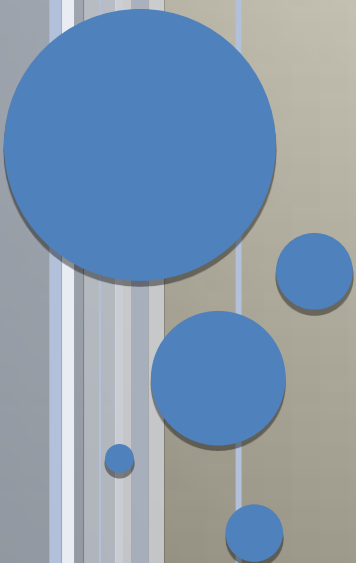


MODULE : ANALYSE DE DONNÉES & STATISTIQUES



IDRISSI NAJLAE
Université Sultan Moulay Slimane
Faculté des Sciences et Techniques
Béni Mellal
Département Informatique
© 2017-2024

PLAN

- Planning
- Objectifs
- Partie 1: ANALYSE DE DONNÉES

Objectifs du module

- ✓ Se familiariser avec le monde statistique en s'initiant aux différentes techniques d'analyse de données ;
- ✓ Être capable de bien mener une analyse statistique de la collecte d'information jusqu'à la prise de décision ;
- ✓ De même, pour le domaine financier, se familiariser avec les différents termes et concepts du domaine ;
- ✓ Être capable de comprendre au mieux le monde des finances et d'effectuer les opérations financières correctement.

- 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

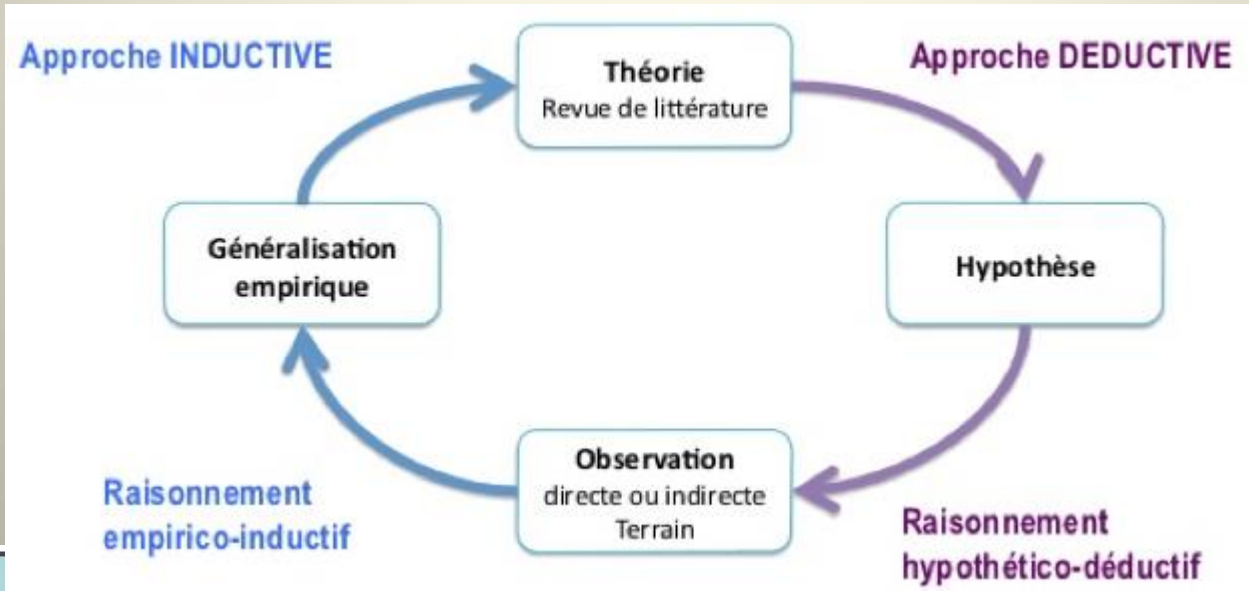


INTRODUCTION

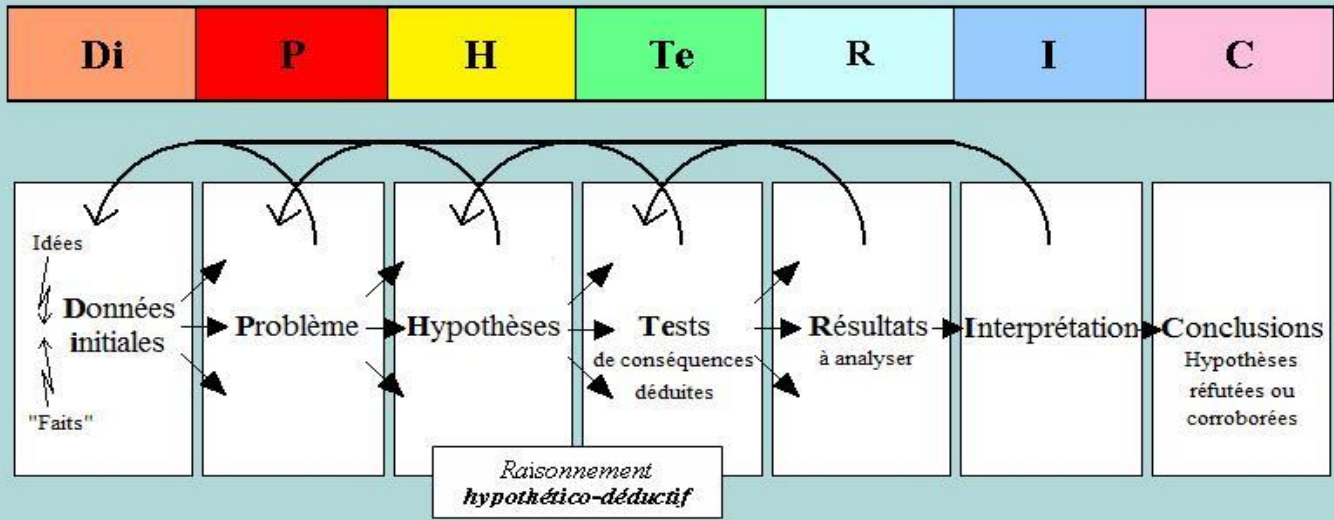
1. Démarches scientifiques
2. Les étapes de l'analyse statistique



