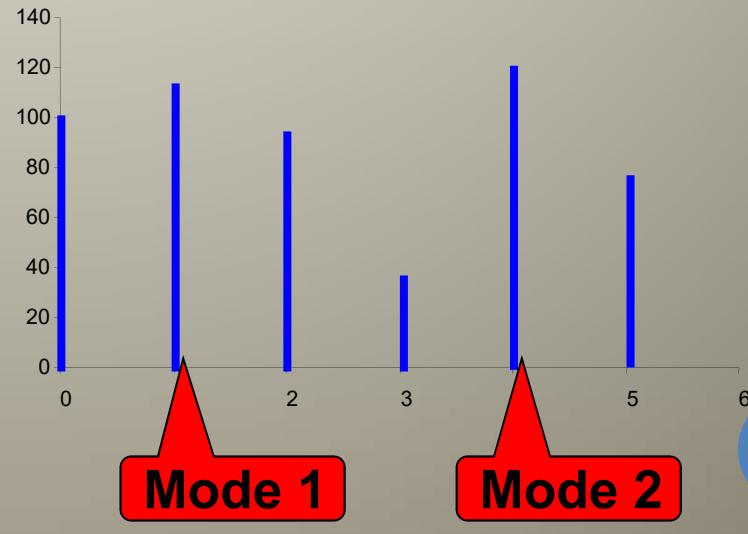
$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{ou} \quad \frac{\sum_{i=1}^n n_i x_i}{\sum_{i=1}^n n_i}$$

- **Mode:** c'est la valeur dont la fréquence est la plus élevée.

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



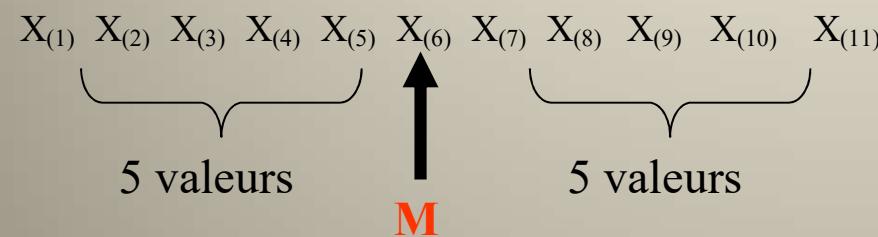
distribution unimodale



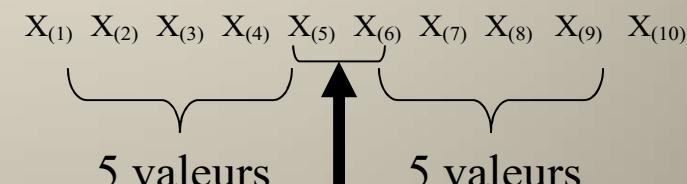
distribution bimodale

- **Médiane**: elle correspond à la valeur du caractère observé (x) pour laquelle 50% des valeurs observées sont supérieures et 50% sont inférieures.

Nombre impair d'observations



Nombre pair d'observations

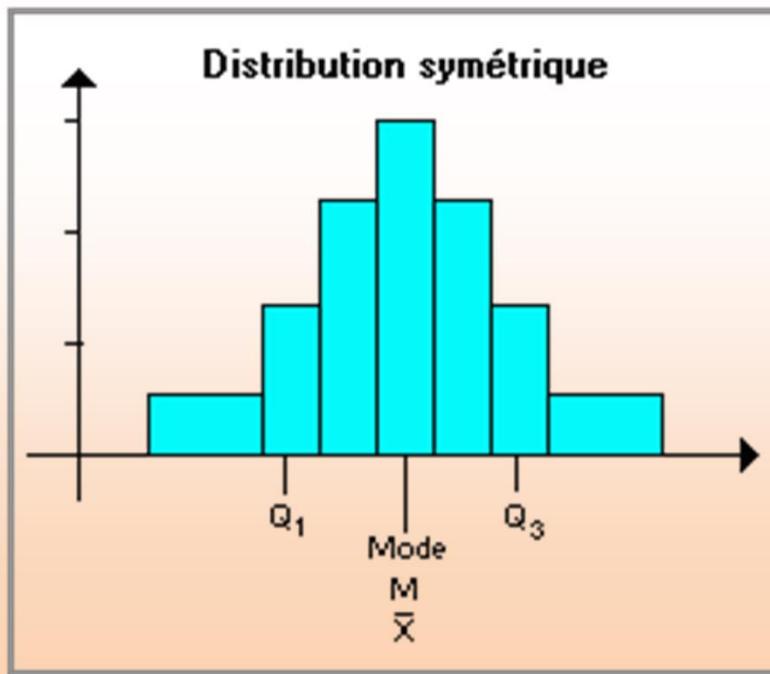
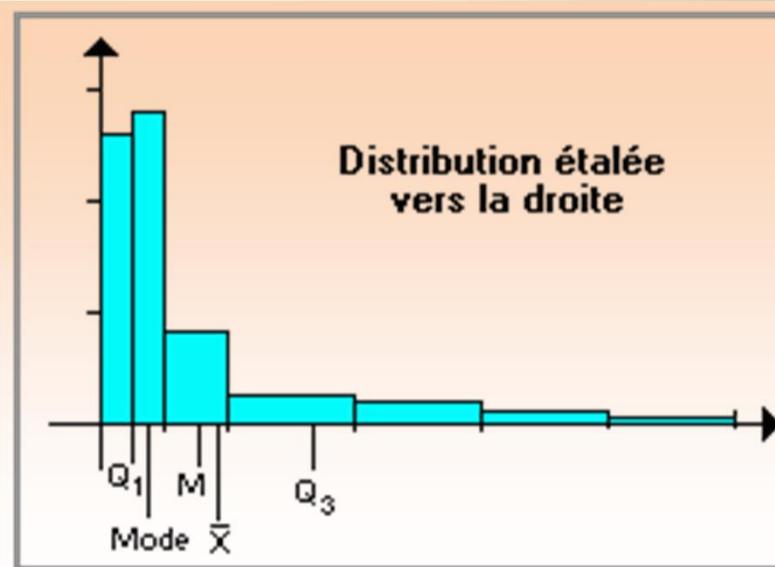
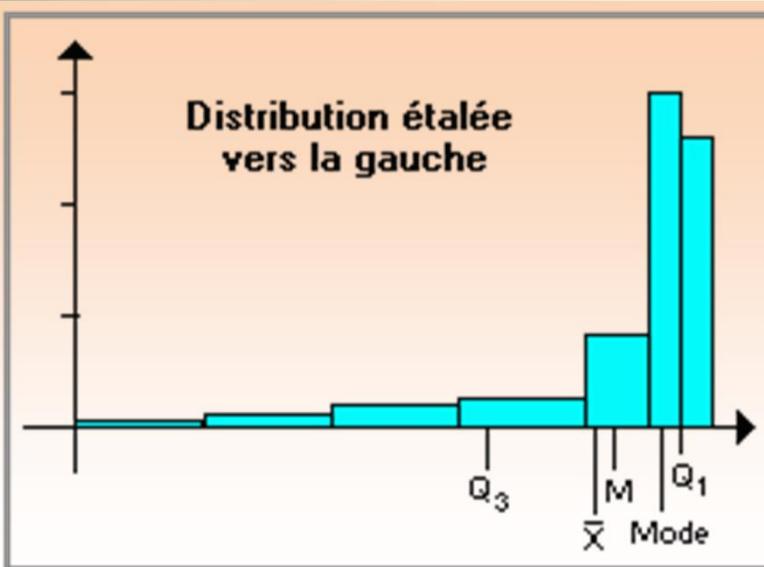


Intervalle médian

$$M = \text{milieu} = (x_{(5)} + x_{(6)})/2$$

La série $X=(x_1, x_2, \dots, x_n)$ originale doit être triée en $X'=(x_{(1)}, x_{(2)}, \dots, x_{(n)})$

Comparaison



Indicateurs de position

- ***Les quartiles sont les valeurs qui partagent la série statistique en quatre parts égales.***
 - Le 1^{er} quartile Q1, est la valeur en dessous de laquelle se situent 25 % des observations;
 - Le 2^{ème} quartile Q2 est la valeur en dessous de laquelle se situent 50 % des observations et au-dessus de laquelle se situent 50 % de la population. Il correspond donc à M;
 - Le 3^{ème} quartile Q3 est la valeur en dessous de laquelle se situent 75 % des observations.

- ***Les déciles sont les valeurs qui partagent la série en 10 parts égales.***
 - Le 1^{er} décile D1 est la valeur en dessous de laquelle se situent 10 % des observations;
 - Le 2^{ème} décile D2 est la valeur en dessous de laquelle se situent 20 % des observations;
 - Etc.
 - Le 9^{ème} décile D9 est la valeur en dessous de la quelle se situent 90 % des observations ou encore au-dessus de laquelle se situent 10 % de ces observations.

Indicateurs de dispersion

Mesure l'écart par rapport à la moyenne

Etendue : $R = x_{\max} - x_{\min}$

Variance :

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

Écart-type:

$$\sigma = \sqrt{V}$$

Coefficient de variation:

$$\frac{\sigma}{\bar{x}}$$

Distance interquartile:

$$DI (IQ) = Q_3 - Q_1$$

TD1

✓ Exercice 1:

Le tableau suivant représente les notes de statistiques de 2 classes différentes dans une école:

Centre classes	Classes x_i	Effectifs n_{1i}	Effectifs n_{2i}	\bar{x}_1	\bar{x}_2
2	[0; 4	0	2	0	4
6]4; 8]	1	2	6	12
10]8; 12]	10	3	100	30
14]12; 16]	2	3	28	42
18]16; 20]	0	2	0	36

1. Représenter le polygone/ histogramme des deux classes
2. Etudier la dispersion pour confirmer vos constats
3. Mode et médiane

✓ Les autres exercices vous sont fournis en papier

Phase 3 –

Analyse statistique bivariée



Chapitre 2

Régression linéaire simple

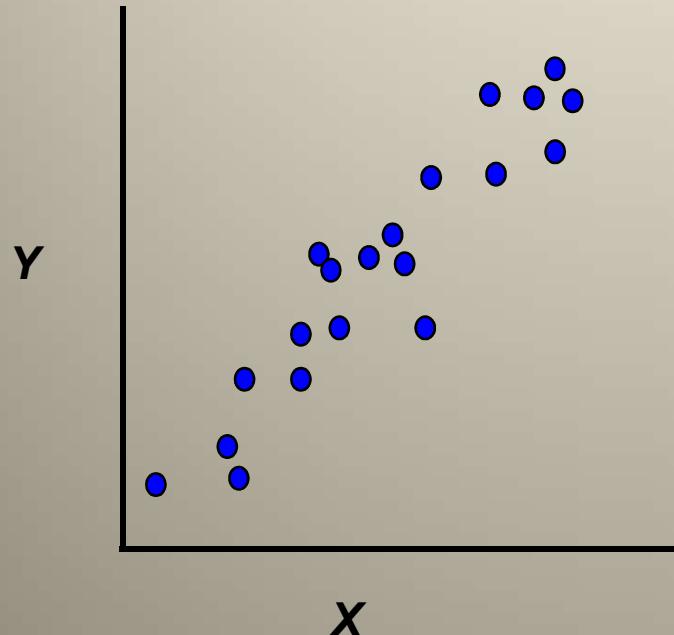


N. IDRISI
FACULTE DES SCIENCES ET TECHNIQUES
DEPARTEMENT INFORMATIQUE
BENI MELLAL
BLOC C RDC

Objectif

Etudier la relation entre deux variables quantitatives:

