



Analyse de données

- 
- Cours 1: INTRODUCTION et DESCRIPTION DE TABLEAU DE Données

Présentée par:

Mme SAIDIS

Introduction

Analyse de données est un processus qui consiste à examiner et interpréter des données afin d'élaborer les réponses à des questions

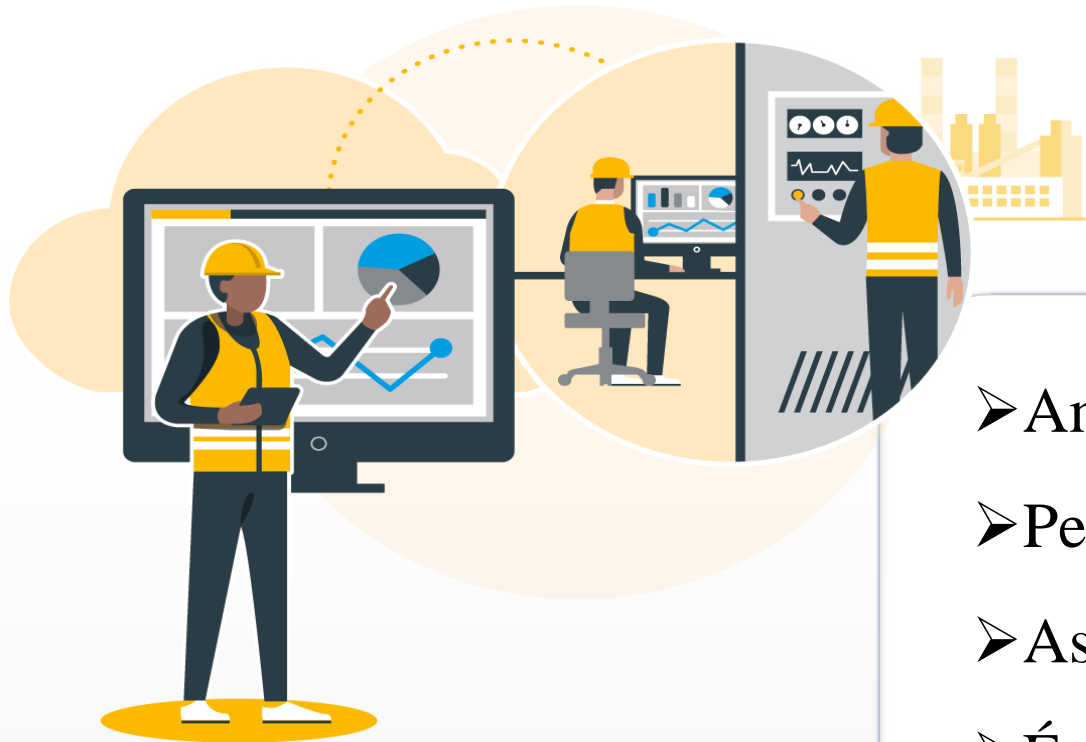
Les principales étapes de l'analyse de données:

- Cerner les sujets de l'analyse
- Déterminer la disponibilité de données
- Le choix de méthode pour répondre aux questions
- Résumer et communiquer le résultat

Dans l'industrie des grandes quantités de données sont générées de plus en plus. Ces données brute provenant de différents domaines

Analyse de données

L'impact de l'analyse des données dans l'industrie



- Améliorer la fabrication
- Personnalisez la conception des produits
- Assurer une meilleure assurance qualité
- Évaluer tout risque potentiel

Analyses de données



```
graph TD; A[Analyses de données] --> B[Les modèles statistiques]; A --> C[Les modèles classification]; A --> D[Les modèles factorielles];
```

Les modèles statistiques

Sont utilisés pour nettoyer les données au début par l'élimination des valeurs aberrantes, et aussi de visualiser les données, afin de construire l'ensemble initial d'exemples.