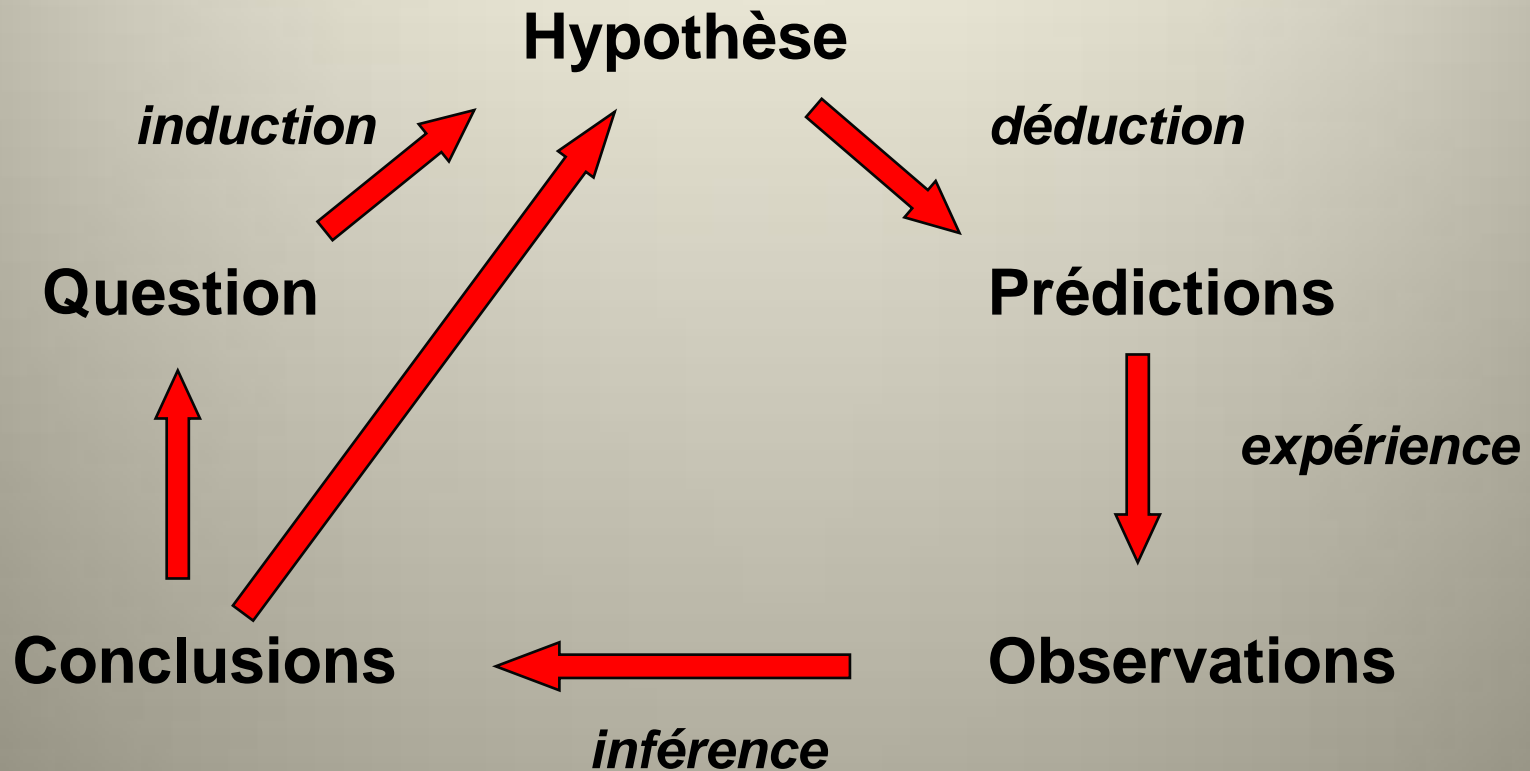
Démarches scientifiques



DiPHTeRIC, modèle de CHEMINEMENT SCIENTIFIQUE



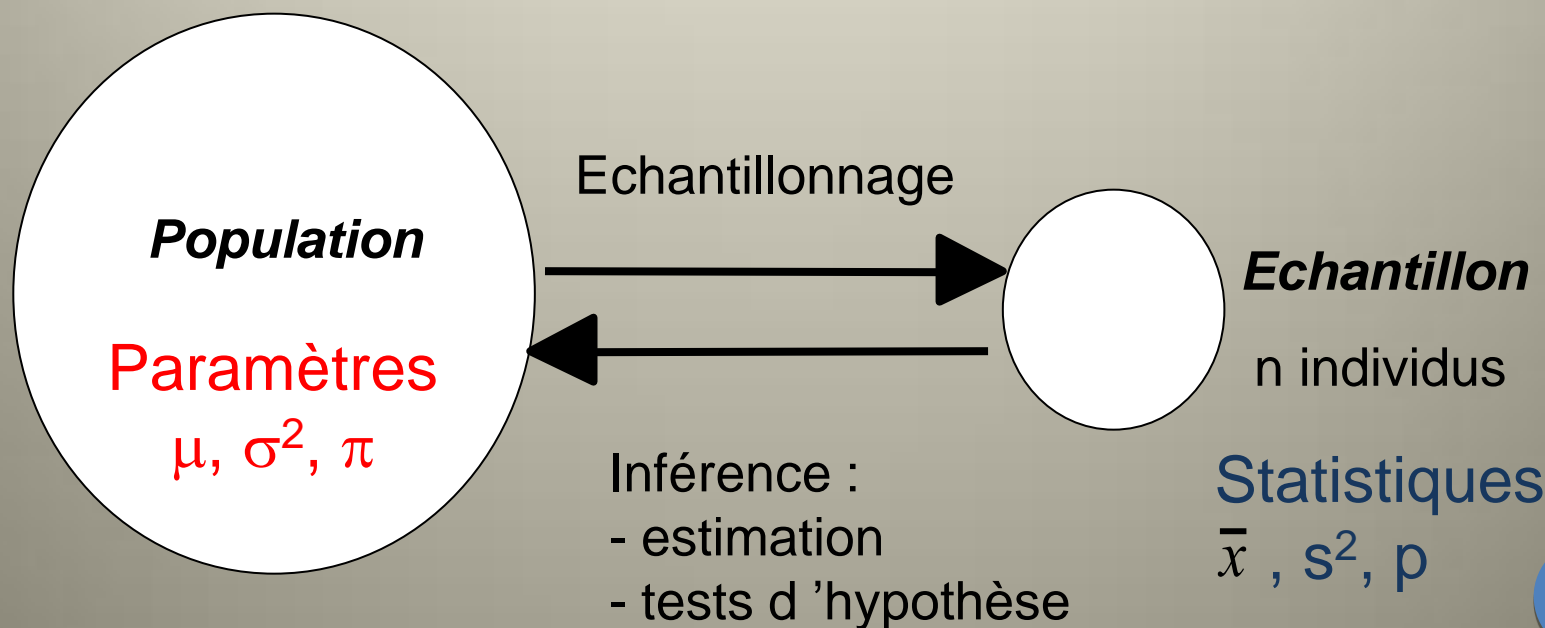
La démarche hypothético-déductive: la falsification d'hypothèses



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 bizzarrite 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

Présupposés : notions de probabilités, notions sur les variables aléatoires

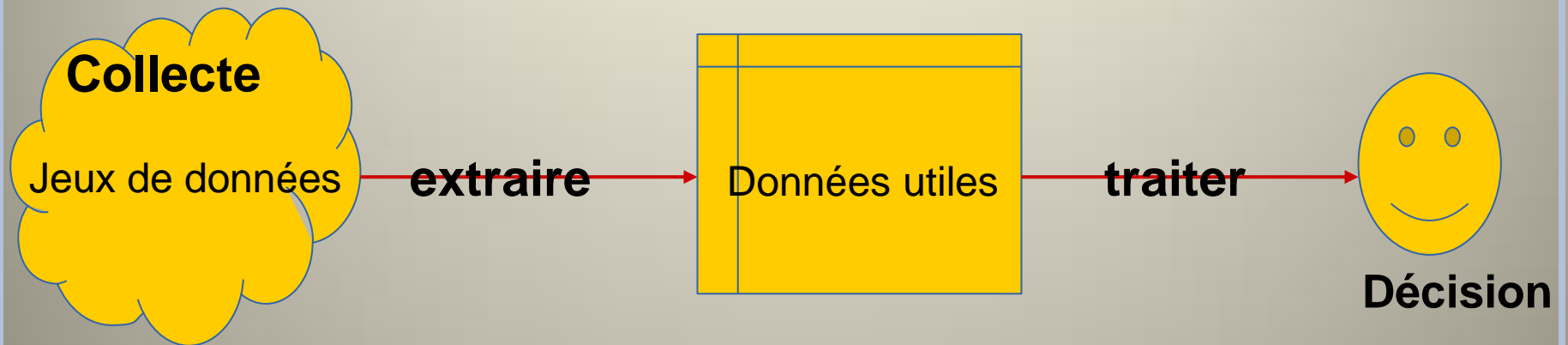
1 - Statistiques descriptives, lois de probabilité, estimation

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, ...

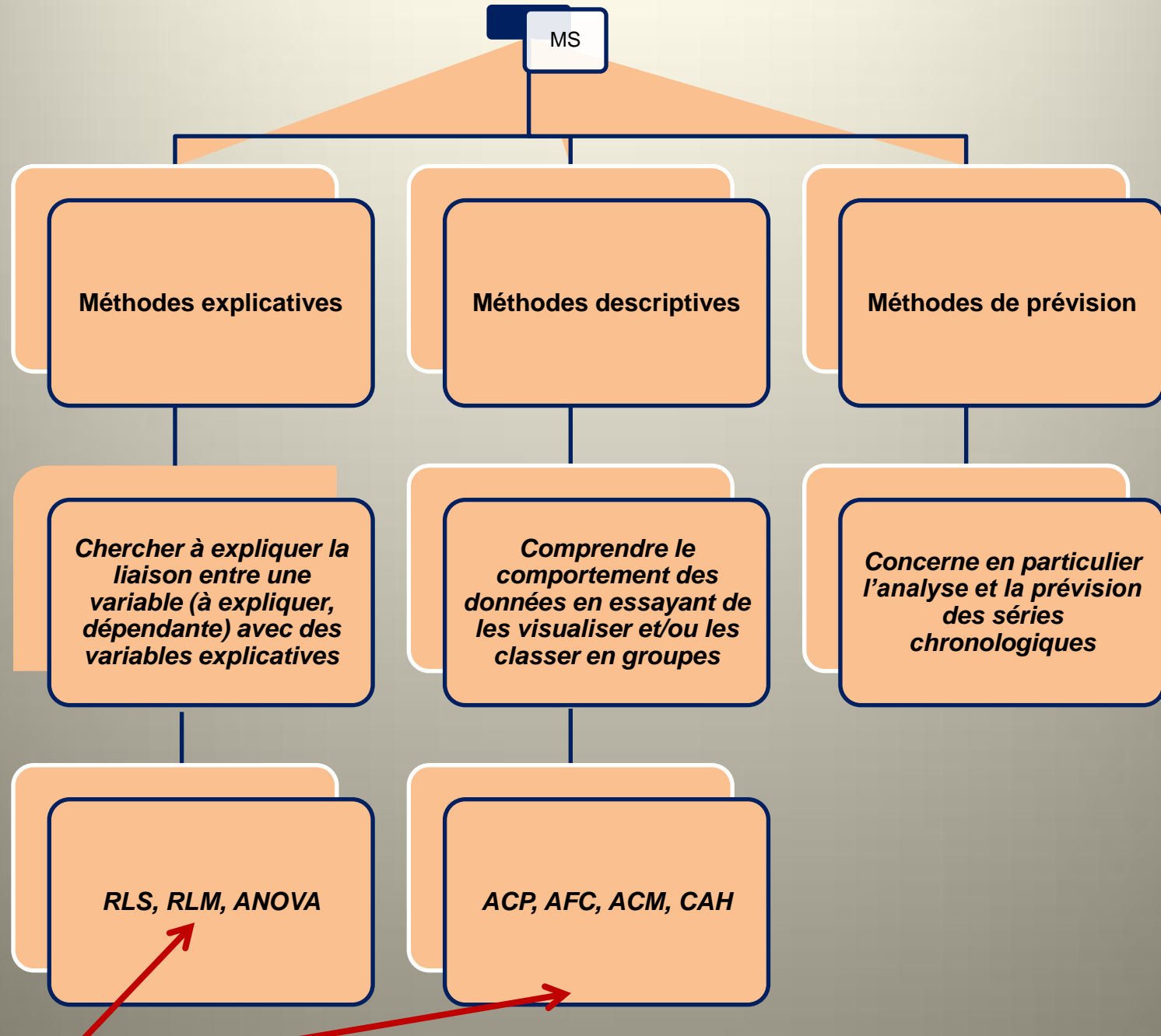
POURQUOI LA STATISTIQUE?