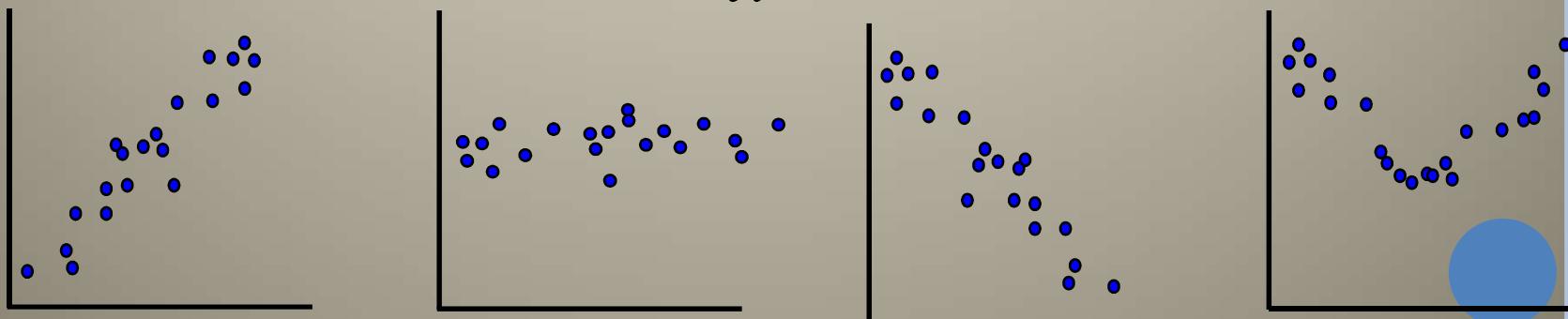
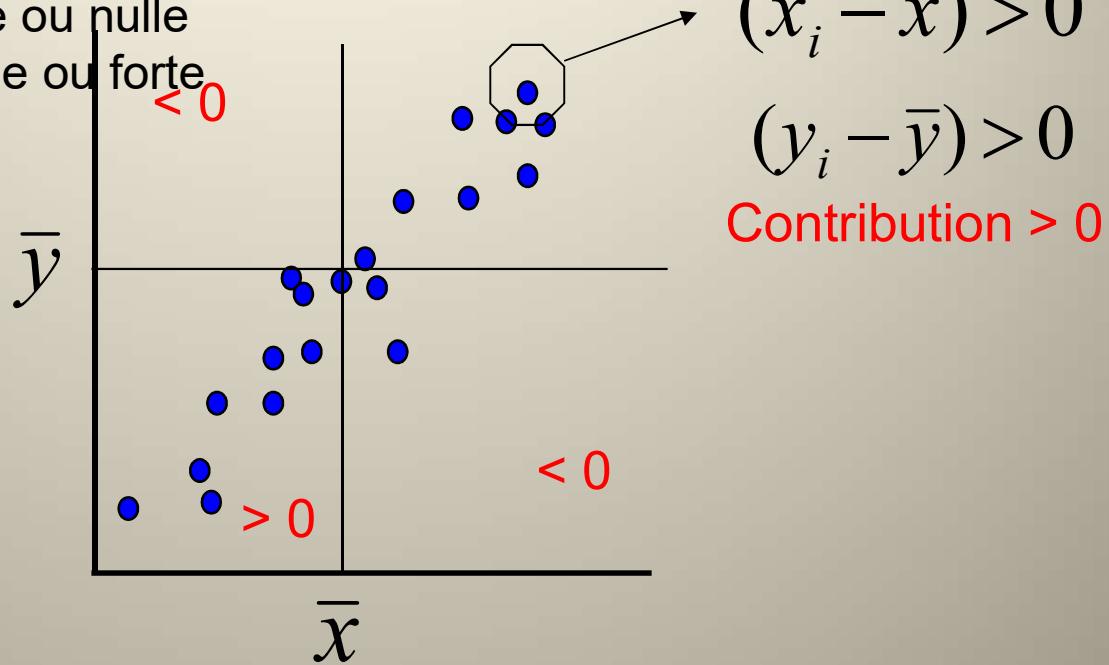
Nuage de points:



1. description de l'association linéaire: corrélation, régression linéaire simple
2. explication / prédiction d'une variable à partir de l'autre: modèle linéaire simple

La corrélation

- Est le degré de dépendance entre deux variables quantitatives
- Est positive, négative ou nulle
- Nulle, faible, moyenne ou forte



La corrélation

Statistique descriptive de la relation entre X et Y: variation conjointe

1. La covariance

Dans l'échantillon:

$$\text{cov}(x,y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$$

Estimation pour la population:

$$\text{cov}(x,y) = \hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$
$$\text{cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \frac{n}{n-1} \bar{x}\bar{y}$$



2. Le coefficient de corrélation linéaire

« de Pearson »

Dans l'échantillon:

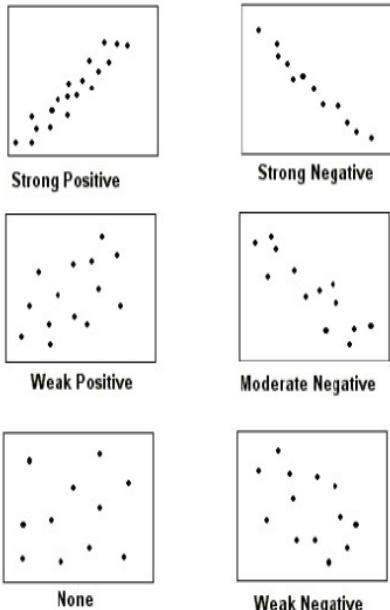
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

Estimation pour la population:

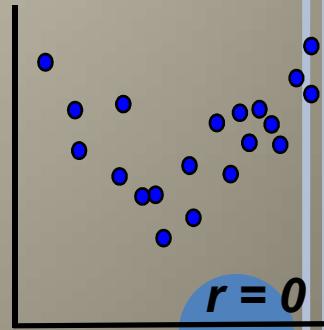
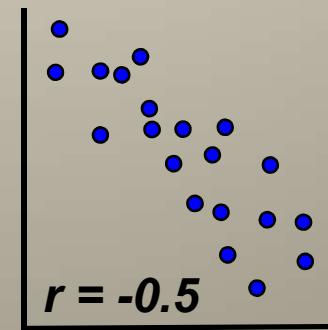
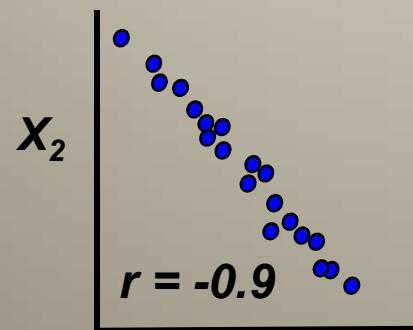
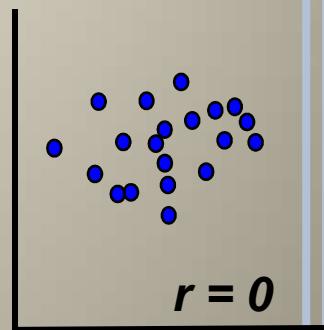
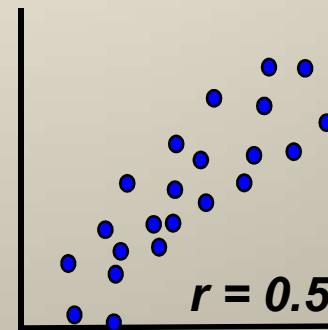
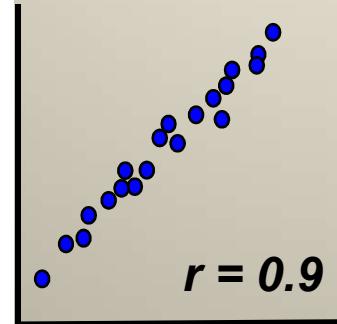
$$\hat{\rho}_{xy} = r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$



Degree of Correlation



de covariance absolue: $-1 \leq r \leq 1$



Tests de la corrélation

b. Test de $\rho = 0$

$$\begin{cases} H_0 : \rho = 0 & \text{Absence de corrélation} \\ H_a : \rho \neq 0 & \text{Corrélation existante} \end{cases}$$

Sous H_0 :

$$\left| t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| < t_{n-2,\alpha}$$

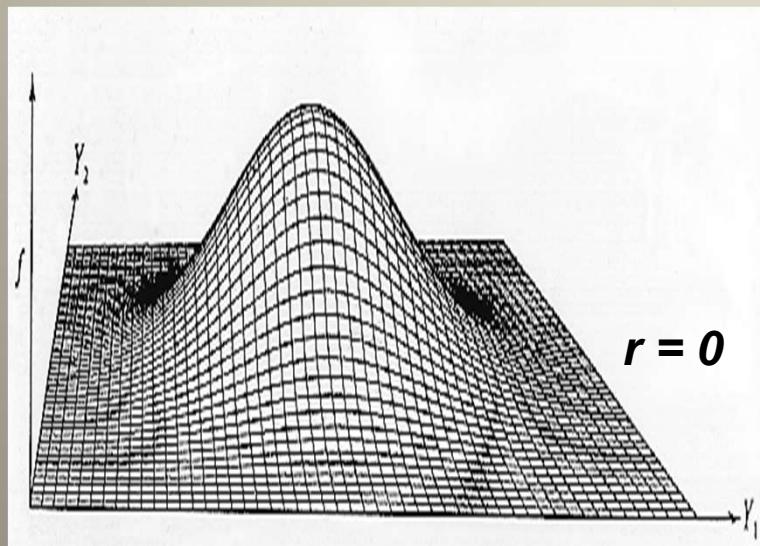
Si H_0 est rejetée ($p\text{-value} <$ corrélation)