Les modèles classification

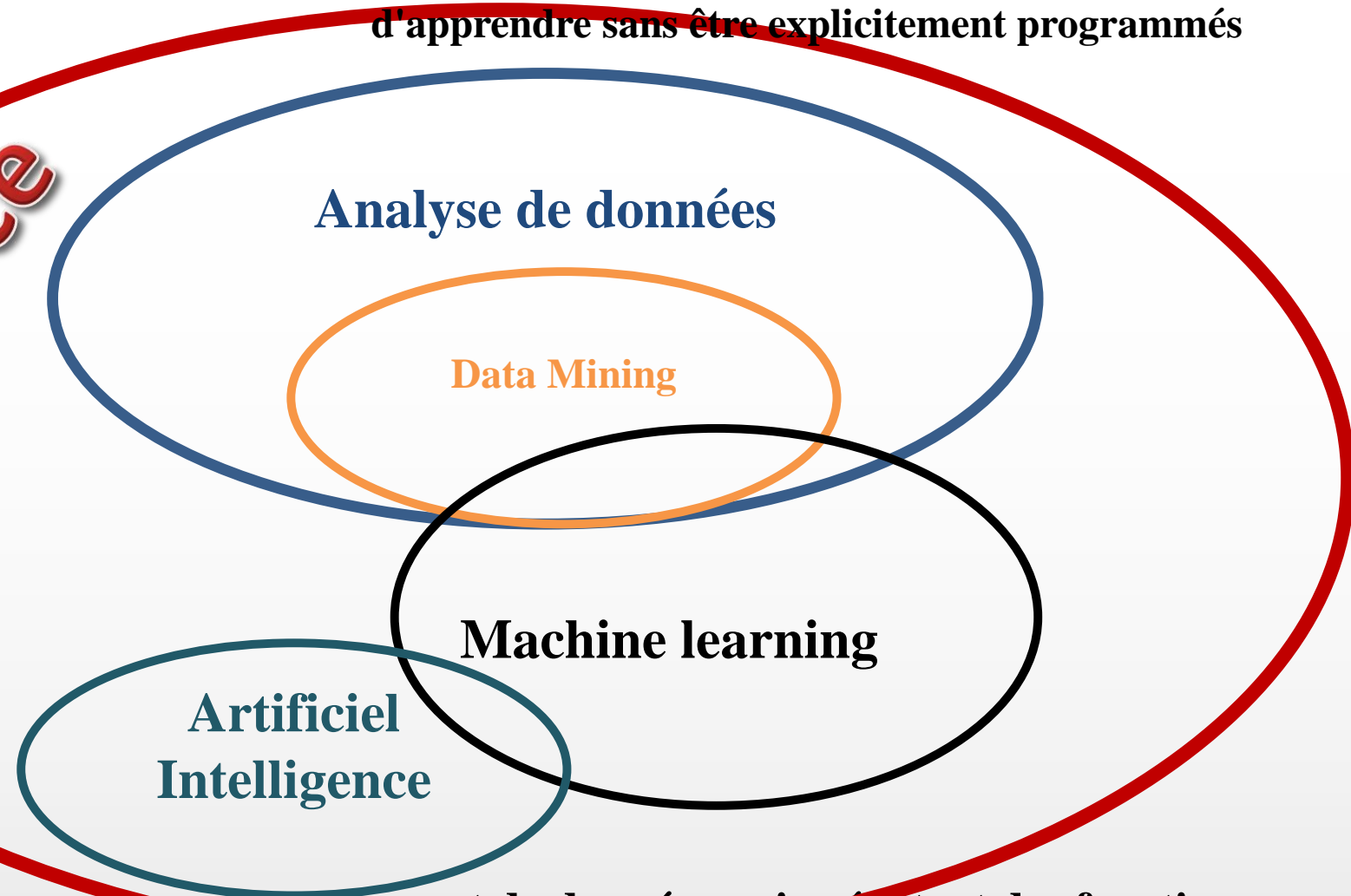
Construisent des règles et des modèles prédictifs pour synthétiser et structurer l'information contenue dans des données.

Les modèles factorielles

Cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l'algèbre linéaire.