- ✓ ***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.

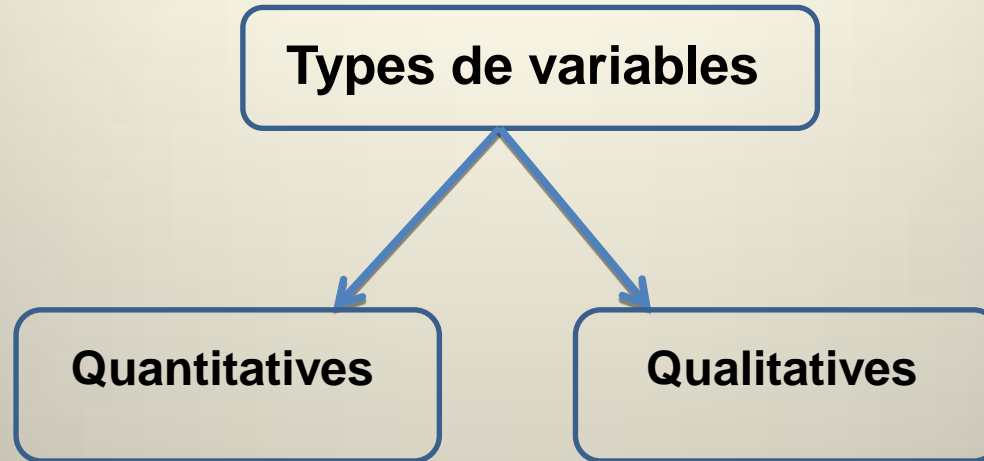
Série statistique :

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

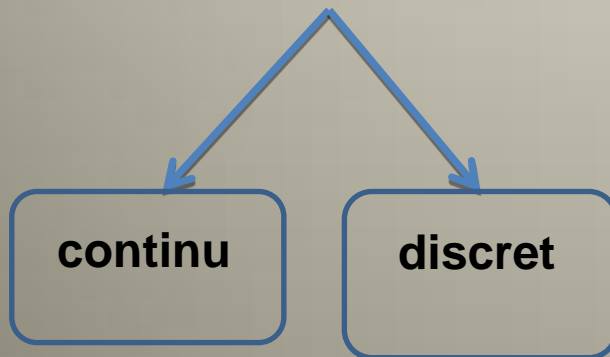
Exemple:

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

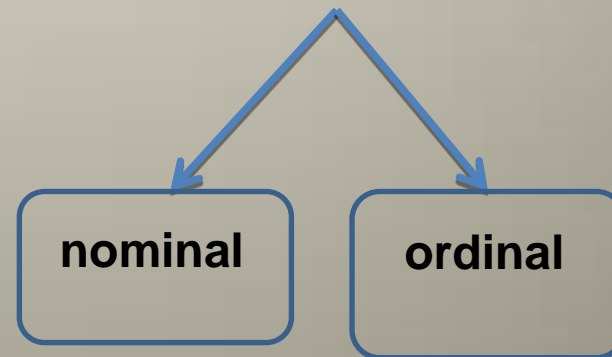


L'ensemble des valeurs numériques ou nombres qu'elle peut de prendre ou mesurer



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

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



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

REPRÉSENTATION DES DONNÉES (PHASE 1)

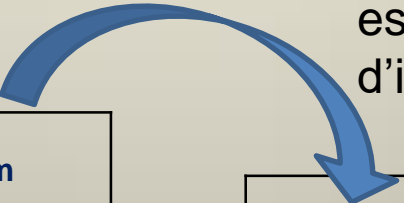
Tableau (ind x vars)

Chaque caractère est observé sur un individu

ind\v ars	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

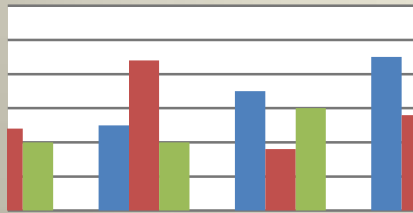
Dans le tableau présenté ci-dessus, il y a :

combien d'individus ?

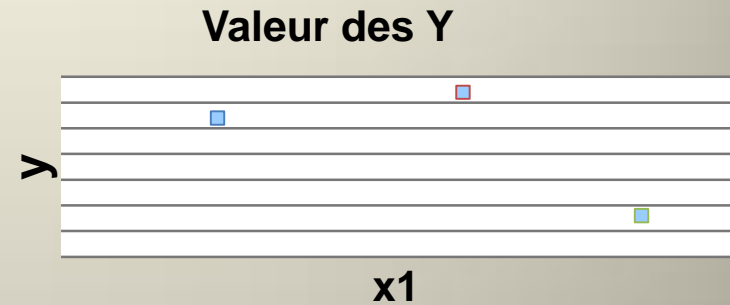
combien de variables ?

1- REPRÉSENTATION GRAPHIQUE

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

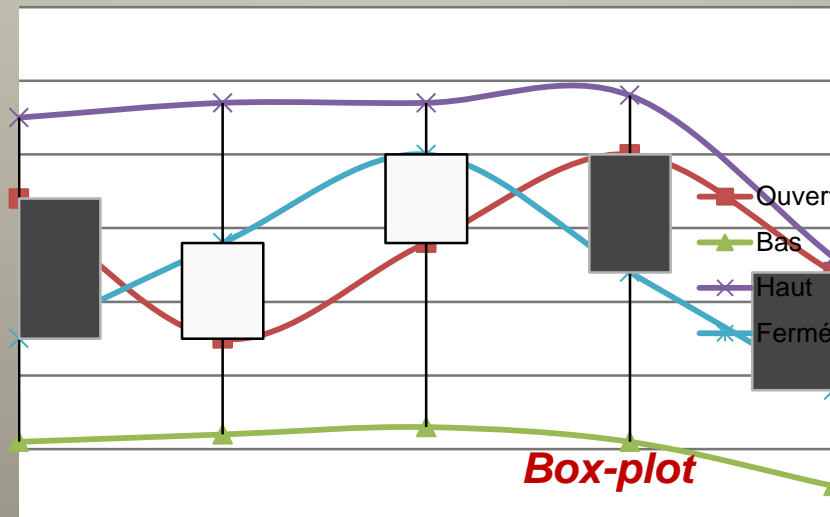


■ Série 1
■ Série 2
■ Série 3



Nuage de points

Histogramme



Box-plot

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

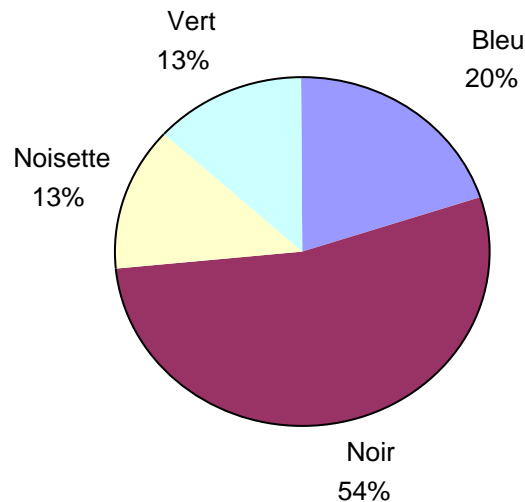


Diagramme circulaire (camembert)