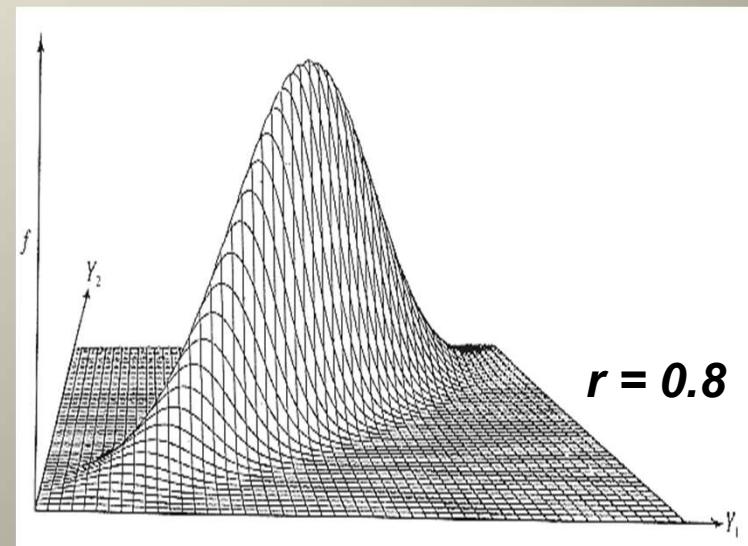
La régression linéaire

Conditions d'utilisation

a. Normalité



- Test de Shapiro-Wilk ou Kolmogorov-Smirnov
- Tracé Q-Q plot

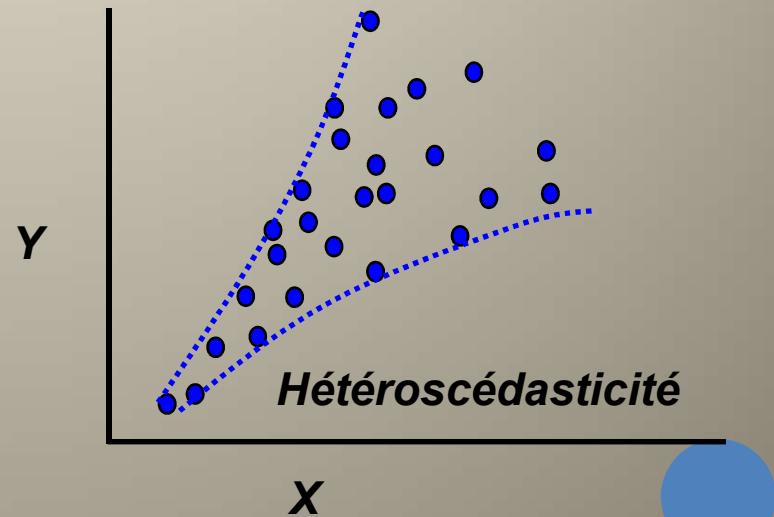
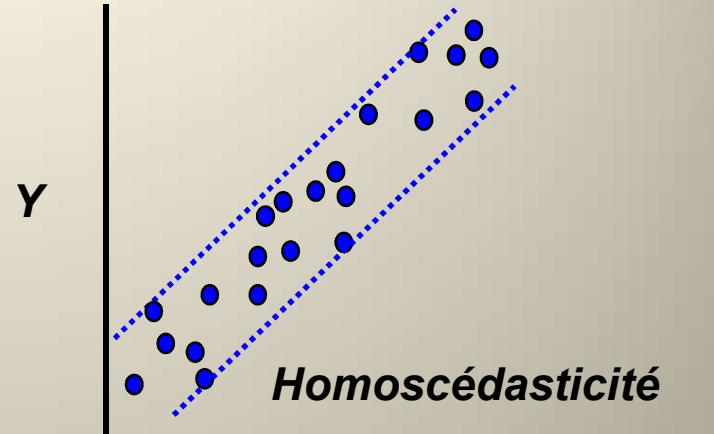


H_0 : les données suivent une distribution normale
 H_1 : les données ne suivent pas une distribution normale

b. Homoscédasticité

La variance de Y est indépendante de X et vice-versa.

*H₀ : les variances des deux groupes sont égales.
H₁ : les variances sont différentes.*



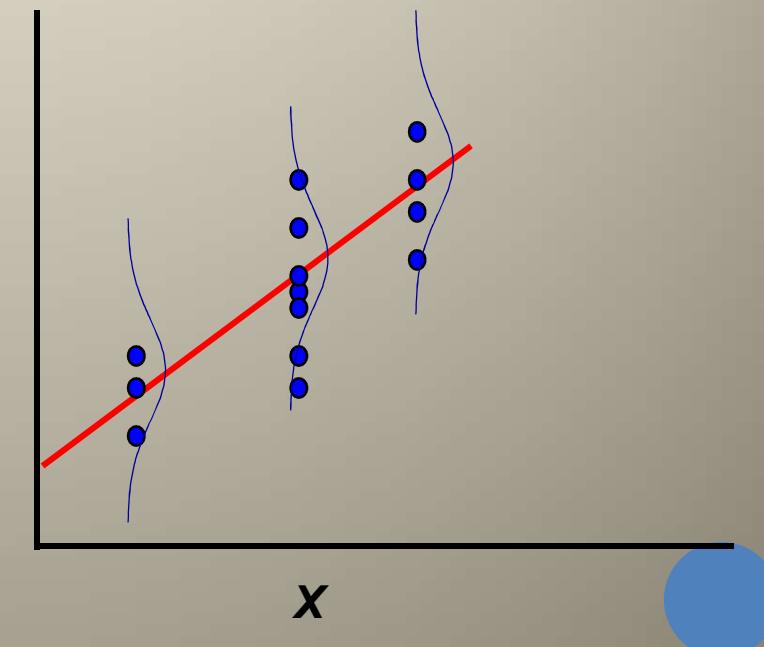
- Test F
- Test de Bartlett
- Test de Levene

Modélisation mathématique

On suppose: $y = f(x) = a + b*x$

Modèle: $Y_i = a + bX_i + e_i$ avec, pour $X = x_i$, $Y_i : N(a+bx_i, \sigma)$

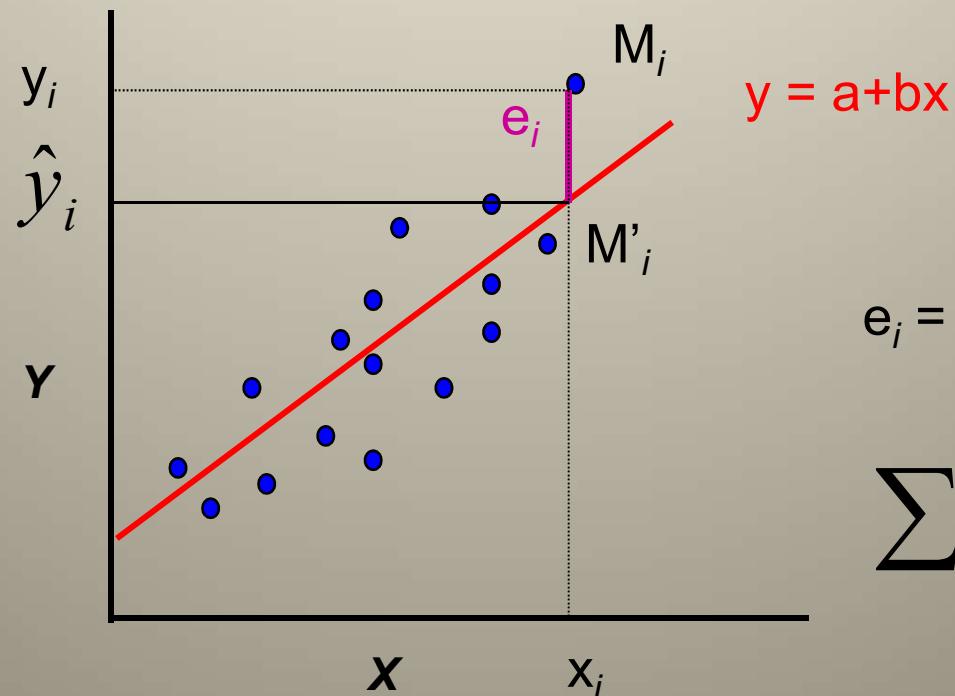
- X = variable explicative
(\ll indépendante \gg)
- Y = variable expliquée (dépendante)



L'estimation des paramètres

a? b?

Méthode d'estimation: les moindres carrés:



$$e_i = y_i - (a + bx_i)$$

$$\sum e_i^2 \text{ minimale}$$



Méthode des moindres carrés

On cherche à minimiser :

$$\sum_{i=1}^n (y_i - (a + bx_i))^2 = E(a, b)$$

$$\begin{cases} \frac{\partial E}{\partial a} = \sum_{i=1}^n 2(y_i - (a + bx_i))(-1) = 0 & (1) \\ \frac{\partial E}{\partial b} = \sum_{i=1}^n 2(y_i - (a + bx_i))(-x_i) = 0 & (2) \end{cases}$$

$$(1) \Rightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n (a + bx_i) = na + b \sum_{i=1}^n x_i$$

$$n\bar{y} = na + nb\bar{x}$$

$$a = \bar{y} - b\bar{x}$$

$$\hat{b} = \frac{\text{cov}(x, y)}{s_x^2}$$

→ On peut alors prédire y pour x compris dans l'intervalle des valeurs de l'échantillon:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$



Qualité de l'ajustement

On a supposé: $Y_i = a + bX_i + e_i$ avec
pour $X = x_i$, $Y_i : N(a+bx_i, \sigma)$

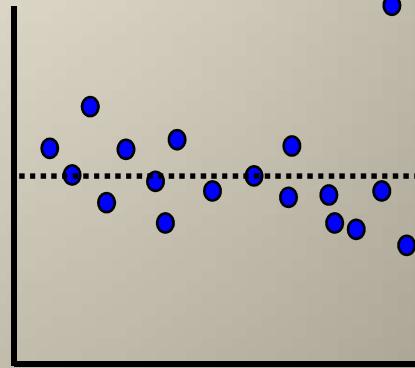
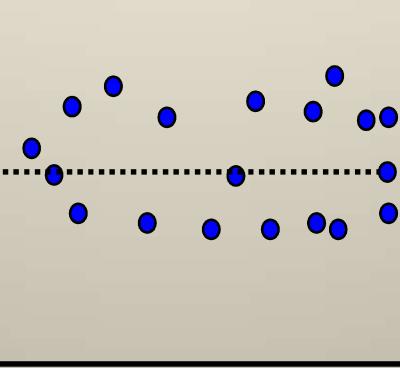
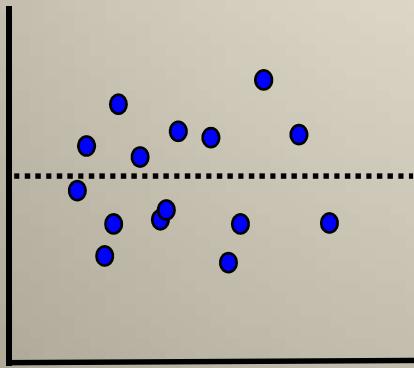
- distribution normale des erreurs
- variance identique (homoscédasticité)
- indépendance:
- linéarité de la relation $\text{cov}(e_i, e_j) = 0$