Domaine d'études qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés

Consiste à explorer (ou fouiller) les données. Il permet d'établir des associations et relations entre les données qui sont cachées ou non évidentes, très souvent réparties sur plusieurs bases de données relationnelles

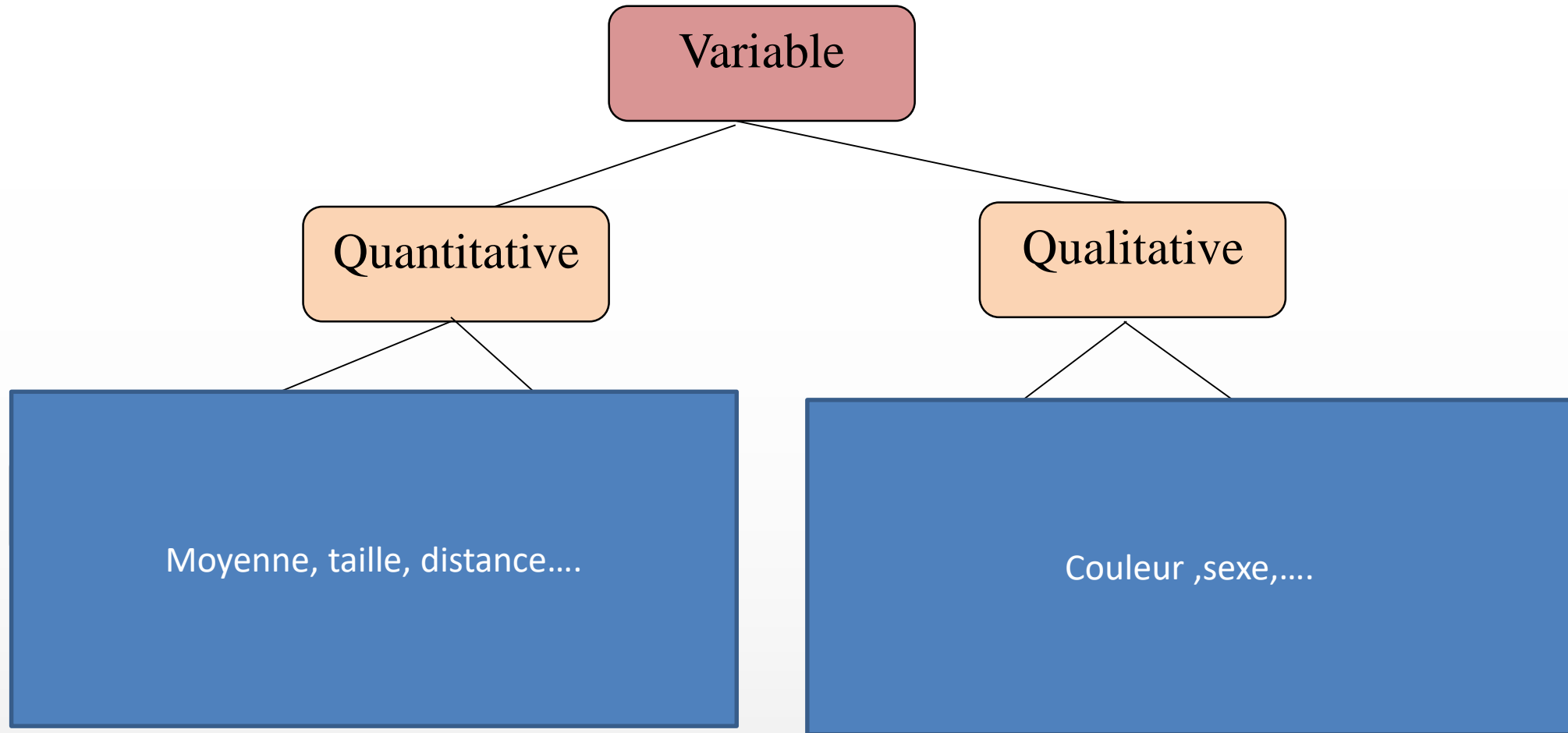


Consacrée au développement de systèmes de traitement de données qui exécutent des fonctions normalement associées à l'intelligence humaine, telles que le raisonnement, l'apprentissage et l'auto-amélioration de données