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

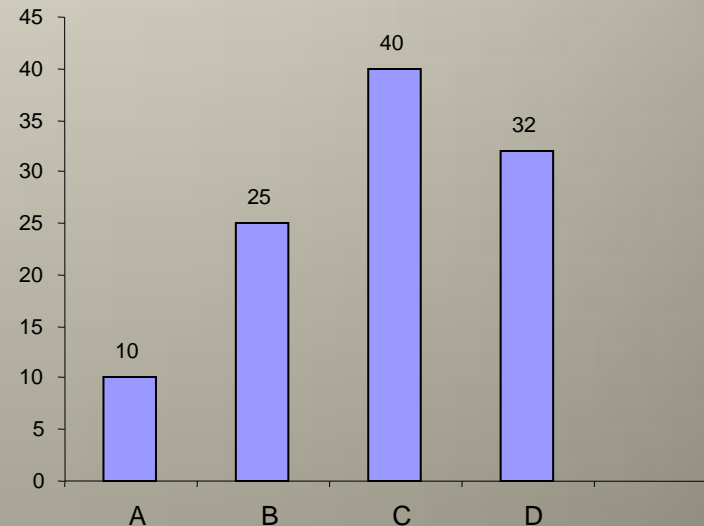


Diagramme en barres ordonné

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

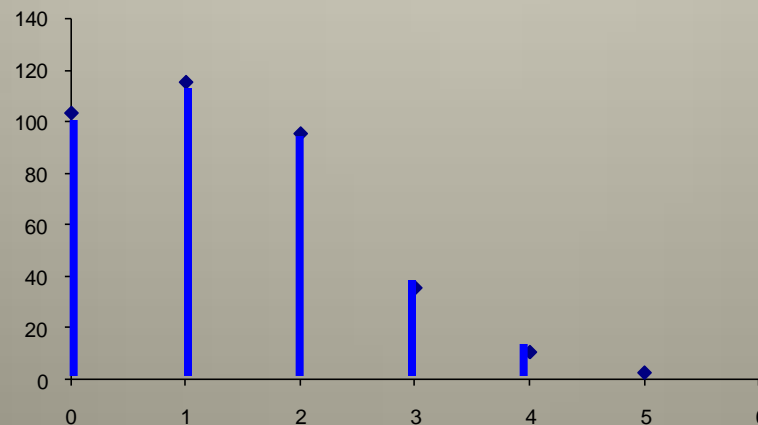


Diagramme en bâtons

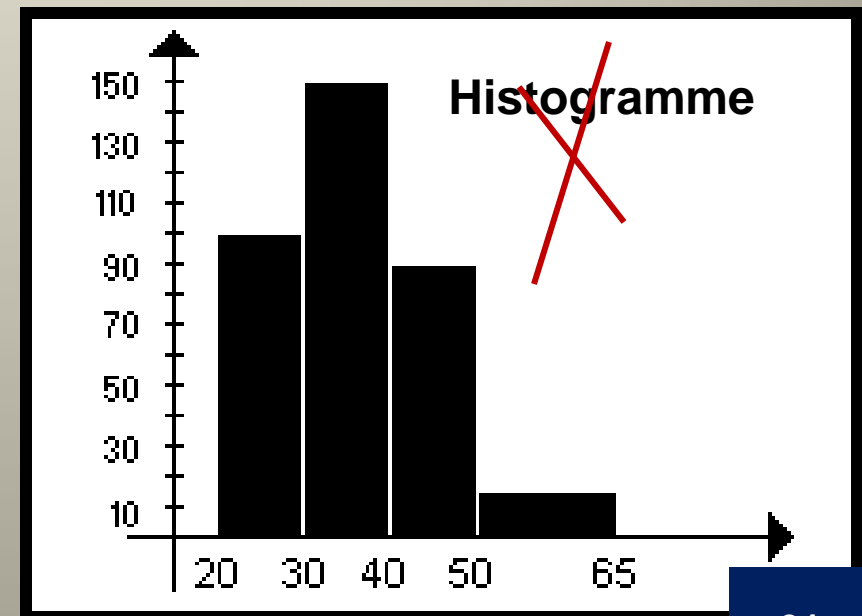
VARIABLES QUANTITATIVES

-- cas continu--

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

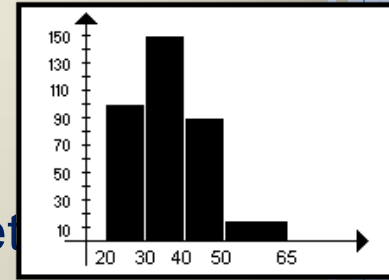
Représentation par intervalles ou classes

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



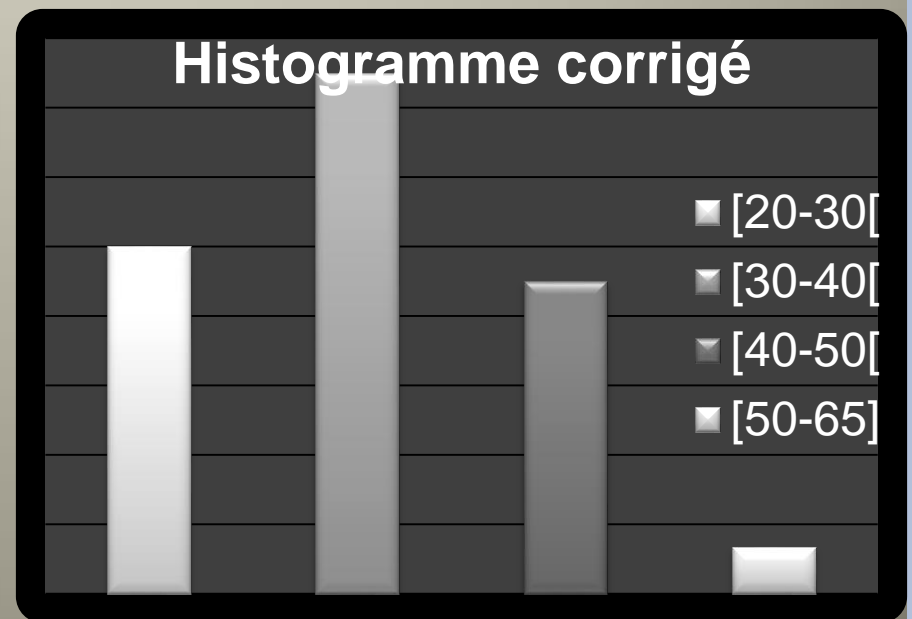
La correction des effectifs ou des fréquences se fait en trois étapes

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

Ordinale

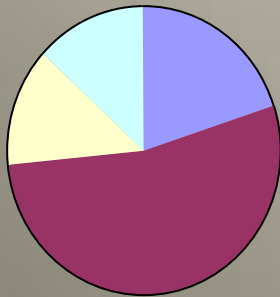
Effectifs ou Fréquences

Diagramme en barres

Diagramme en barres

Modalités dans l'ordre

Diagramme circulaire



VARIABLE QUANTITATIVE

Discrète

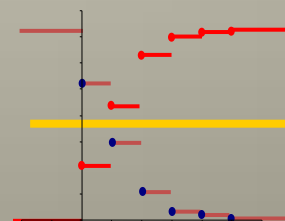
Continue

Effectifs ou Fréquences

Diagramme en bâtons

Histogramme

Courbes cumulatives des effectifs ou des fréquences



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



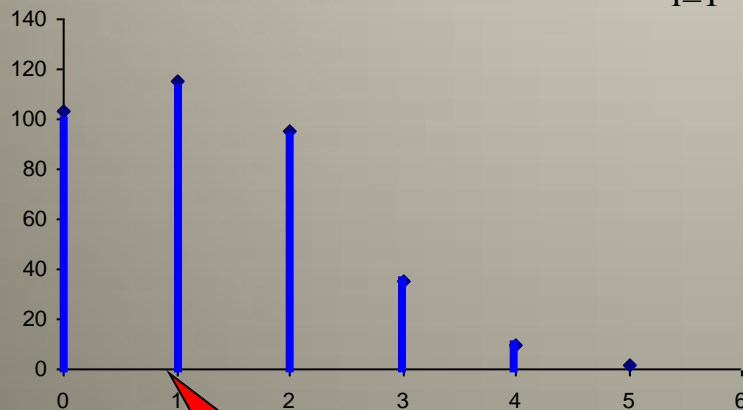
Indicateurs centrales

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

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

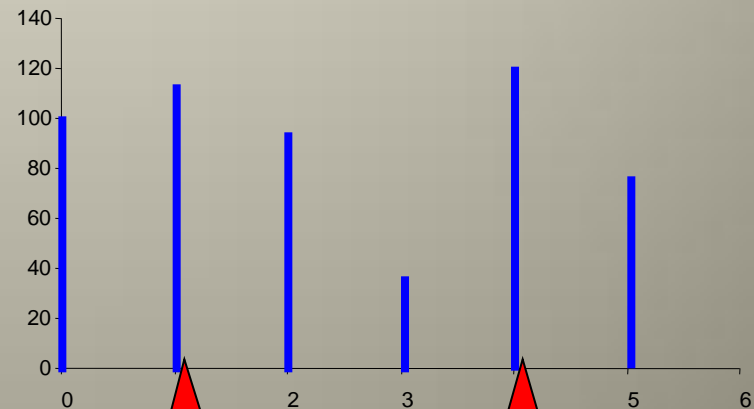
- Mode:** c'est la valeur dont la fréquence est la plus élevée.