Test *a posteriori*: étude du nuage de points/ du graphe des résidus



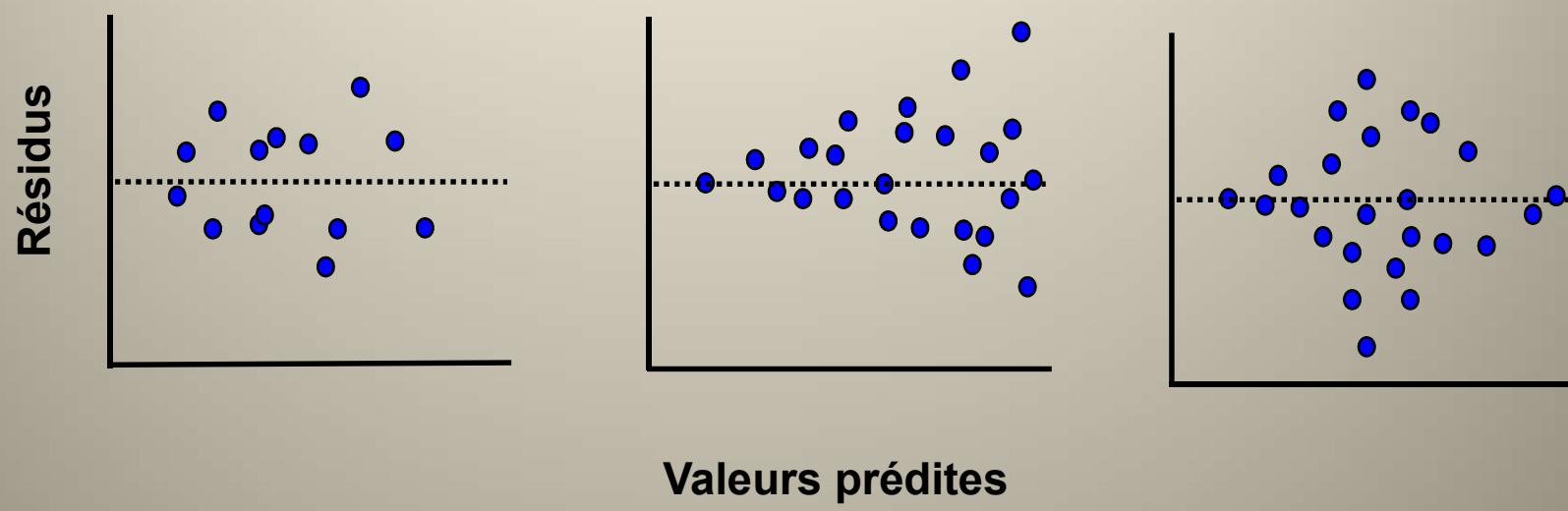
Normalité de l'erreur

Résidus



Valeurs prédictes



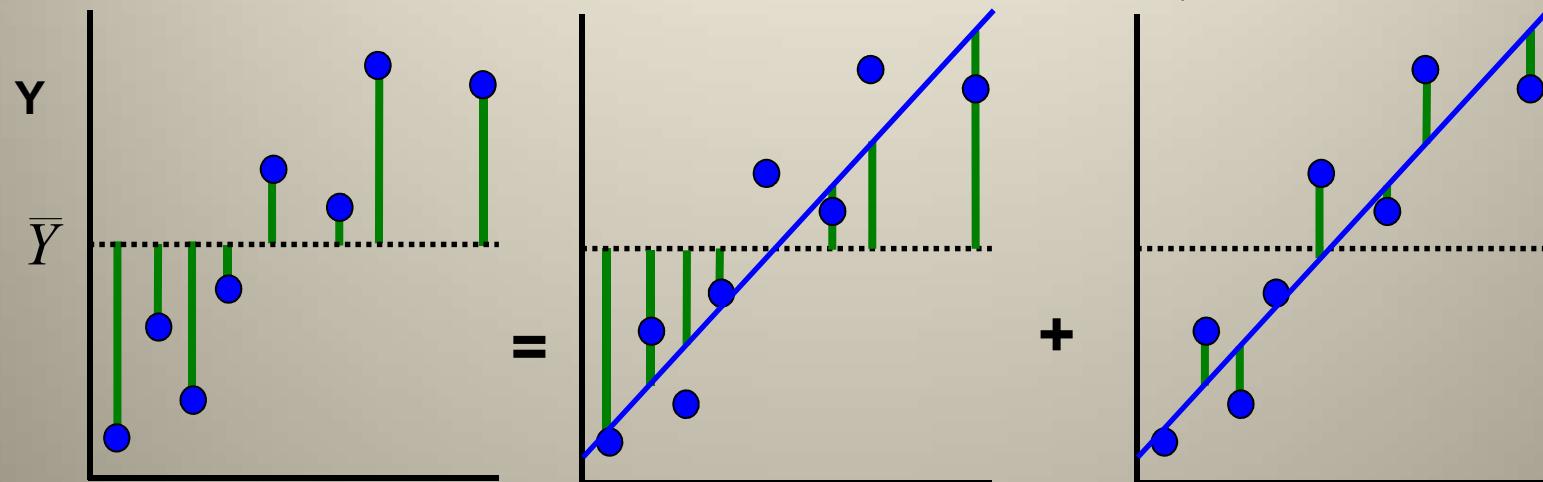


Possibilité de transformation: attention aux transformations *ad hoc*

Décomposition de la variation

Quelle part de la variabilité de Y est expliquée par la relation linéaire avec X?

Variabilité? Somme des Carrés des Ecarts SCE: $SCE_T = \sum_{i=1}^n (y_i - \bar{y})^2 = ns_y^2$



SCE Totale

$$\sum_{i=1}^N (Y_i - \bar{Y})^2$$

SCE reg.lin. (Expliquée)

$$= \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2$$

SCE hors reg.lin. (erreur)

$$+ \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

Relation entre r et r^2

Coefficient de détermination

$$\begin{aligned} SCE_{reg.lin.} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n ((a + bx_i) - (a + b\bar{x}))^2 \\ &= b^2 \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 n s_x^2 = b^2 SCE_x \end{aligned}$$

Donc $r^2 = \frac{b^2 n s_x^2}{n s_y^2} = \left(\frac{\text{cov}(x, y)}{s_x^2} \right)^2 \frac{s_x^2}{s_y^2} = \frac{(\text{cov}(x, y))^2}{s_x^2 s_y^2} = (r)^2$

$$r^2 = \frac{SCE_{reg.lin.}}{SCE_T} \quad \text{En particulier, } r = 0 \Leftrightarrow r^2 = 0$$



Tests

Test de la décomposition de la variation ou analyse de variance (ANOVA): $H_0 : \rho^2 = 0$

$$\frac{\sigma_{reg.lin.}^2}{\sigma_{horsreglin.}^2} = \frac{SCE_{reg.lin.}/1}{SCE_{horsreglin.}/(n-2)} : F_{n-2}^1$$

NB: $\frac{SCE_{reg.lin.}/1}{SCE_{horsreglin.}/(n-2)} = \frac{r^2 SCE_T}{(1-r^2)SCE_T/(n-2)} = \left(\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right)^2$

$$\frac{SCE_{reg.lin.}/1}{SCE_{horsreglin.}/(n-2)} : F_{n-2}^1$$

numériquement
équivalent à

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} : T_{n-2}$$



Autres tests

- comparaison de la pente à une valeur non nulle
- comparaison de l'ordonnée à l'origine à une valeur quelconque

Chapitre 3

Analyse en Composantes Principales A.C.P.

N. IDRISI
Faculté des Sciences et Techniques
Beni Mellal
Département d'Informatique



Introduction

L'ACP, introduite par K. Pearson et Thurston (années 20), est une technique des statistiques descriptives destinée à l'analyse des données multidimensionnelles.

- ➔ Comprendre la structure d'un ensemble de variables (regroupement, points isolés, ...)
- ➔ Réduire la dimension de l'espace des descripteurs (variables) avec le minimum de perte d'information

Position du Problème

Sujet	Math	Sciences	Français	Latin	Musique
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

Rappels

Matrice de variance-covariance : mesure la liaison entre les différents descripteurs

$$\Sigma = \left(\text{cov}(X_i, X_j) \right)_{i,j}$$

où $\text{cov}(X_i, X_i) = \text{Var}(X_i)$.

Matrice de corrélation : $R = (R_{ij})_{i,j} = \Sigma / \delta_{xi} \delta_{xj}$

Matrice de corrélation

1	0,970	-0,064	0,094
--	1	-0,102	0,037
--	--	1	0,986
--	--	--	1

Commentaires

Le tableau initiale (ind x var) est difficile à lire (en particulier lorsqu'on a plusieurs variables et sujets, n,p >>>).

Par conséquent les relations entre les différentes variables sont indétectables à première vue.

La matrice de corrélation montre les variables qui sont fortement corrélées entre elles.

Comment se fait la réduction de la dimension tout en préservant les liaisons entre les différentes variables?

- Les variables de départ sont remplacées par « des vecteurs propres » de la matrice Σ ou de la matrice R , appelés **Composantes principales**.
- **Y-a-t-il un critère d'arrêt ?** généralement on s'arrête quand au moins 75% de la variance est expliquée par la variance cumulée par les CP.
 - Ou en appliquant le critère de Kaiser (80% de l'information est gardée ou de valeur propre $>=1$)



Qu'est-ce qu'un vecteur propre ?

λ est une **valeur propre** de la matrice A si et seulement si $A\mathbf{v} = \lambda\mathbf{v}$

Le vecteur \mathbf{v} dans la relation ci-dessus est appelé **vecteur associé à λ** .
Les valeurs propres s'obtiennent en résolvant le système d'équations $\det(A - \lambda I) = 0$.

Le nombre de valeurs propres, $\lambda_1 > \dots > \lambda_p$, est égal au nombre de colonnes de la matrice A

→ **La somme** des valeurs propres de A est égale à la **variance** contenue dans l'ensemble des données.



Expression des composantes principales

D'un point de vue pratique les composantes principales s'écrivent

$$F_j = \lambda_1 X_1 + \dots + \lambda_p X_p$$

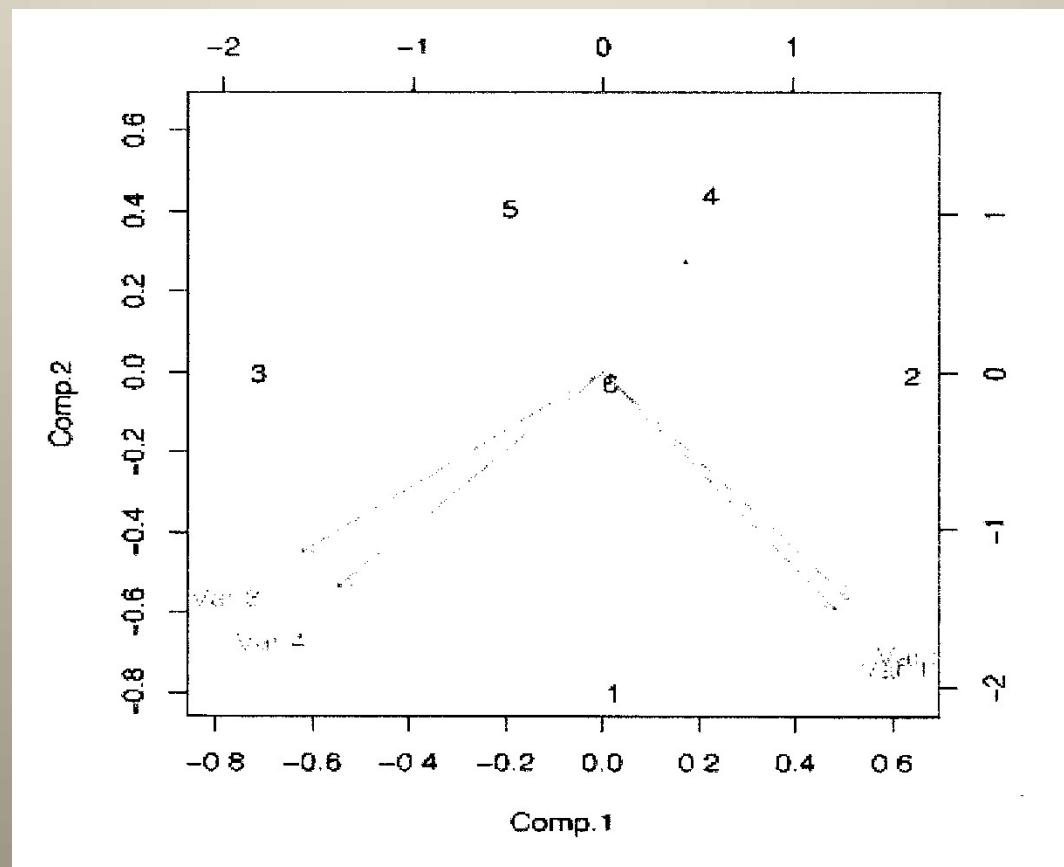
c'est-à-dire que F_j est une combinaison linéaire des variables initiales X_1, \dots, X_p .

En plus de cet aspect calculatoire on doit pouvoir faire des affirmations sur la qualité de la réduction et la qualité de la représentation graphique.



Représentation graphique

Lorsque les différentes CP ont été trouvées on peut représenter les différentes variables et les différents individus dans le plan CP1, CP2 comme illustré ci-dessous



Interprétation

Chaque valeur propre représente la variance prise en compte par la composante principale correspondante.

Par exemple:

	CP_1	CP_2	CP_3	CP_4
Valeur propre	2.0011	1.8668	0.0317	0.0003
Prop. variance	0.5003	0.4917	0.0079	0.0001
Prop. cumulée	0.5003	0.9920	0.9999	1.0000

Ici les deux premières composantes rendent compte de $0,5003+0,4917 = 0,9920 = 99,2\%$ de la variance totale.

Ce qui veut dire que les 4 variables peuvent être remplacées par les 2 premières composantes (CP_1 , CP_2) tout en préservant la quasi-totalité de l'information (réduction).

Résultats des calculs

Scores des individus : il s'agit des valeurs prises par les composantes principales sur les individus.

