

# Statistique Descriptive

- › Statistique descriptive à une variable
- › Statistique descriptive à deux variables

# Statistique descriptive à une variable

# Introduction

La statique est un ensemble de méthode et techniques mathématiques basées sur la collecte de données numériques, l'organisation, la présentation l'analyse et la modélisation de ces données, dans le but de comparaison, de prévision, de constat...



# Introduction

Les plus gros consommateurs de statistiques sont les assureurs (risques d'accident, de maladies des assurés), les médecins (épidémiologie), les démographes (populations et leur dynamique), les économistes (emploi, conjoncture économique), les météorologue...

# Introduction

Remarque : Il ne faut pas confondre entre la statistique, une statistique et les statistique car

- **Une statistique**: (ou une série statistique) peut être fabriquée (à partir des matériaux de base que constituent les statistiques) et présentée sous la forme d'un tableau statistique.
- **Les statistiques**: sont les résultats obtenus lors d'une étude statistique

# Terminologie de base

## 1. Population

A la base de toute étude statistique, il y a une population formée d'éléments de même nature qu'on appelle individus (ou unités statistique). Les individus peuvent être humains, des êtres vivants, des objets etc...

La population est notée  $\Omega$ , le nombre de ses éléments est noté  $N$

( $\text{card}(\Omega) = N$ ) le nombre  $N$  appelé effectif total de la population.

# Terminologie de base

Une partie de la population( càd un sous-ensemble de  $\Omega$ ) est appelée échantillon.

## Exemple1:

les 40000 étudiants de l'université Mohammed 1<sup>er</sup>

- La population  $\Omega = \{\text{des étudiants de l'université Mohammed 1}^{\text{er}}\}$ ,  $\text{card } (\Omega) = 40000$
- Un individu (ou unité statistique) est un élément de  $\Omega$  c'est donc un étudiant de l'université Mohammed 1<sup>er</sup>
- Les 5000 étudiants de la faculté pluridisciplinaire de Nador est un échantillon de La population  $\Omega$

# Terminologie de base

## 2. Caractère

**Définition:** On appelle caractère, toute propriété (ou spécificité) étudiée sur les individus.

Un caractère s'appelle aussi variable statistique.

**Exemple 2:** Un abonné de l'ONE peut être étudié selon son âge , sa nationalité, son sexe, son salaire etc...

L'âge, le sexe, le salaire et la nationalité sont des **caractères**.

**Exemple 3:** Un étudiant de la faculté pluridisciplinaire peut être étudié selon son groupe sanguin, sa taille, son poids etc ... Le groupe sanguin, la taille, le poids Sont des caractères.

# Terminologie de base

## 3. Modalité d'un caractère.

On appelle modalité d'un caractère, une situation dans laquelle se trouve un individu d'un caractère étudié. Les modalités sont donc les différentes spécificités du caractère.

### Exemple 4:

Soit le caractère " état matrimonial ". Les modalités sont: Marié, divorcé, célibataire, veuf.

### Exemple 5:

Soit le caractère " groupe sanguin ". Les modalités sont: A, B, AB, O.

# Terminologie de base

## Exemple 6:

Soit le caractère

taille d'un individu  $\begin{cases} \textit{grand} \\ \textit{moyen} \\ \textit{petit} \end{cases}$

Caractère: taille

Modalités: grand, moyen, petit

## Exemple 7:

Soit le caractère "note d'examen de mathématique" des étudiants de SMP. Les modalités sont 0, 1, 2, 3..., 20.

# Terminologie de base

## Remarque:

Un caractère présente plusieurs modalités, par contre, un individu n'admet qu'une et une seule modalité.

On distingue deux types de caractères: caractères qualitatif et caractères quantitatif.



# Terminologie de base

## a. caractères qualitatif

**Définition:** Un caractère est dit qualitatif lorsque ses différentes modalités ne sont pas mesurables ( c'àd ne sont pas des nombres).

### Exemple 8

Profession ( médecin, enseignant, avocat etc...); Sexe (masculin, féminin), groupe sanguin (A, B, AB, O), Nationalité (marocaine, algérienne, tunisienne, etc...).

# Terminologie de base

## b. caractères quantitatif

**Définition:** Un caractère est dit quantitatif lorsque ses différentes modalités sont numériques (càd mesurables) dans ce cas est appelé variable quantitative.

### Exemple 9:

L'âge, le poids, la température, la vitesse, le salaire sont des variables quantitatives.