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



Mode

distribution unimodale



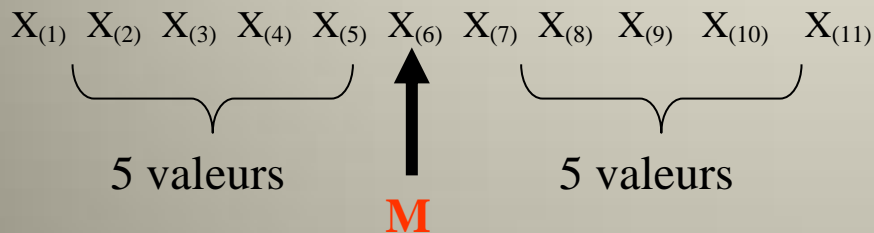
Mode 1

Mode 2

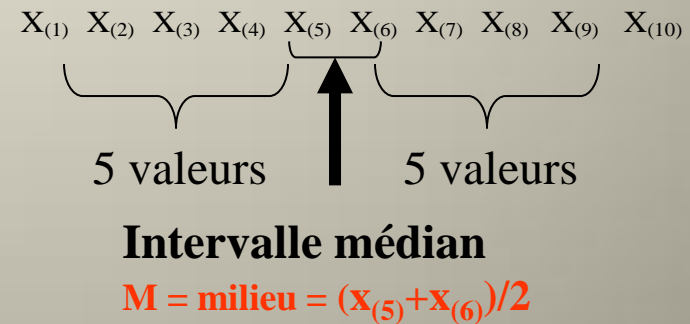
distribution bimodale

- **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

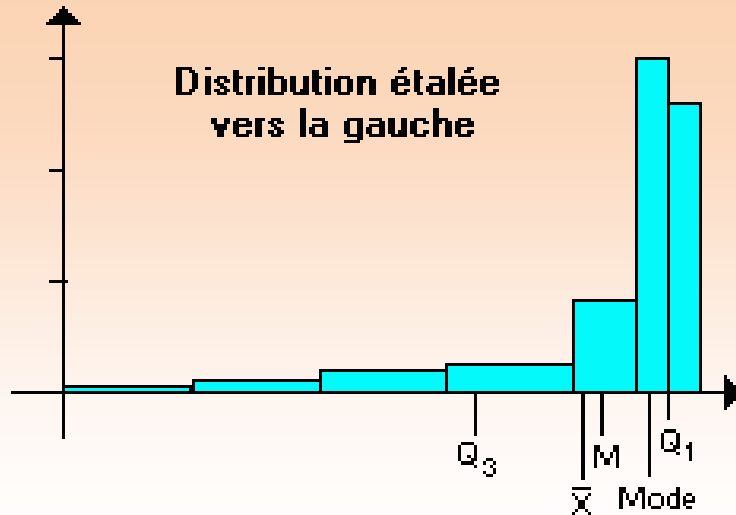


La série $X=(x_1, x_2, \dots, x_n)$ originale doit être triée en $X'=(x_{(1)}, x_{(2)}, \dots, x_{(n)})$

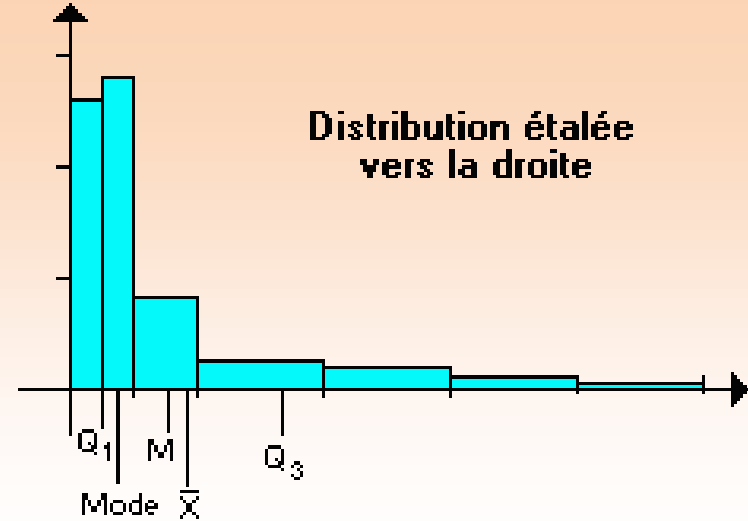


Comparaison

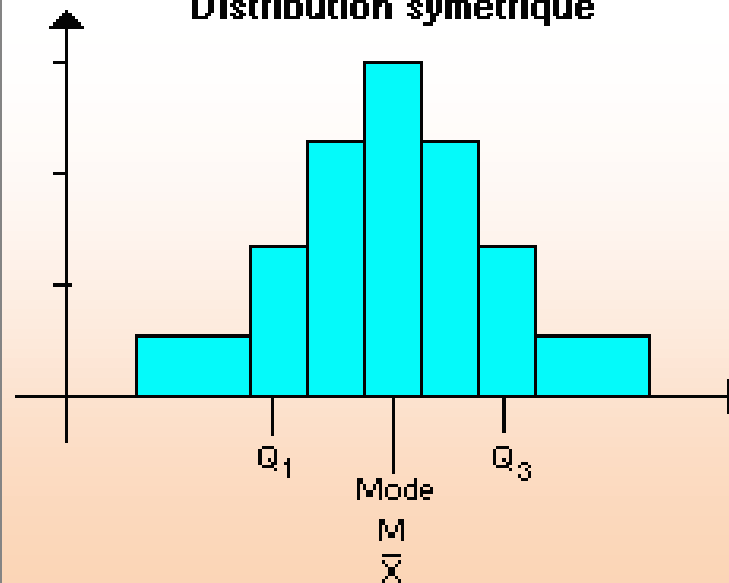
**Distribution étalée
vers la gauche**



**Distribution étalée
vers la droite**



Distribution symétrique



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 laquelle 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 bivariable



Chapitre 2 Régression linéaire simple

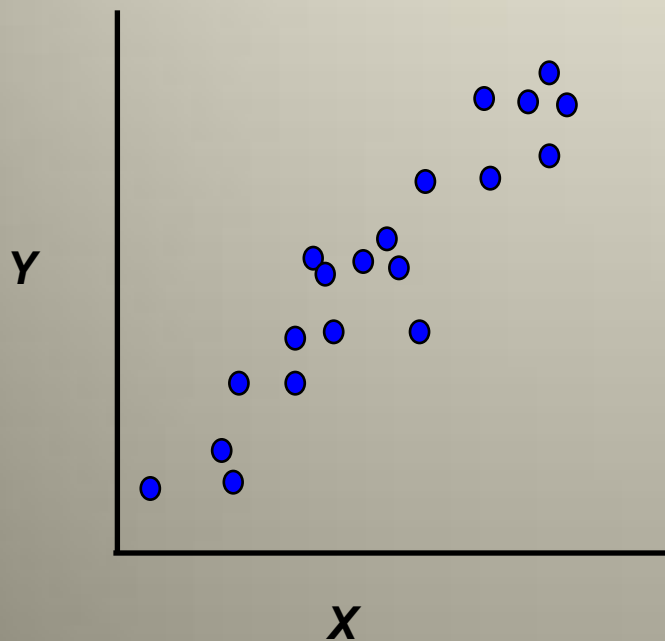


N. IDRISI
FACULTE DES SCIENCES ET TECHNIQUES
DEPARTEMENT INFORMATIQUE
BENI MELLAL
BLOC C RDC

Objectif

Etudier la relation entre deux variables quantitatives:

Nuage de points:

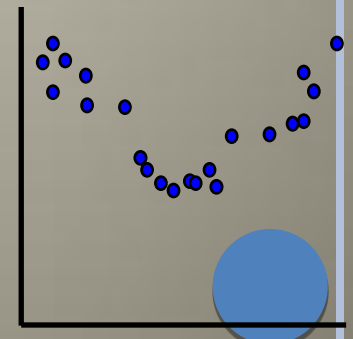
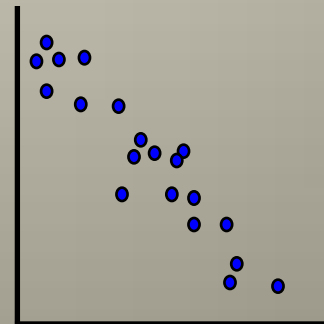
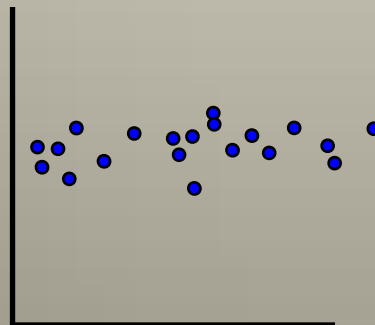
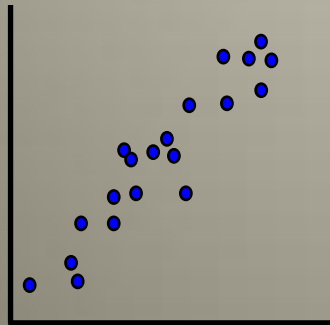
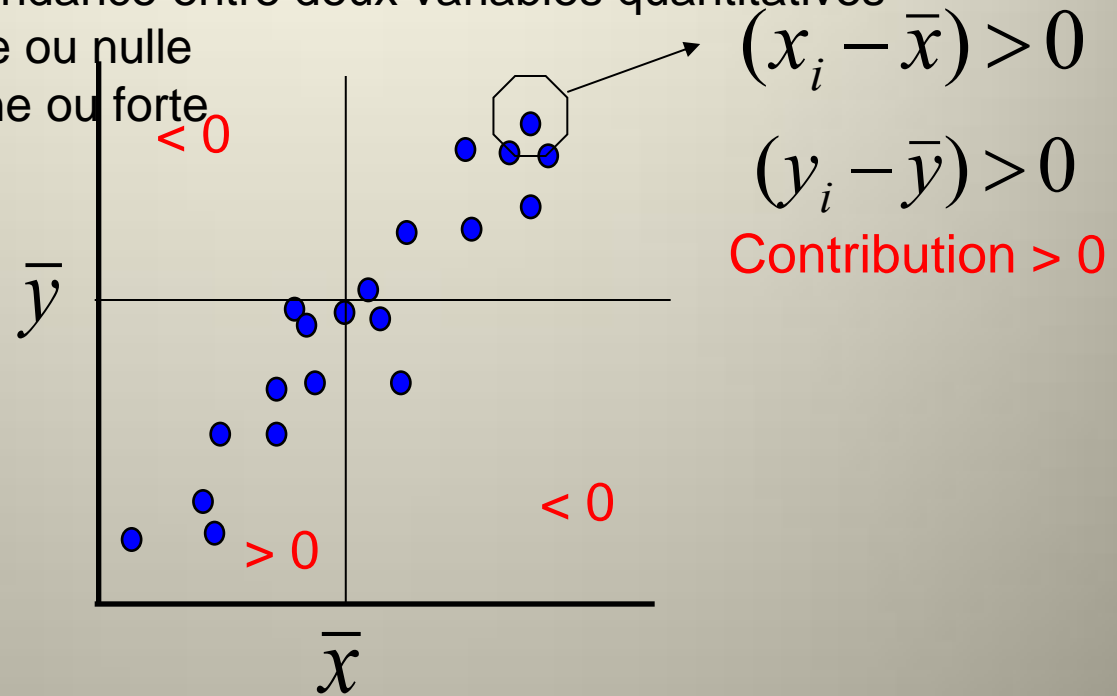