Ici

Suj	CP_1	CP_2	CP_3	CP_4
s1	0.0771	-2.7515	-0.0935	0.0166
s2	2.2153	-0.0327	0.1778	-0.0095
s3	-2.4608	-0.0173	0.2445	-0.0036
s4	0.7734	1.5097	0.0664	0.0219
s5	-0.6606	1.3926	-0.2592	0.0064
s6	0.0556	-0.1008	-0.1360	-0.0319

Résultats (suite I)

Saturations des variables : il s'agit des coefficients de corrélation entre les variables et les composantes principales.

Var	CP_1	CP_2	CP_3	CP_4
Z_1	0.6288	-0.7687	-0.1169	-0.0048
Z_2	0.6651	-0.7366	0.1228	0.0030
Z_3	-0.8094	-0.5857	0.0413	-0.0119
Z_4	-0.7129	-0.7002	-0.0355	0.0121

La première composante est surtout corrélée avec les deux dernières variables;

La deuxième composante est corrélée avec les deux premières variables et la dernière;

Résultats (suite II)

Contribution (relative) d'un individu à la formation d'une composante principale :

CTR(sujet 1, CP1)=

$$\frac{0,0771^2}{0,0771^2 + \dots + 0,0556^2} = 0,64\%$$

Qualité de la représentation :

pour sujet 1 et CP2

$$QLT = \frac{2751^2}{0,0771^2 + \dots + 0,0166^2} = 0,998$$

Résultats (suite II)

Qualité de la représentation d'une variable à la formation d'une CP :
contribution de la première variable à la formation de la première composante principale

$$\text{CTR} = \frac{0,6288^2}{0,6288^2 + 0,6651^2 + \dots + 0,7129^2} = 0,1976$$

A retenir

Interpréter chaque axe : part de la variance, variables avec lesquelles il est corrélé, contribution.

Individus proches de l'origine : ils ont peu contribué à l'inertie.

Interpréter :

Les regroupements d'individus et variables;

Les oppositions marquées;

Les points isolés



TP

A vos machines !



Chapitre 4

ANALYSE DE LA VARIANCE (ANOVA)



L'Analyse de Variance à un facteur (ANOVA 1)

- A. Introduction**
- B. Formulation du modèle**
- C. Conditions d'application de l'ANOVA**
- D. Mesure de la décomposition de la SCE**
- E. Test de Tukey**



Introduction

21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	A	B	C
Notes	10,11,11 12,13,15	8,11,11,13 14,15,16,16	10,13,14,14 15,16,16
Effectif	6	8	7
Moyenne	12	13	14

"effet d'examinateur"?

Forêt 1	Forêt 2	Forêt 3
23,3	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24,0
25,0	22,5	24,0
26,2	23,5	24,5

"effet de plantation"?



Introduction

Objectif = Quand utiliser l'ANOVA?

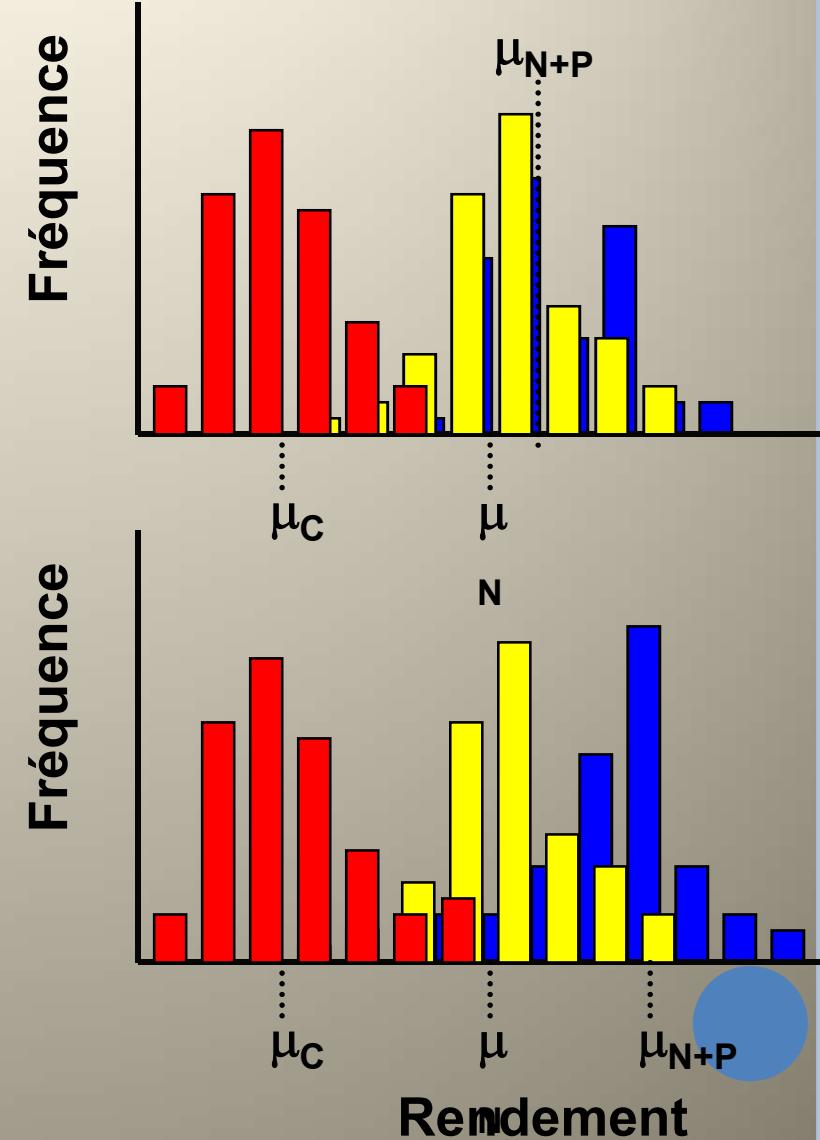
- Pour tester l'effet d'une variable explicative dite **facteur** contrôlé (chaque facteur a k niveaux ou modalités) sur les moyennes d'une variable quantitative Y
- l'ANOVA teste si toutes les moyennes sont égales

ANOVA 1

Possibilités et limites

Permet de tester si toutes les moyennes sont égales (au niveau α)...

...mais si on rejette H_0 , l'ANOVA ne dit pas lesquelles



Les erreurs

		Statistical Decision	
		Reject Null	Retain Null
True Population Status	Null is True	Type I Error α	Correct Decision $1-\alpha$
	Null is False	Correct Decision $1-\beta$	Type II Error β



Types d'ANOVA

Fixe : les traitements sont déterminés (manipulés) par le chercheur

Aléatoires : les modalités sont choisies au hasard dans une population de modalités: on peut estimer l'effet du facteur pour d'autres modalités non étudiées

Données identiques, modèles différents, calculs identiques mais seulement pour l'ANOVA à un critère de classification!



ANOVA fixe : rendement agricole

sable argile terreau

21 16 23

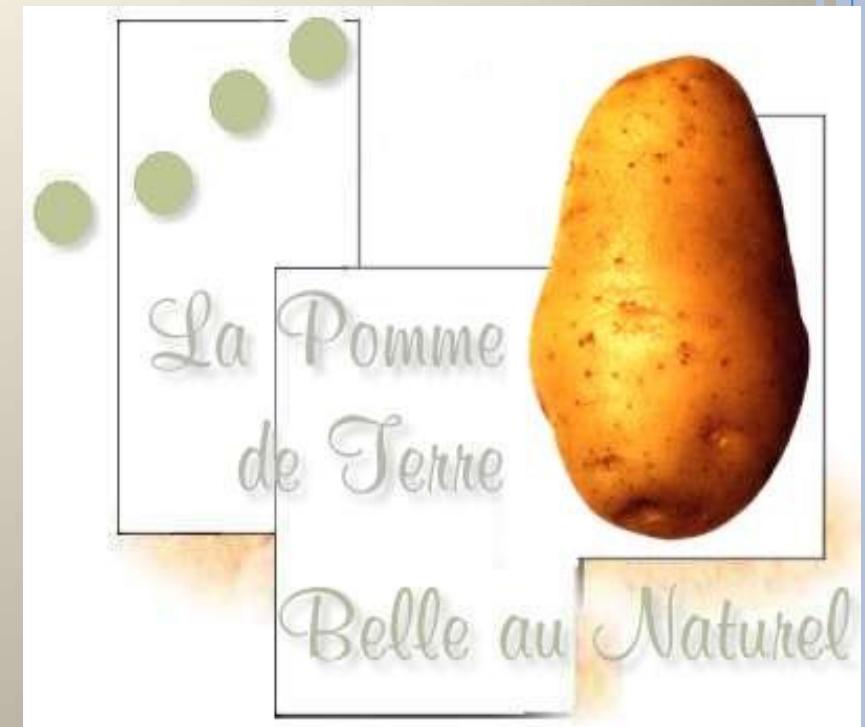
20 18 31

16 11 24

$$n_i \quad [3 \quad 3 \quad 3 \quad 9] = N$$

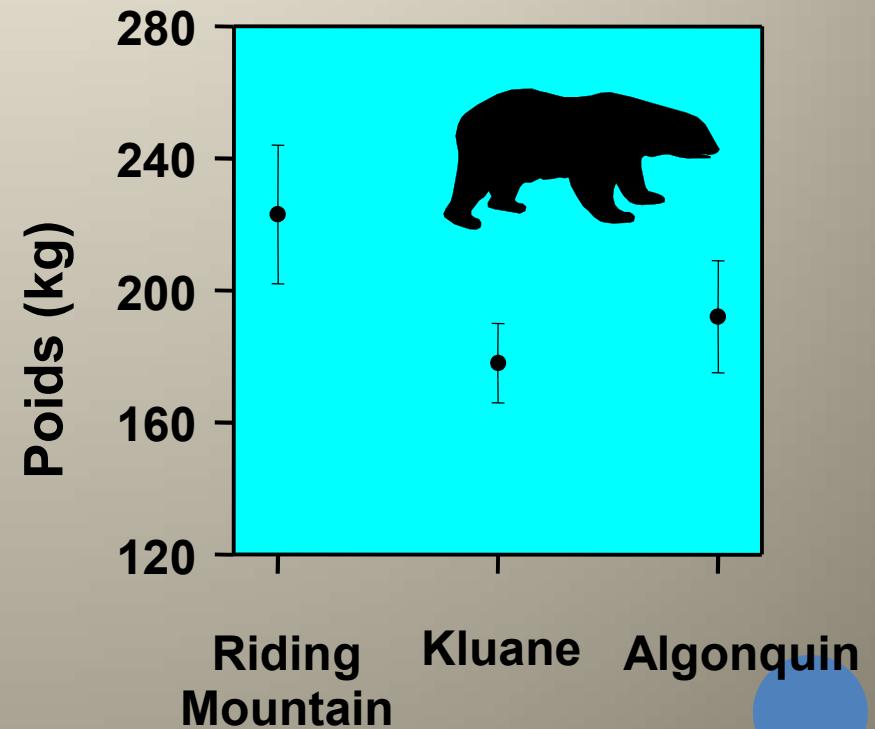
$$T_{i.} \quad 57 \quad 45 \quad 78 \quad 180 = T$$

$$\bar{y}_{i.} \quad 19 \quad 15 \quad 26 \quad 20$$



ANOVA aléatoire: poids de l'ours noir

- variable dépendante est le poids,
- facteur (X) = site, $p=3$
- Question = effet site, au-delà des sites étudiés