Lorsque le caractère (la variable) est quantitatif(ve), on distingue deux cas:

**Les variables discrètes et les variables continues**

# Terminologie de base

## i. Variable statistique discrète

**Définition:** Une Variable statistique est dite variable discrète si l'ensemble de ses valeurs est fini ou dénombrable.

### Exemple 10:

Les résultats d'examen, le nombre d'abstentions à un vote, le nombre d'enfants par ménage sont des variables discrètes.

# Terminologie de base

## ii. Variable statistique continue

**Définition:** Une **Variable statistique** est dite variable continue si l'ensemble de ses valeurs est infini (càd la variable peut prendre toutes les valeurs d'un intervalle).

### Exemple 11:

Les longueurs, le poids, le pourcentage d'abstentions à un vote sont des variables continues.

### Remarque:

Si la variable est discrète mais prend beaucoup de valeurs, on la traite comme une variable continue

# Terminologie de base

## 6. Série ou distribution statistique

**Définition:** On appelle Série statistique (ou distribution statistique), une liste  $N$  d'observations faites pour un caractère de la population  $\Omega$ .

- Une Série statistique quantitative est donc une liste de valeurs de variable.

### Exemple 12:

Voici Une Série statistique quantitative indiquant le nombre d'appels téléphoniques réalisés au moyen d'un GSM au cours d'une journée pour un échantillon de 15 personnes

0, 1, 0, 0, 1, 2, 1, 3, 1, 0, 2, 2, 3, 2, 1.

# Terminologie de base

- Une Série statistique qualitative est une liste de variétés du caractère ( c à d modalités).

## Exemple 20

Voici Une Série statistique qualitative indiquant le groupe sanguin de 15 enseignants.

O, A, A, AB, AB, O, O, O, AB, B, B, B, A, AB, B.

# Terminologie de base

## 7. Les tableaux statistiques

L'un des objectifs de la statistique descriptive est de résumer les données brutes recueillies sur une population dans les tableaux statistiques.

Avantages:

- Présentation des données de façon lisible.
- En ligne: informations relatives à chaque individu.
- En colonne: critère ou caractères étudiés.

**Exemple:**

Enquête au près d'un échantillon de 56 familles marocaines sur le nombre d'enfants par ménage.

5 - 4 - 0 - 2 - 1 - 5 - 3 - 4 - 2 - 0 - 4 - 5 - 7 - 6 - 2 - 8 - 1 - 4 - 6

3 - 5 - 7 - 4 - 1 - 2 - 4 - 6 - 3 - 5 - 2 - 1 - 0 - 4 - 6 - 5 - 4 - 2 - 1

3 - 6 - 4 - 2 - 5 - 3 - 4 - 5 - 4 - 3 - 9 - 2 - 4 - 6 - 5 - 4 - 3 - 4

# Terminologie de base

Les données brutes ne sont pas lisibles c'est pourquoi on va grouper les données dans un tableau pour faciliter le traitement et les interprétations.

Nombre d'enfants par famille	Nombre de famille
0	3
1	5
2	8
3	7
4	14
5	9
6	6
7	2
8	1
9	1
Total	56



# Terminologie de base

La présentation d'un tableau statistique doit respecter des principes généraux.

- Le tableau doit porter un titre précisant son contenu: le phénomène étudié, la façon dont il est étudié, le lieu, la date, etc...
- Le tableau doit porter des intitulés de lignes et de colonnes clairement définis.
- Le tableau doit préciser les unités utilisés: ne pas confondre le mètre avec le mètre carré, le millier avec le million, le DH avec l'Euro.
- Le tableau doit préciser la source des informations lorsque les données sont empruntées à une publication ou à un organisme.

# Distributions d'effectifs et de fréquences

Soit  $\Omega$  une population de taille  $N$  et soit  $X$  une variable quantitative discrète ( ou qualitative) définie sur  $\Omega$  dont les valeurs possibles, rangées dans l'ordre croissant, sont  $x_1, x_2, \dots, x_n$ .

Notation:  $X(\Omega) = \{x_1, x_2, \dots, x_n\}$

## 1. Effectifs et fréquences

**Définition:** L'**effectif** d'une modalité est le nombre  $n_i$  d'individus présentant cette modalité. On l'appelle aussi **effectif partiel**

# Distributions d'effectifs et de fréquences

## Exemple22:

Voici une série statistique quantitative indiquant le nombre d'appels téléphoniques réalisés au moyen d'un GSM au cours d'une journée pour un échantillon de 100 personnes

Nombre d'appel	0	2	3	5	7	8	9
Effectif	15	20	25	15	10	8	7

L'effectif de la valeur 0 est 15, l'effectif de la valeur 7 est 10, l'effectif de la valeur 3 est 25 etc...