Tout un ensemble de méthodes mathématiques et informatiques

Analyse de données

Types de données



Objectif	Variables quantitative	Variables qualitative/mixtes
Repérer et visualiser les corrélations multiples entre variables et/ou les ressemblances entre individus	Analyse en composantes principales (ACP)	Analyse factorielle des correspondances (AFC) et Analyse factorielle des Correspondances Multiples , (AFCM)
Réaliser une typologie des individus	Méthodes de classification	AFC ou AFCM et classification
Caractériser de groupes d'individus à l'aide de variables	Analyse factorielle discriminante (AFD)	Analyse factorielle discriminante (AFD)

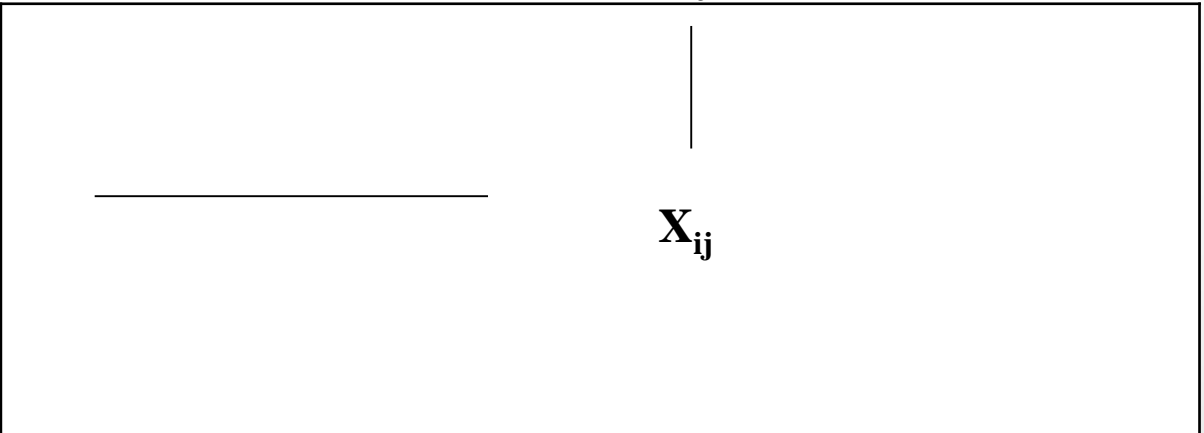
Analyse de données

Cas d'application

Les données se présentent généralement sous la forme d'un tableau

- Les lignes correspondent à des individus ou objets
- Les colonnes correspondent à des attributs ou caractéristiques.

Le tableau de variables quantitatives

		Variables		
individus		X_1	X_j	X_m
	1			
			
	i			
			
	N			

La méthode d'analyse porte le nom d'analyse en composantes principales : ACP

Exemple: Le tableau de variables quantitatives

Le tableau de notes

	Math	science	physique	Histoire
JE	6	9	12	14
LA	8	10	17	12
SA	14	6	10	13
REB	10	9	12	11
JIMS	5	9,5	6	9

Description de données

The diagram illustrates the components of a data table. A blue arrow labeled 'Individu' points to the first column, 'Nom des Fournisseurs'. Another blue arrow labeled 'Variable' points to the 'Age' column. A third blue arrow labeled 'Modalité' points to the 'Chiffre d'affaire' column.

Nom des Fournisseurs	Sexe	Age	Chiffre d'affaire
Mohamed	H	40	Modéré
Sarah	F	50	Important
Ismail	H	44	Moyen
Ilyes	H	50	Modéré
Hanane	F	35	Important
Ghouti	H	60	Moyen
Yasmina	F	55	Modéré
Fatima	F	35	Moyen

Population : Fournisseurs

Le tableau de contingence

		Modalités de Y				
Modalités de X		1	j	n		
1			n_{ij}			
.....						
i						
.....						
N						
		M				

Plus généralement on ajoute en dernière ligne et en dernière colonne les sommes par lignes et par colonnes appelés effectifs marginaux.