1. description de l'association linéaire: corrélation, régression linéaire simple
2. explication / prédiction d'une variable à partir de l'autre: modèle linéaire simple



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:

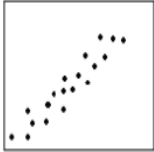
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

Estimation pour la population:

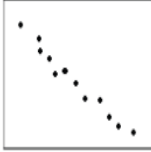
$$\hat{\rho}_{xy} = r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$



Degree of Correlation



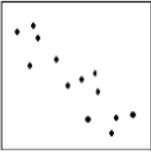
Strong Positive



Strong Negative



Weak Positive



Moderate Negative

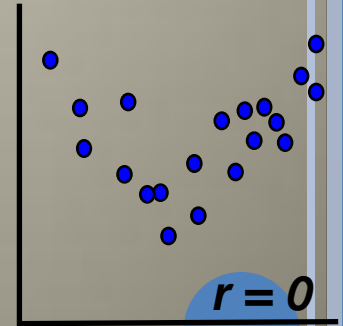
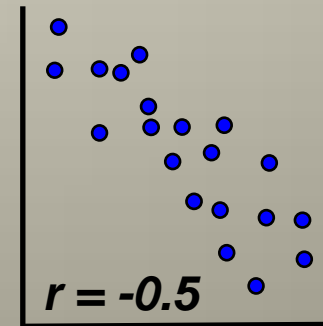
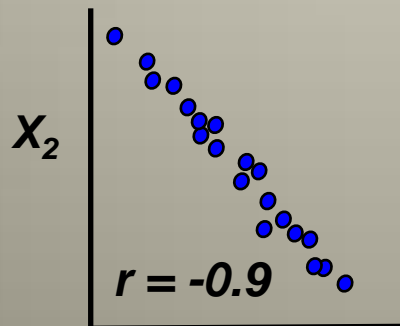
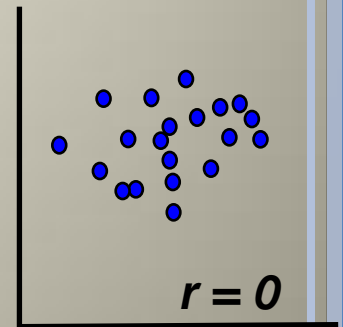
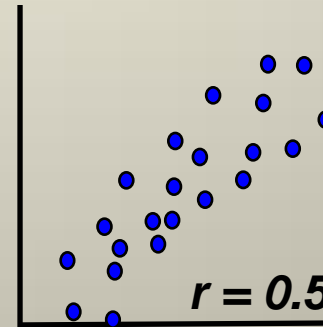
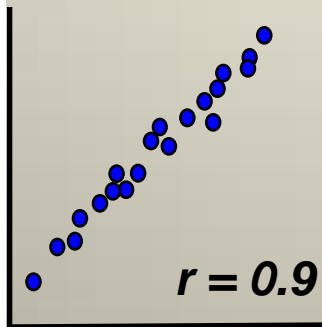


None



Weak Negative

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



X_2

X_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 :

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

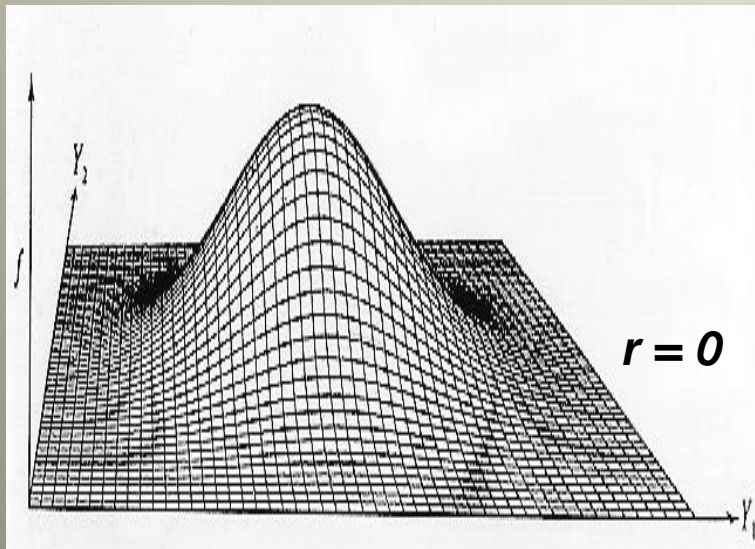
Si H_0 est rejetée (p-value < α)
corrélation



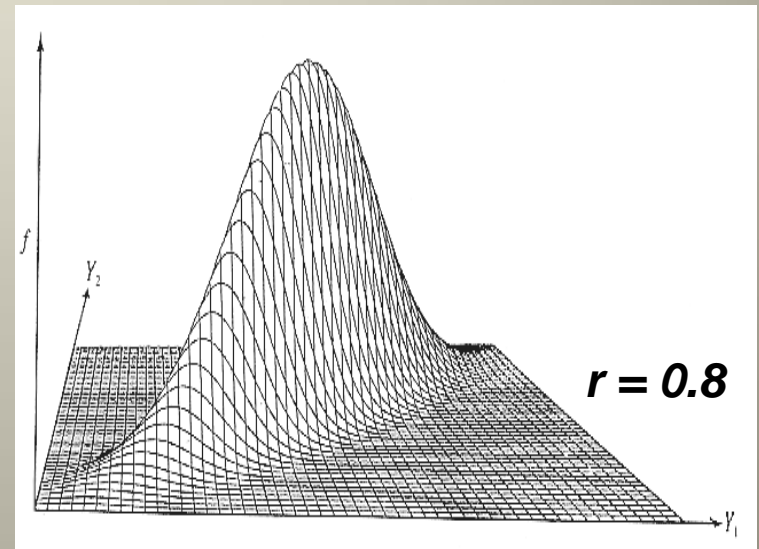
La régression linéaire

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*

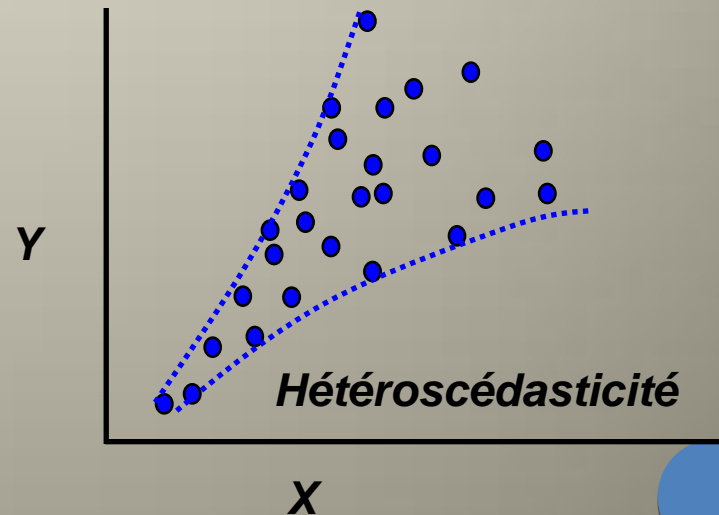
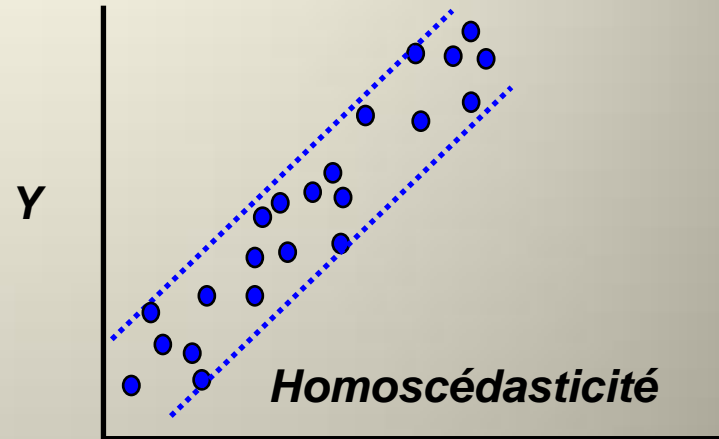