# Distributions d'effectifs et de fréquences

## Exemple22:

Voici une série statistique qualitative indiquant le groupe sanguin de 15 étudiants de la section SMP

Groupes sanguins	A	B	AB	O
Effectifs	2	3	6	4

Par exemple, l'effectif de la modalité "A" est 2, l'effectif de la modalité "O" est 4 etc...

# Distributions d'effectifs et de fréquences

**Remarque:**  $\sum_{i=1}^p n_i = n_1 + n_2 + \dots + n_p = N$ .

$N$  s'appelle **l'effectif total**.

**Définition:** La fréquence d'une valeur  $x_i$ , notée  $f_i$  est le rapport

$$f_i = \frac{n_i}{N}$$

La fréquence  $f_i$  peut être exprimé en pourcentage

$$f_i = 100 \times \frac{n_i}{N}$$

Dans ce cas on dit que  $f_i$  représente le pourcentage de  $x_i$

# Distributions d'effectifs et de fréquences

## Exemple 23:

Voici une série statistique qualitative indiquant le groupe sanguin de 500 étudiants de la faculté pluridisciplinaire de Nador

Groupe sanguin	A	B	AB	O
Effectifs	150	70	180	100
Fréquences	0,3	0,14	0,36	0,2

- Population : Les 500 étudiants de la faculté pluridisciplinaire.
- Caractère étudié : Groupe sanguin.
- Modalités : A, B, AB et O.

# Distributions d'effectifs et de fréquences

$$\left\{ \begin{array}{l} \text{La fréquence de la modalité } A \text{ est } f_1 = \frac{150}{500} \\ \text{La fréquence de la modalité } B \text{ est } f_2 = \frac{70}{500} \\ \text{La fréquence de la modalité } AB \text{ est } f_3 = \frac{180}{500} \\ \text{La fréquence de la modalité } O \text{ est } f_4 = \frac{100}{500} \end{array} \right.$$

On vérifie que:  $f_1 + f_2 + f_3 + f_4 = 1$

# Distributions d'effectifs et de fréquences

## Exemple 24:

Voici une série statistique qualitative indiquant l'état matrimonial de 1000 salariés de l' IAM:

- Population : Les 1000 salariés de l' IAM.
- Caractère étudié : l'état matrimonial .
- Modalités : célibataire, marié, veuf et divorcé.

Etat matrimonial	célibataire	Marié	veuf	divorcé
Effectifs	125	850	15	10
Fréquences	0,125	0,85	0,015	0,01



# Distributions d'effectifs et de fréquences

Remarque:

$$0 \leq f_i \leq 1 \quad \sum_{i=1}^p f_i = f_1 + f_2 + \cdots f_p = 1$$

Exemple 25:

Reprenons l'exemple 22 c'est-à-dire le nombre d'appels téléphoniques réalisés au moyen d'un GSM au cours d'une journée pour un échantillon de 100 personnes.

Nous obtenons le tableau suivant:

Nombre d'appels	0	2	3	5	7	8	9
Effectifs	15	20	25	15	10	8	7
Fréquences	0,15	0,2	0,25	0,15	0,1	0,08	0,07

# Distributions d'effectifs et de fréquences

## Définition:

- L'effectif **cumulé croissant** d'une valeur  $x_i$  est la somme des effectifs de cette valeur et des valeurs inférieures. Autrement dit:

L'effectif cumulé croissant de  $x_i$  est:

$$n_1 + n_2 + \dots + n_i$$

La fréquence **cumulée croissante**, notée  $F_i$  d'une valeur  $x_i$ , est la somme des fréquences de cette valeur et des valeurs inférieures. Autrement dit:

La fréquence **cumulée croissante** de  $x_i$  est :

$$F_i = f_1 + f_2 + \dots + f_i$$

Il est commode de présenter une série statistique sous forme d'un tableau contenant les valeurs possibles de cette variable, rangées dans l'ordre croissant, et pour chacune de ces valeurs, l'effectif (ou fréquence) correspondant(e) .