EXEMPLE: Le tableau de contingence

OBSEVATION

Individu	bac	sexe
1	S	HOMME
2	TM	FEMME
3	TM	HOMME
4	L	FEMME
5	S	FEMME
6	TM	Femme

TABLEAU DE CONTINGENCE

n_{ij}

MODALITE	HOMME	FEMME	
S	1	1	2
TM	1	2	3
L	0	1	1
	2	4	6

$n_{11}=1, n_{31}=0, \dots$

L'objectif de la Statistique Descriptive est de décrire les données observées pour mieux les analyser

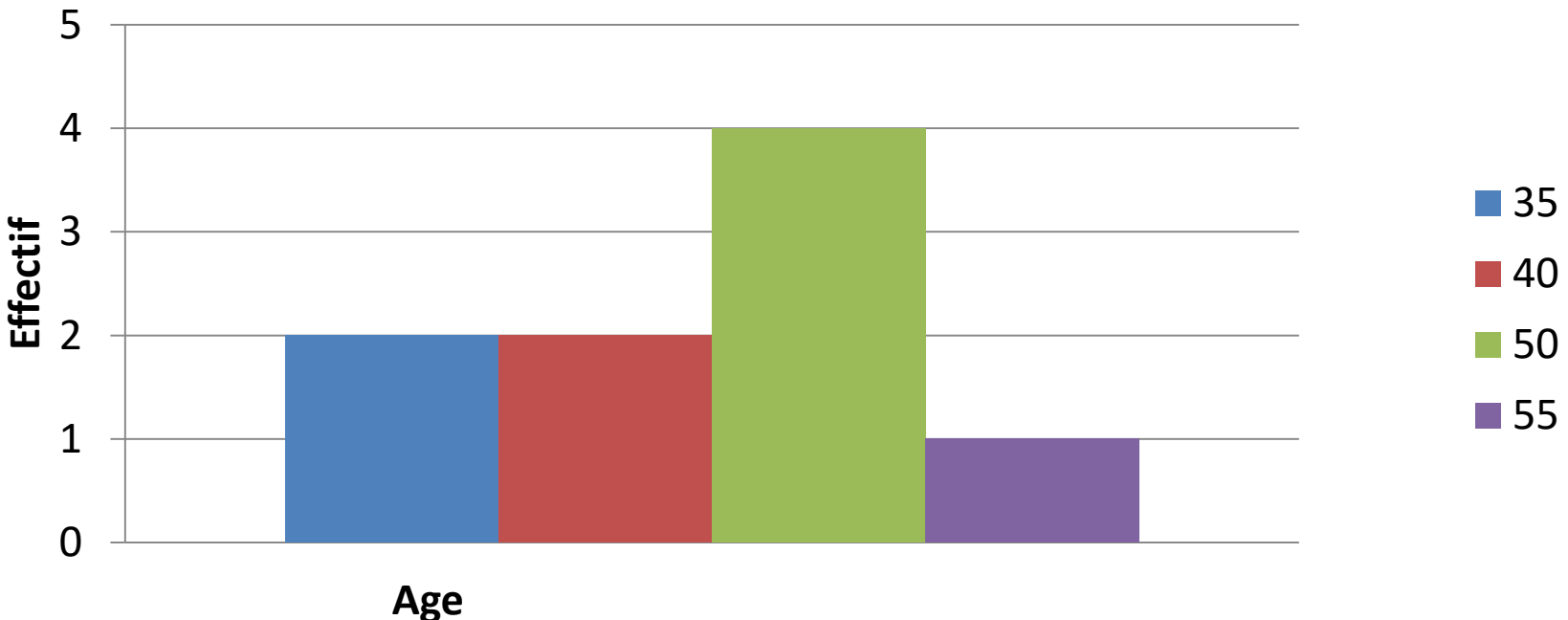
- **Description de données**
- **Valeurs centrales**
- **Indicateurs de dispersion**