b. Homoscédasticité

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

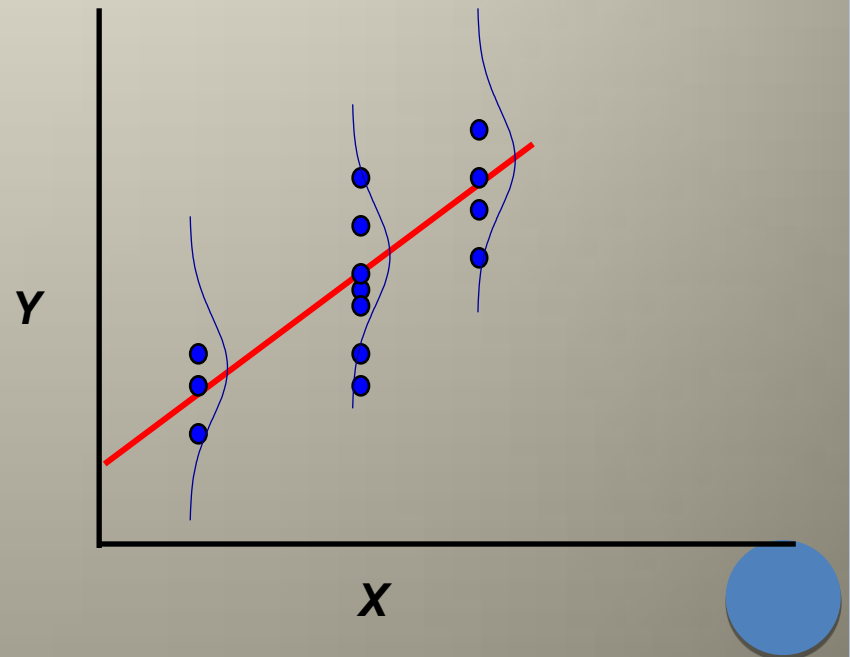


Modélisation mathématique

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

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

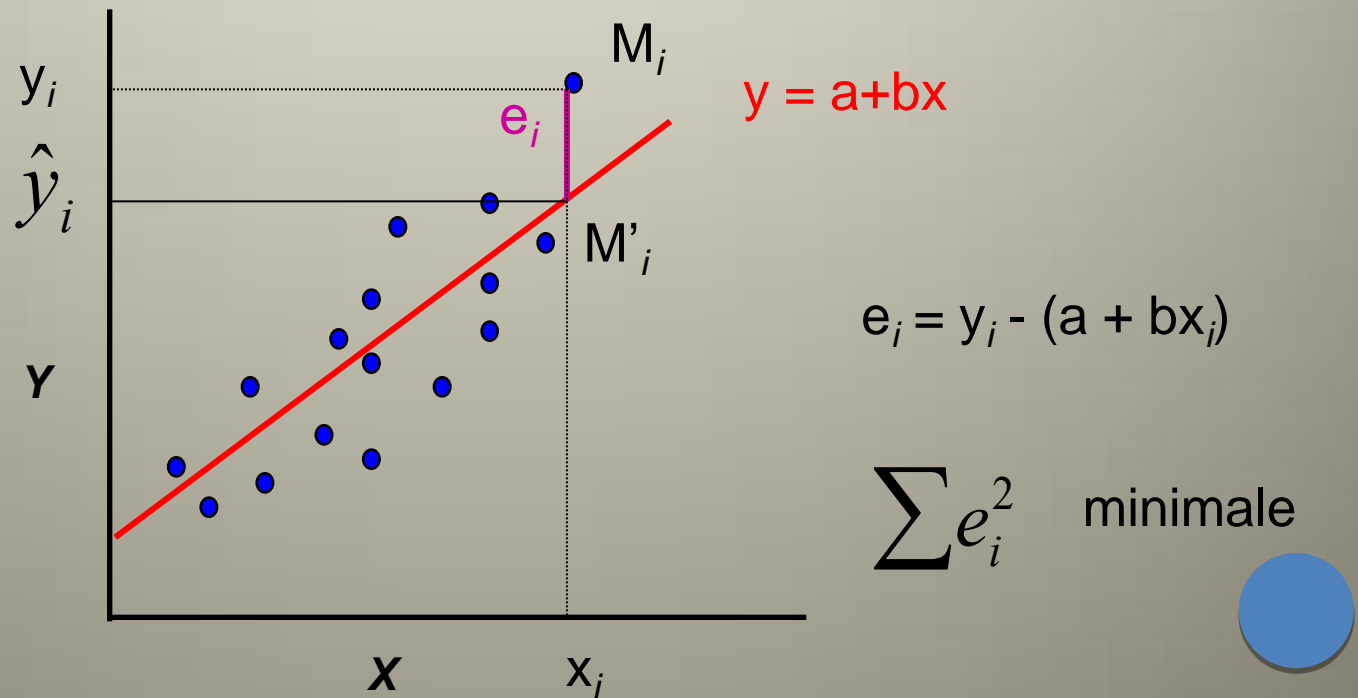
- X = variable explicative
(« indépendante »)
- Y = variable expliquée (dépendante)



L'estimation des paramètres

a? b?

Méthode d'estimation: les moindres carrés:



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 $X = x_i$, $Y_i : N(a+bx_i, \sigma)$

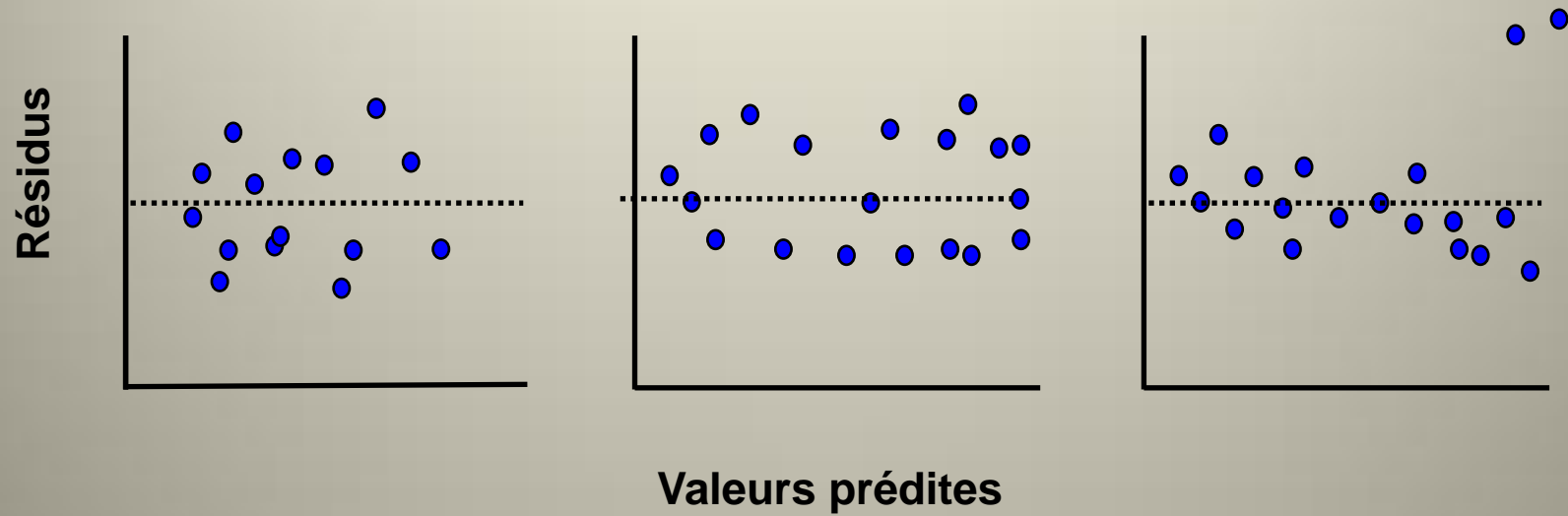
- distribution normale des erreurs
- variance identique (homoscédasticité)
- indépendance:
- linéarité de la relation