Modèle : données

Un seul facteur F

k niveaux

k échantillons de tailles respectives n_1, \dots, n_k

Effectif total

$$n = \sum_{i=1}^k n_i$$

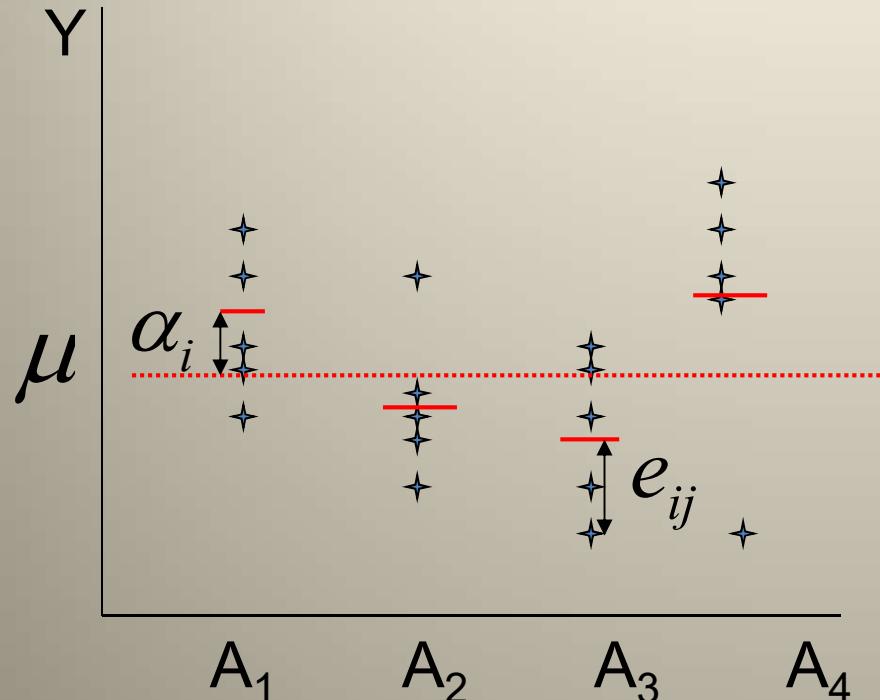
$n_i = n_j \forall i, j \rightarrow$
expérience équilibrée

À chaque expérience, on mesure la valeur de la variable Y .

Données

Niveau (population)	Nb. obs.	Valeurs de Y
1	n_1	$y_{11}, y_{12}, \dots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \dots, y_{2n_2}$
:	:
k	n_k	$y_{k1}, y_{k2}, \dots, y_{kn_k}$

Modèle



Les p moyennes sont-elles identiques?

Les modalités de A influencent-elles Y?

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

μ moyenne de Y

α_i effet de la $i^{\text{ème}}$ modalité (constante). $H_0 : \alpha_i = 0$

e_{ij} erreur aléatoire

- Modèle:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad i=1,\dots,I \text{ et } j=1,\dots,J$$

- Test de comparaison des moyennes :

Hypothèse nulle (H0) : $\mu_1 = \mu_2 = \dots = \mu_I$

Contre (H1) : Les μ_i ne sont pas tous égaux.

=> Utilisation de l'**analyse de la variance à un facteur**.

Conditions d'application de l'ANOVA

les k échantillons sont indépendants et de loi Normale.

Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \sim \mathcal{N}(m_i, \sigma^2)$ et $Y_{ij}, Y_{i'j'}$ indépendantes pour $i \neq i'$ ou $j \neq j'$.

Test de shapiro-wilks (sur les résidus)

Autrement dit, pour chaque i , $(y_{ij})_{j \leq n_i}, \dots, y_{in_i}$ est un échantillon standard.

L'écart-type (théorique) est le même pour tous les niveaux. La moyenne (théorique) peut varier avec le niveau.

Homogénéité des variances ou homoscédasticité.

Test de Bartlett

1. Indépendance :

- Pas de test statistique simple pour étudier l'indépendance.
- Les conditions de l'expérience choisie nous déterminent si nous sommes dans le cas de l'indépendance.

2. Normalité :

Test de **Shapiro-Wilk** sur l'ensemble des résidus

(H0) : les résidus suivent une loi normale

(H1) : les résidus ne suivent pas une loi normale

- Statistique de test :

$$W = \frac{\left(\sum_{i=1}^{[n/2]} a_i (x_{(n-i+1)} - x_{(i)}) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$x_{(i)}$ correspond à la série des données triées, et a_i sont des constantes fournies par des tables spécifiques.

- Décision : On rejette H0 si $W < W_{crit}$.

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk.

3. Homogénéité :

Test de Bartlett :

- Comparaison multiple de variances

$$(H0) : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_I^2$$

$$(H1) : \text{les } \sigma_i^2 \text{ ne sont pas toutes égales}$$

- Statistique de test : $B_{obs} = \frac{1}{C} [(n-1) \ln(s_R^2) - \sum_{i=1}^I (n_i - 1) \ln(s_{c,i}^2)]$

avec $C = 1 + \frac{1}{3(I-1)} \left(\left(\sum_{i=1}^I \frac{1}{n_i - 1} \right) - \frac{1}{n-1} \right)$

et B_{obs} suit une loi du Khi-Deux à $I-1$ ddl.

- Décision : Si $B_{obs} < c \rightarrow (H0)$ vraie

Exemple : forêt

Application à l'exemple :

$$\bar{y}_1 = 24,75$$

$$\bar{y}_2 = 21,53$$

$$\bar{y}_3 = 23,6$$

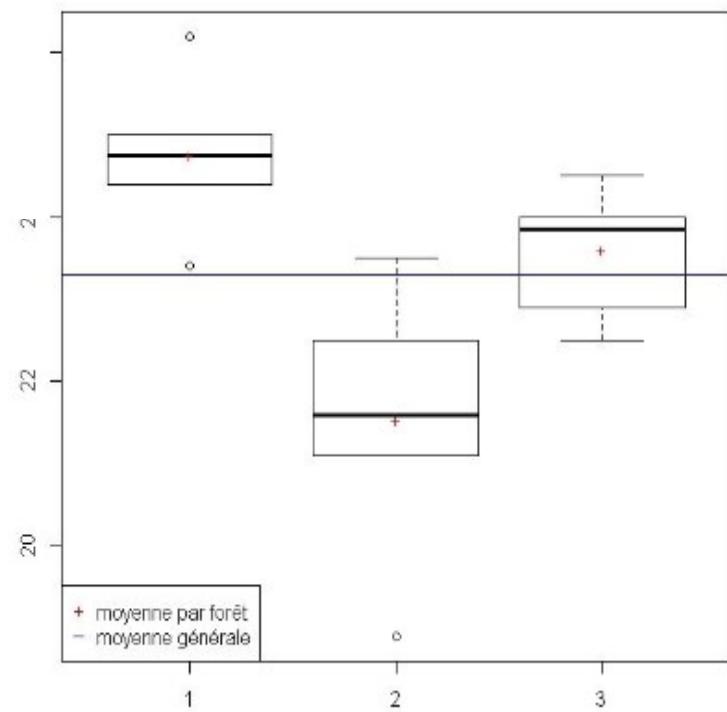
$$s_1 = 0,83$$

$$s_2 = 2,49$$

$$s_3 = 0,57$$

Nombre d'observations : $n = I \cdot J = 6 \cdot 3 = 18$

Hauteur des arbres en fonction des forêts



- **Normalité (Shapiro)** : nombre d'observations trop faible pour tester sur chaque forêt donc on va tester sur tout l'échantillon.

Test de Shapiro-Wilk

W=0.9748

P-value=0.882

p-value = 0.882 > 0.05 donc on accepte H0 => normalité.

- **Homogénéité (Bartlett)** : nombre d'observations trop faible pour tester sur chaque forêt donc on va tester sur tout l'échantillon.

Test de Bartlett

B=2.8279

Df=2

P-value= 0.2432

p-value = 0.2432 donc on accepte H0 => homogénéité des variances

ANOVA 1

Propriété fondamentale

Dans une ANOVA, la variance totale est répartie en deux composantes:

intergroupe: variance des moyennes des différents groupes (modalités)

intragroupe (erreur): variance des observations autour de la moyenne du groupe



Propriété fondamentale

Variation due au facteur :

dispersion des moyennes autour de la moyenne générale.



$$SC_{tot} = SC_F + SC_R$$



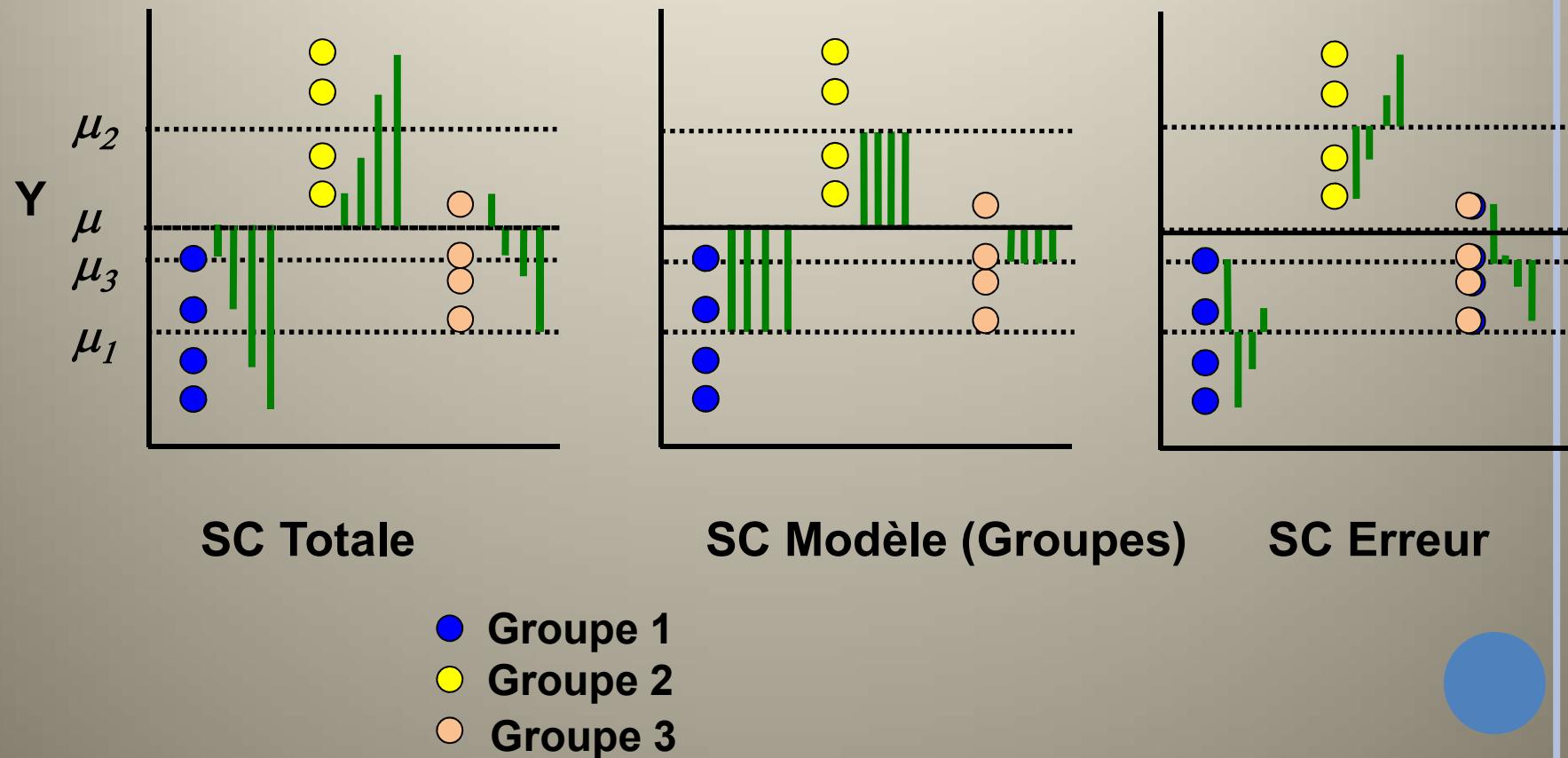
Variation totale :

dispersion des données autour de la moyenne générale.