Description de données

Effectifs

La variable Age

Age	35	40	50	55
Effectifs	2	2	4	1



Nom des Fournisseurs	Sexe	Age	Chiffre d'affaire
Mohamed	H	40	Modéré
Sarah	F	50	Important
Ismail	H	40	Moyen
Ilyes	H	50	Modéré
Hanane	F	35	Important
Ghouti	H	50	Moyen
Yasmina	F	55	Modéré
Fatima	F	35	Moyen
Karima	F	50	Important

Description de données

Fréquence

Fréquence de la modalité « M » d’une variable qualitative (FM)

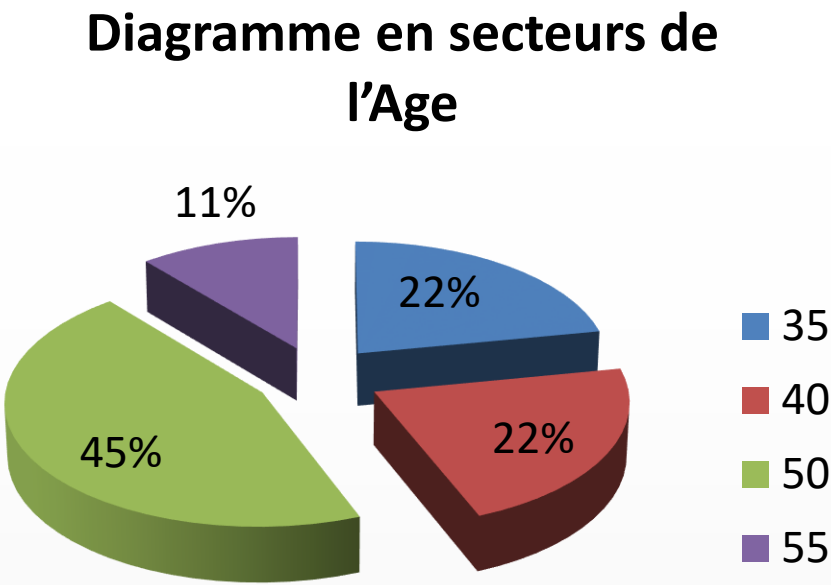
$$F_M = \frac{Effe_corresp_M}{Effe_Total}$$

Pourcentage

Pourcentage des individus correspondant à la modalité « M »

$$P_M = F_M \times 100$$

Age	Effectifs	Fréquences	Pourcentage
35	2	2/9=0.22	22%
40	2	2/9=0.22	22%
50	4	4/9=0.45	45%
55	1	1/9=0.11	11%
	Total Effectifs « 9 »	Total Fréquences « 1 »	Total Pourcentage 100



Valeurs centrales

La moyenne arithmétique

On dispose d'une population de N individus et on observe X_1, X_2, \dots, X_n les valeurs d'une variable quantitative discrète X pour ces individus.

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

Age	35	40	50	55
Effectifs	2	2	4	1

$$\bar{x} = 45$$

Indicateurs de dispersion

L'étendue

L'étendue E_x de la variable quantitative discrète X est la différence entre la plus grande et la plus petite des valeurs observées

$$E_x = \max(x_i) - \min(x_i)$$

La variance

- La variance est une mesure de la dispersion d'une série de données.
- Une variance faible indique que les nombres de la série de données sont proches l'un de l'autre.
- Une variance élevée indique que les nombres sont très distants.

Indicateurs de dispersion

La variance

Age	35	40	50	55
Effectifs	2	2	4	1

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$



$$\sigma^2 = \frac{1}{N} (\sum n_i x_i^2) - \bar{x}^2$$

Application:

μ : la moyenne

$$\sigma^2 = 1/9(2*35^2 + 2*40^2 + 4*50^2 + 1*55^2) - 45^2$$

L'écart-type

- ➔ L'écart-type est une mesure de la dispersion d'une série statistique autour de sa moyenne.
- ➔ Plus la distribution est dispersée c'est-à-dire les valeurs ne sont pas concentrées autour de la moyenne, plus l'écart-type sera élevé.

$$Ecart_type(x) = \sqrt{\text{var}(x)}$$