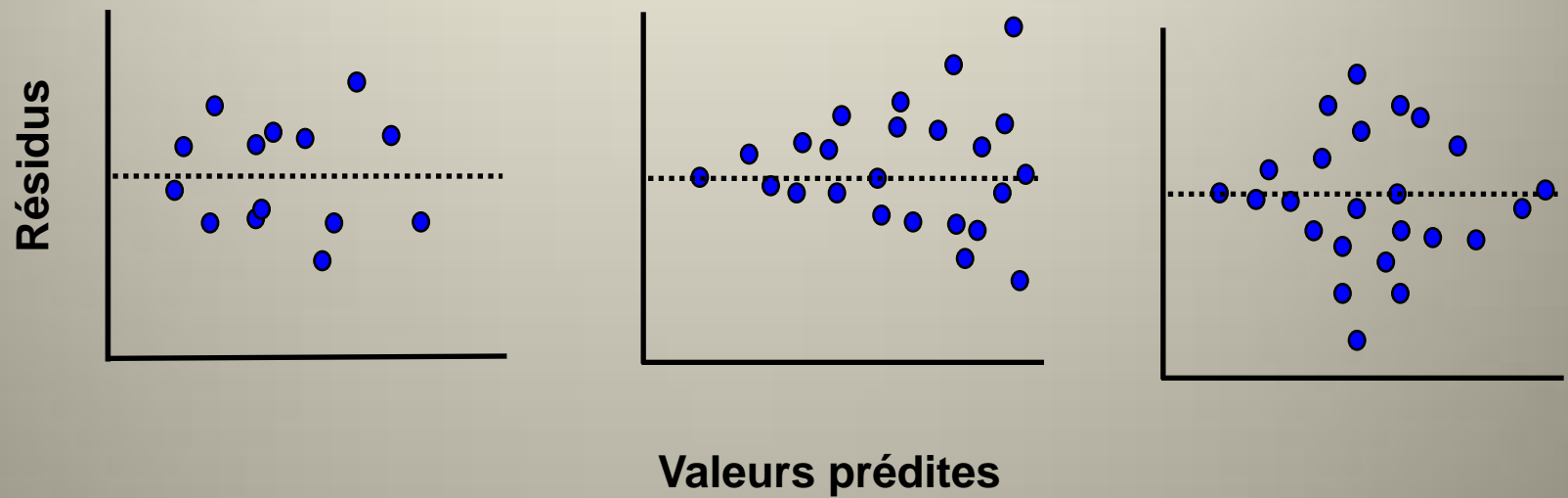
$$\text{cov}(e_i, e_j) = 0$$

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



Normalité de l'erreur





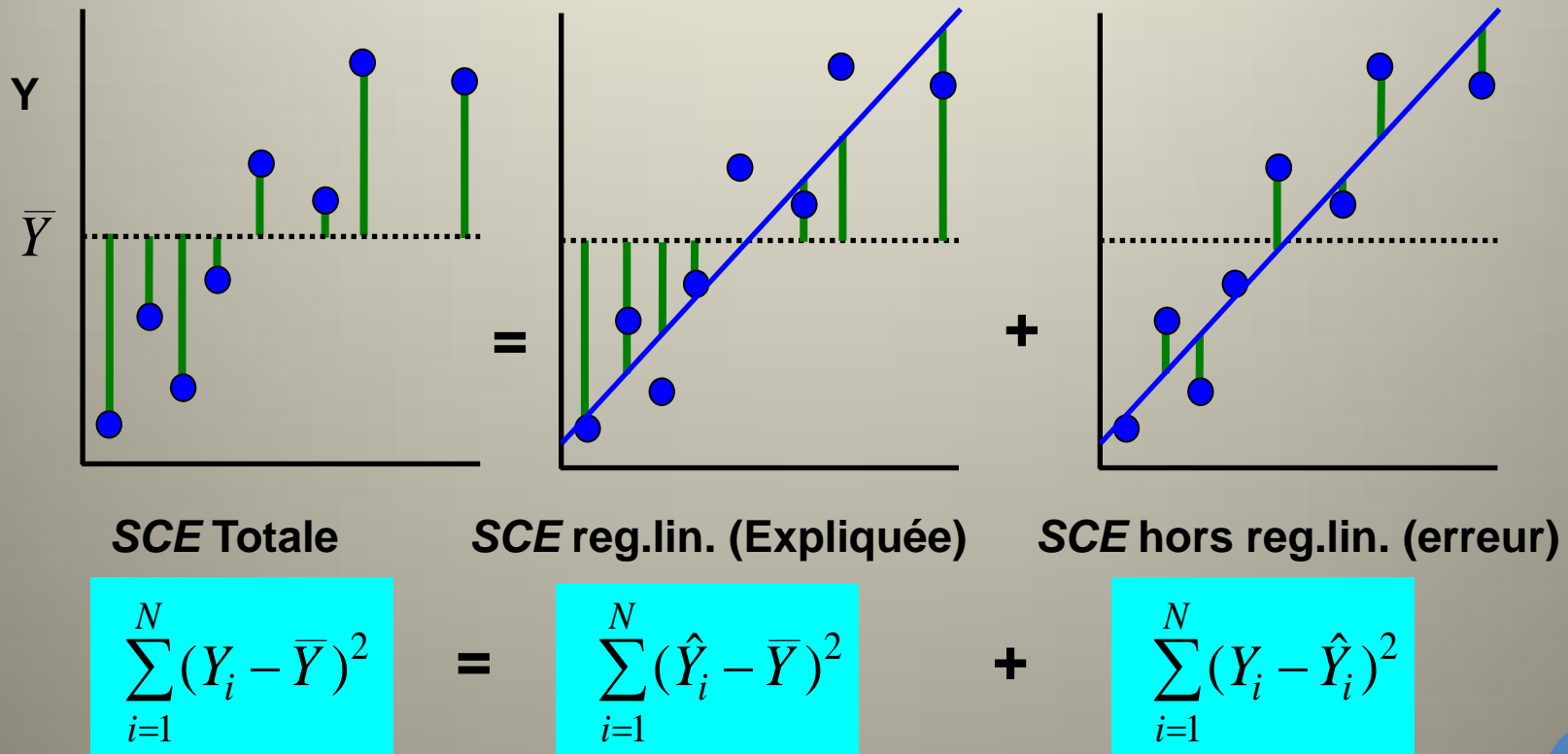
Possibilité de transformation: attention aux transformations *ad hoc*



Décomposition de la variation

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

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



Relation entre r et r^2

Coefficient de détermination

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

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

$$r^2 = \frac{SCE_{reg.lin.}}{SCE_T}$$

En particulier, $r = 0 \Leftrightarrow r^2 = 0$



Tests

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

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

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

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

numériquement
équivalent à

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



Autres tests

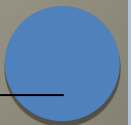
- comparaison de la pente à une valeur non nulle
- comparaison de l'ordonnée à l'origine à une valeur quelconque



Chapitre 3

Analyse en Composantes Principales A.C.P.

N. IDRISI
Faculté des Sciences et Techniques
Bordj Mouloud
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

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

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

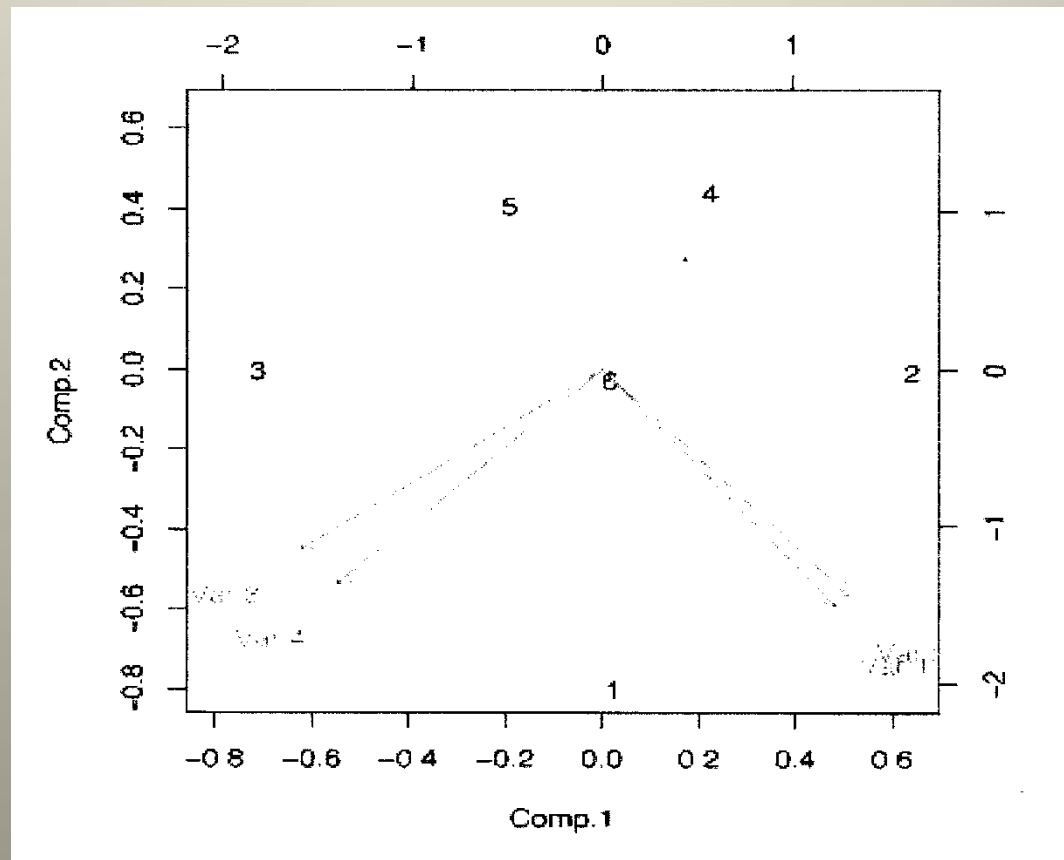
c'est-à-dire que F_j est une combinaison linéaire des variables initiales X_1, \dots, X_p .

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).



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



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 :

$$\text{CTR}(\text{sujet 1, CP1}) = \frac{0,0771^2}{0,0771^2 + \dots + 0,0556^2} = 0,64\%$$

Qualité de la représentation :
pour sujet 1 et CP2

$$\text{QLT} = \frac{2,7515^2}{0,0771^2 + \dots + 0,0166^2} = 0,998$$

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

A vos machines !