Variation résiduelle :

dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

Répartition de la somme des carrés totale



$$\begin{aligned}
 \text{SCE}_T &= \text{SCE}_A + \text{SCE}_{R(=E)} \\
 \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 &= \sum_{i=1}^p \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \\
 &= \text{SCE}_{\text{inter}} + \text{SCE}_{\text{intra}} \\
 &= \text{SCE}_B + \text{SCE}_W
 \end{aligned}$$

Tableau d'ANOVA

Sources de variation	Somme des carrés	Degré de liberté (ddl)	Carré moyen (CM)	F
Total	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$n - 1$	SCE_T / dd	
Facteur	$\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$p - 1$	SCE_A / dl	$\frac{CM_A}{CM_R}$
Résidus	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$n - p$	SCE_R / dl	

Décision

TEST DE FISHER:

$$(H_0) : \mu_1 = \mu_2 = \dots = \mu_I$$

(H1) : Les μ_i ne sont pas tous égaux.

Si les 3 conditions (Indépendance, Normalité et Homogénéité) sont vérifiées
et si (H0) est vraie,

Alors :

$$F_{obs} = \frac{CM_F}{CM_R} \sim F_{I-1, n-I}$$

Décision : Pour un seuil donné α (5% en général) les tables de Fisher nous fournissent une valeur critique c telle que :

$$P_{H_0}(F_{I-1, n-1} < c) = 1 - \alpha$$

Alors:

si $F_{obs} < c \rightarrow H_0$ est vraie

si $F_{obs} \geq c \rightarrow H_1$ est vraie

Exemple

	sable	argile	terreau	
$T_i.$	21	16	23	
	20	18	31	
	16	11	24	
	57	45	78	180

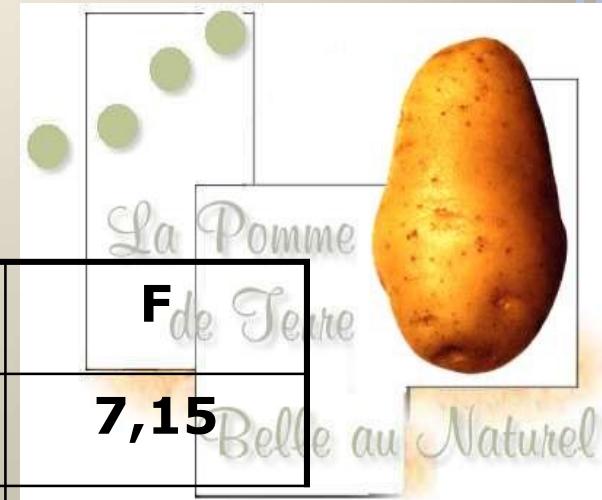
$$SCE_T = 21^2 + 20^2 + \dots - \frac{180^2}{9} = 264$$

$$SCE_A = \frac{57^2}{3} + \frac{45^2}{3} + \dots - \frac{180^2}{9} = 186$$



Exemple

SV	SCE	ddl	CM	F
A	186	2	93	7,15
R	78	6	13	
T	264	8		



$$F_{6,\alpha=0,05}^2 = 5,14$$

→ Conclusion ?



Comparaison multiple

But : classer les traitements par groupes qui sont significativement différents.

- Test de Tukey : test de la différence franchement significative (HSD= honestly significant difference)

- S'applique sur un facteur si :
 - Les 3 conditions fondamentales sont vérifiées,
 - Le facteur est à effet fixe, avec au moins 3 modalités,
 - Le facteur a un effet significatif sur la réponse.

Méthode :

- Pour chaque paire i et l de groupes, on calcule un IC de niveau $(1-\alpha)\%$ de la différence $(\mu_i - \mu_l)$.
- Si zéro appartient à l'IC, les moyennes ne sont pas jugées significativement différentes au niveau α .

Exemple :

	Diff	Lower	Upper	P-value
2-1	-3.22	-4.92	-1.51	0.0005
3-1	-1.15	-2.86	0.56	0.22
3-2	2.07	0.36	3.77	0.02

0 est dans l'intervalle de confiance de 3-1 → les hauteurs moyennes dans les forêts 1 et 3 ne sont pas significativement différentes.

CHAPITRE 5

ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)

L'Analyse Factorielle des Correspondances (AFC)

- A. Introduction & données
- B. Objectifs de l'AFC
- C. Tableaux de profils
- D. Interprétation de l'AFC
- E. Graphiques



109

A. Introduction

	none	light	medium	heavy
SM	4	2	3	2
JM	4	3	7	4
SE	25	10	12	4
JE	18	24	33	13
SC	10	6	7	2

Smoke dataset

Roman	!	?	,	;	:	-	-
1. Thérèse Raquin	3468	236	138	76	6195	691	168	285	543
2. Madeleine Ferrat	5131	362	236	245	8012	922	291	518	1115
3. La fortune des Rougon	6157	238	534	229	11346	936	362	711	1301
4. La curée	4958	443	357	232	11164	738	364	679	1200
5. Le ventre de Paris	5538	534	426	232	13234	1015	318	734	1201
6. La conquête de Plassans	6292	943	756	512	11585	1285	402	1432	1916
7. La faute de l'abbé Mouret	6364	679	859	462	13948	634	377	1067	1564
8. Son excellence Eugène Rougon	7258	728	1002	496	14295	889	543	1469	1907
9. L'assommoir	7820	769	1929	443	19244	1399	436	995	2272
10. Une page d'amour	6206	843	918	492	11953	647	347	1235	1409
11. Nana	7821	1007	1796	611	17881	1087	509	1523	1797
12. Pot Bouille	6875	1045	1873	651	17044	912	675	1669	1935
13. Au bonheur des dames	6916	808	1313	651	18402	972	642	1531	2114
14. La joie de vivre	5803	710	972	623	13917	602	420	1142	1590
15. Germinal	7944	606	1463	729	21388	908	621	1362	2083
16. L'Œuvre	5000	774	1692	668	18292	811	566	1107	1489
17. La terre	6979	957	2307	796	23417	947	657	1681	2113
18. Le rêve	3052	292	385	237	9551	345	230	416	650
19. La bête humaine	5484	601	929	557	18264	673	467	957	1721
20. L'argent	5022	850	1235	569	19267	684	399	1049	1677
21. La débâcle	7440	860	1833	690	26482	832	564	1398	2197
22. Le docteur Pascal	4586	621	1072	464	15598	462	315	955	1218

Les signes de ponctuation chez Zola

Introduite par Guttman, 1941 & Benzécri, 1973, permet une visualisation en 2 dimensions des tableaux de contingence

Généralisation de l'ACP appliquée aux données qualitatives



Données

Tableau de contingence

= Croisement de deux variables qualitatives X (à n modalités) et Y (à p modalités)

		Y					
		1	...	j	...	J	
1		n_{11}		n_{1j}		n_{1J}	
:				:			
X i		n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
:				:			
I		n_{I1}		n_{Ij}		n_{IJ}	
				n_j			n

n_{ij} = Nombre d'observations ayant la modalité x_i de X et y_j de Y.

$n_{i.}$ = effectif marginal : Nombre d'observations ayant la modalité x_i de Y

$n.j$ = effectif marginal : Nombre d'observations ayant la modalité y_j de Y.

Exemple

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	195	62	1	499	44	141	49	65	1056
Patrons	700	354	229	959	185	292	119	140	2978
Cadres.sup	961	471	633	1580	305	360	162	148	4620
Cadre.moy	572	537	279	1689	206	748	155	112	4298
Employes	441	404	166	1079	178	434	178	92	2972
Ouvriers	783	1114	387	4052	497	1464	525	387	9209
Autres.actifs	142	103	210	1133	132	181	46	59	2006
Inactifs	741	332	327	1789	311	236	102	102	3940
Total	4535	3377	2232	12780	1858	3856	1336	1105	31079

- X=CSP, I=8
- Y=hébergement, J=8

B. Objectif de l'AFC?

- ➔ Étudier les correspondances entre les modalités des deux variables qualitatives;
- ➔ Mettre en évidence des liaisons contenues dans un tableau de contingence;
- ➔ Résumer et représenter les principales liaisons pouvant exister entre les modalités de deux variables qualitatives.
écart à l'indépendance, proximité des profils

➔➔ Réduction de la dimension en effectuant la décomposition factorielle des nuages de points associés aux profils lignes et aux profils colonnes du tableau de contingence croisant les modalités des deux variables
(L'AFC est une double ACP sur les deux tableaux de profils).



Le test du khi-deux d'indépendance

Test :

H_0 : Les variables X et Y sont indépendantes (pas de correspondance)

vs

sous H_0 , H_A : Les variables X et Y sont liées entre elles (liaison significative)

Statistique utilisée :

$$\frac{n_{ij}}{ni \cdot nj}$$

$$\frac{n_{ij} - \frac{ni \cdot nj}{n}}{\sqrt{\frac{ni \cdot nj}{n}}}$$

$$\chi^2$$

= Effectif attendu sous l'hypothèse d'indépendance

= Résidu standardisé (moyenne 0, écart-type 1)

$$= \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \frac{ni \cdot nj}{n})^2}{\frac{ni \cdot nj}{n}}$$

Khi-deux

- **L'indice du chi2** : une mesure classique de la liaison entre deux variables qualitatives d'un tableau de contingence ;
- **Coefficient Beta** (β) = $\frac{\chi^2 - (I-1)(J-1)}{\sqrt{(I-1)(J-1)}}$
 - Si $\beta > 3 \rightarrow$ liaison significative



✓ *Mesure de la liaison entre X et Y*

- En probabilité, si il y a indépendance entre X et Y, on a:

$$(1) \quad \forall (i, j) \quad P(X = i \text{ et } Y=j) = P(X = i)P(Y=j)$$

$$(2) \quad \forall (i, j) \quad P(X = i / Y=j) = P(X = i)$$

$$(3) \quad \forall (i, j) \quad P(Y = j / X=i) = P(Y = j)$$

- En statistiques, ces relations équivalent à

$$(1) \quad \forall (i, j) \quad f_{ij} = f_{i\cdot} \times f_{\cdot j} \quad (\Leftrightarrow f = p, \quad f = (f_{ij}); p = (f_{i\cdot} \times f_{\cdot j}))$$

$$(2) \quad \forall (i, j) \quad f_i^j = f_{i\cdot} \quad (\Leftrightarrow Y^{(j)} = p_X, \quad p_X = (f_{i\cdot}))$$

$$(3) \quad \forall (i, j) \quad f_i^j = f_{\cdot j} \quad (\Leftrightarrow X^{(i)} = p_Y, \quad p_Y = (f_{\cdot j}))$$

- Conclusion : lorsque X et Y sont indépendants
 - Toutes les distributions conditionnelles de X/Y (profils colonnes) sont égales et égales à la loi marginale de X (profil moyen)
 - Toutes les distributions conditionnelles de Y/X (profils lignes) sont égales et égales à la loi marginale de Y (profil moyen)
 - La distribution jointe (fréquence) est égale au produit des marginales (fréquences marginales)

La mesure de la liaison entre X et Y se fait en évaluant un de ces écarts entre distributions

Tableaux utilisés

119

- Tableau de fréquence
- Tableau Profils-lignes
- Tableau Profils-colonnes



Fréquences

$$\text{Fréquence de } (i,j) : f_{ij} = \frac{n_{ij}}{n_{i\cdot}}$$

$$\text{Fréquence marginale de la modalité } i : f_{i\cdot} = \frac{n_{i\cdot}}{n}$$

$$\text{Fréquence marginale de la modalité } j : f_{\cdot j} = \frac{n_{\cdot j}}{n}$$

$$\text{Fréquence conditionnelle } j \text{ sachant } i : f_i^j = \frac{n_{ij}}{n_{i\cdot}} = \frac{f_{ij}}{f_{i\cdot}}$$

$$\text{Fréquence conditionnelle } i \text{ sachant } j : f_j^i = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}$$

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij} = \sum_{i=1}^I n_{i\cdot} = \sum_{j=1}^J n_{\cdot j} = n \quad ; \quad \sum_{i=1}^I \sum_{j=1}^J f_{ij} = \sum_{i=1}^I f_{i\cdot} = \sum_{j=1}^J f_{\cdot j} = \sum_{j=1}^J f_i^j = \sum_{i=1}^I f_j^i = 1$$

$$\sum_{j=1}^J n_{ij} = n_{i\cdot} ; \sum_{i=1}^I n_{ij} = n_{\cdot j} ; \sum_{j=1}^J f_{ij} = f_{i\cdot} ; \sum_{i=1}^I f_{ij} = f_{\cdot j}$$

Tableau profil-lignes

			Y		
	1	...	j	...	J
X	1				
i				:	
			n_{ij}/n_i		
				:	
I					

$\} f_j^i$

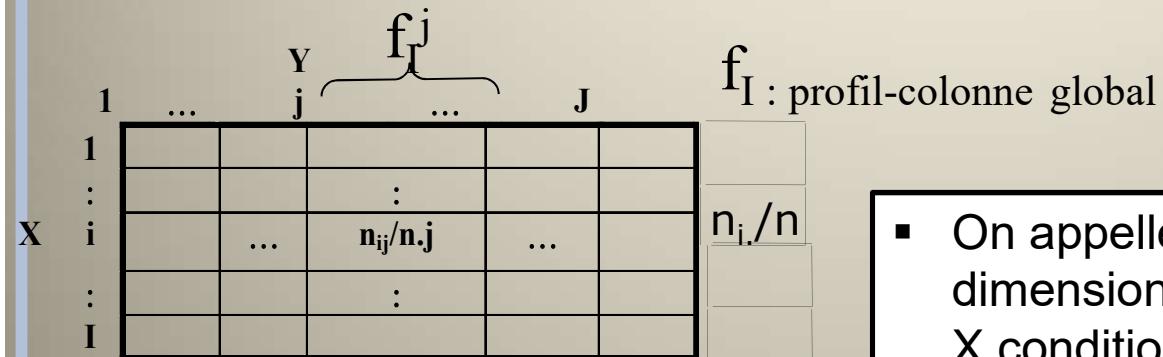
- On appelle i° profil ligne, le vecteur de dimension J des fréquences de la variable Y conditionnellement à la valeur x_i de X ;
- → La répartition des valeurs de Y dans les différentes modalités de X .

$\} f_J : \text{profil-ligne global}$

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	0,184659091	0,058712121	0,00094697	0,472537879	0,041666667	0,133522727	0,046401515	0,06155303	1
Patrons	0,235057085	0,118871726	0,076897246	0,322028207	0,06212223	0,098052384	0,039959704	0,047011417	1
Cadres.sup	0,208008658	0,101948052	0,137012987	0,341991342	0,066017316	0,077922078	0,035064935	0,032034632	1
Cadre.moy	0,133085156	0,124941833	0,064913913	0,392973476	0,047929269	0,174034435	0,036063285	0,026058632	1
Employes	0,148384926	0,135935397	0,055854643	0,363055182	0,059892328	0,14602961	0,059892328	0,030955585	1
Ouvriers	0,085025519	0,120968618	0,042024107	0,440004344	0,053968943	0,158974916	0,057009447	0,042024107	1
Autres.actifs	0,070787637	0,051345962	0,104685942	0,564805583	0,065802592	0,090229312	0,022931206	0,029411765	1
Inactifs	0,188071066	0,084263959	0,082994924	0,454060914	0,07893401	0,059898477	0,025888325	0,025888325	1
Total	1,253079138	0,796987669	0,565330733	3,351456926	0,476333356	0,938663939	0,323210747	0,294937493	8

- 23.5% des patrons vont à l'hôtel. Les patrons vont plus fréquemment à l'hôtel que dans les autres professions.. Dans toutes les professions on va majoritairement chez les parents.

Tableau profil-colonnes



- On appelle j° profil colonne, le vecteur de dimension I des fréquences de la variable X conditionnellement à la valeur j de Y ;
- → La répartition des valeurs de X dans les différentes modalités de Y .

CSP	Hotel	Location	Res.Second	Parents	Amis	Camping	Sej.org	Autres	Total
Agriculteurs	0,042998897	0,018359491	0,000448029	0,039045383	0,023681378	0,03656639	0,036676647	0,058823529	0,256599744
Patrons	0,154355017	0,104826769	0,102598566	0,075039124	0,099569429	0,075726141	0,089071856	0,126696833	0,827883735
Cadres.sup	0,211907387	0,139472905	0,283602151	0,123630673	0,164155005	0,093360996	0,121257485	0,133936652	1,271323253
Cadre.moy	0,126130099	0,159016879		0,125	0,132159624	0,110871905	0,193983402	0,116017964	0,101357466
Employes	0,09724366	0,11963281	0,07437276	0,084428795	0,095801938	0,112551867	0,133233533	0,083257919	0,800523282
Ouvriers	0,172657111	0,32987859	0,173387097	0,317057903	0,267491927	0,37966805	0,392964072	0,350226244	2,383330994
Autres.actifs	0,031312018	0,030500444	0,094086022	0,088654147	0,071044133	0,046939834	0,034431138	0,053393665	0,450361401
Inactifs	0,16339581	0,098312111	0,146505376	0,139984351	0,167384284	0,06120332	0,076347305	0,092307692	0,94544025
Total	1	1	1	1	1	1	1	1	8

- 15.4% des personnes allant à l'hôtel sont des patrons. Parmi les personnes allant à l'hôtel on trouve une majorité de cadres sup, bien que ces derniers aillent préférentiellement chez leurs parents (cf profils lignes).
- La part des cadres sup est plus importante parmi ceux-qui vont en résidence secondaires qu'ailleurs.

A. Indice d'attraction / répulsion:

$$d_{ij} = \frac{f_{ij}}{f_i \cdot f_j}$$

$$d_{ij} \in \begin{cases} > 1 : \text{les modalités } i \text{ et } j \text{ s'attirent} \\ = 1 : \text{indépendance parfaite} \\ < 1 : \text{les modalités } i \text{ et } j \text{ se repoussent} \end{cases}$$



Interprétation = Nuage de points

126

Chaque profil ligne (point--ligne) représente un point dans l'espace de dimension J des profils--colonnes

Chaque profil colonne (point--colonne) représente un point dans l'espace de dimension I des profils--colonnes

Les tableaux des profils lignes et colonnes définissent chacun un nuage de points:

- ❖ Le nuage de points--lignes **N(I)** est constitué des I points--lignes dans l'espace de dimension J des points--colonnes.
- ❖ le nuage de points--colonnes **N(J)** est constitué des J points--colonnes dans l'espace de dimension I des points--lignes.

Notions

- **Poids des points-ligne et points-colonne**

- ✓ Chaque point-ligne $X^{(i)}$ est doté d'un poids relatant l'importance de la modalité i de X:

$$f_{i\cdot} = \frac{n_{i\cdot}}{n}$$

- ✓ Chaque point colonne est doté d'un poids relatant l'importance de la modalité j de Y:

$$f_{\cdot j} = \frac{n_{\cdot j}}{n}$$

- **Centre de gravité**

- ✓ Centre de gravité du nuage N(I) = distribution marginale de Y= p_Y

$$G_X = (g_{X1}, \dots, g_{XJ})$$

$$g_{Xj} = \sum_{i=1}^I f_{i\cdot} f_i^j = \sum_{i=1}^k \frac{n_{i\cdot}}{n} \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} = f_{\cdot j}$$

- ✓ Centre de gravité du nuage N(J) : distribution marginale de X= p_X

$$G_Y = (g_{Y1}, \dots, g_{YI})$$

$$g_{Yi} = \sum_{j=1}^J f_{\cdot j} f_j^i = \sum_{j=1}^k \frac{n_{\cdot j}}{n} \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{n} = f_{i\cdot}$$

- Inertie

- ✓ Distance entre points-lignes : les points lignes étant des distributions, on utilise la métrique du chi2 centrée sur la distribution moyenne $G_X (=p_Y)$

$$\chi^2_{G_x}(X^{(i)}, X^{(i')}) = \sum_{j=1}^J \frac{(f_i^j - f_{i'}^j)^2}{f_{.j}}$$

La distance est d'autant plus grande que i et i' sont réparties de façon différente dans les modalités de Y

Ex : La distance entre deux CSP est d'autant plus grande que ces deux CSP sont réparties de façon différentes dans les lieux de vacances (que les structures diffèrent).

✓ Inertie des nuage $N(I)$ et $N(J)$:

$$I_X = \sum_{i=1}^I f_{i\cdot} \chi_{G_X}^2(X^{(i)}, G_X) = \sum_j \sum_i f_{i\cdot} \frac{(f_i^j - f_{\cdot j})^2}{f_{\cdot j}} = \sum_i \sum_j \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \frac{\chi^2}{n}$$

$$I_Y = \sum_{j=1}^J f_{\cdot j} \chi_{G_Y}^2(Y^{(j)}, G_Y) = \sum_i \sum_j f_{\cdot j} \frac{(f_j^i - f_{i\cdot})^2}{f_{i\cdot}} = \sum_i \sum_j \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \frac{\chi^2}{n}$$

- L'inertie est nulle lorsque tous les profils-lignes (resp. colonnes) sont égaux au centre de gravité \Leftrightarrow lorsque toutes les distributions de Y sachant X=i (resp. X sachant Y=j) sont égales et égales à la distribution marginale de Y (resp. de X) \Leftrightarrow lorsque X et Y sont indépendantes
- Au plus la dépendance est forte, au plus l'inertie est grande

AFC : décomposition factorielle

- Les objectifs de l'analyse factorielle des correspondances (AFC) :
 - ✓ comparer les profils-lignes entre eux (les distributions de Y dans les différentes modalités de X),
 - ✓ comparer les profils-colonnes entre eux (les distributions de X dans les différentes modalités de Y),
 - ✓ Repérer les cases du tableau de contingence où les effectifs observés n_{ij} sont nettement différents des effectifs théoriques (effectifs sous l'hypothèse d'indépendance) pour mettre en évidence les modalités I de X et j de Y qui s'attirent ($f_{ij} > p_{ij}$) et celles qui se repoussent ($f_{ij} < p_{ij}$)
- L'AFC est une ACP sur le tableau de contingence constitué des deux variables X et Y, en utilisant la métrique du chi2.

De façon équivalente, elle consiste en la décomposition factorielle des nuages de points $N(I)$ (analyse directe) et $N(J)$ (analyse duale)

→ Il faut chercher les axes orthogonaux passant le mieux possible par le milieu du nuage de points $N(I)$.

- Chaque axe factoriel supporte une part de l'inertie totale. Cette part est mesurée par les valeurs propres, inférieures ou égales à 1.
- Des valeurs proches de 1 indiquent d'intéressants liens entre modalités de variables différentes.
- Nombre axe $k \leq \min(J-1, I-1)$



MERCI POUR VOTRE ATTENTION!!!