$\pi$ 

Valeur de la variable	Effectif:	Fréquence	Effectif cumulés croissant	Fréquences cumulées croissant
$x_1$	$n_1$	$f_1$	$n_1$	$f_1$
$x_2$	$n_2$	$f_2$	$n_1+n_2$	$F_1 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_i$	$f_i$	$n_1+n_2+\dots+n_i$	$F_i = f_1 + f_2 + \dots + f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_p$	$n_p$	$f_p$	$n_1+n_2+\dots+n_p = N$	$F_p = f_1 + f_2 + \dots + f_p = 1$

# Distributions d'effectifs et de fréquences

## Exemple 26:

Le nombre d'enfants dans 1000 ménages ayant au moins un enfant est indiqué dans le tableau ci-dessous

Nombre d'enfants	Effectifs: $n_i$	Fréquence : $f_i$	Fréquences cumulées: $F_i$
1	282	0,282	0,282
2	273	0,273	0,555
3	248	0,248	0,803
4	120	0,120	0,923
5	35	0,035	0,958
6	42	0,042	1

# Représentations graphiques

Il est souvent préférable de représenter graphiquement une série statistique.

Un graphique permet de visualiser le comportement de la variable étudiée. Les représentations graphiques facilitent également les comparaisons des séries statistiques.

## 1. Diagrammes en bâtons

Ce diagramme consiste à porter en abscisse les valeurs observées et à tracer en regard de chacune d'elles un segment vertical de longueur égale à son effectif ou à sa fréquence. Ce diagramme est utilisé pour les variables discrètes.

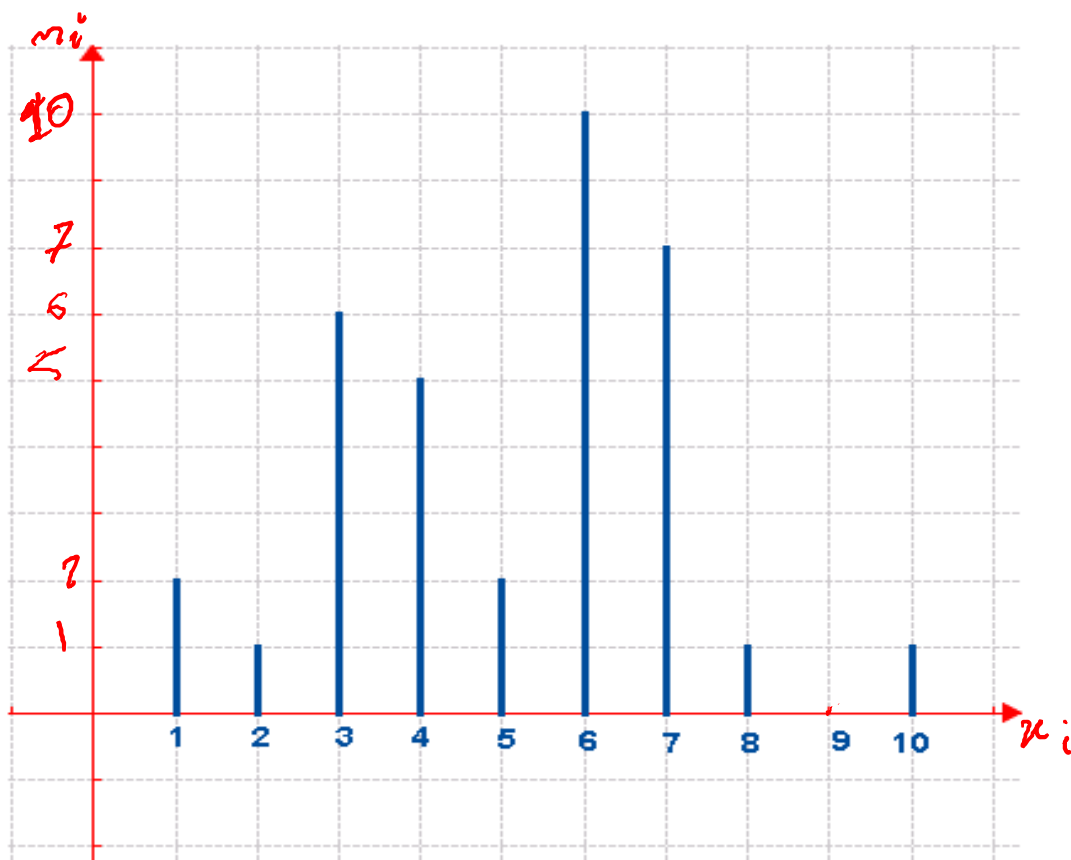
# Représentations graphiques

**Exemple 27:** Voici le diagramme en bâtons représentant la série :

Notes obtenues à un contrôle dans une classe de 34 étudiants

Note	1	2	3	4	5	6	7	8	9	10
Effectif	2	1	6	5	2	9	7	1	0	1

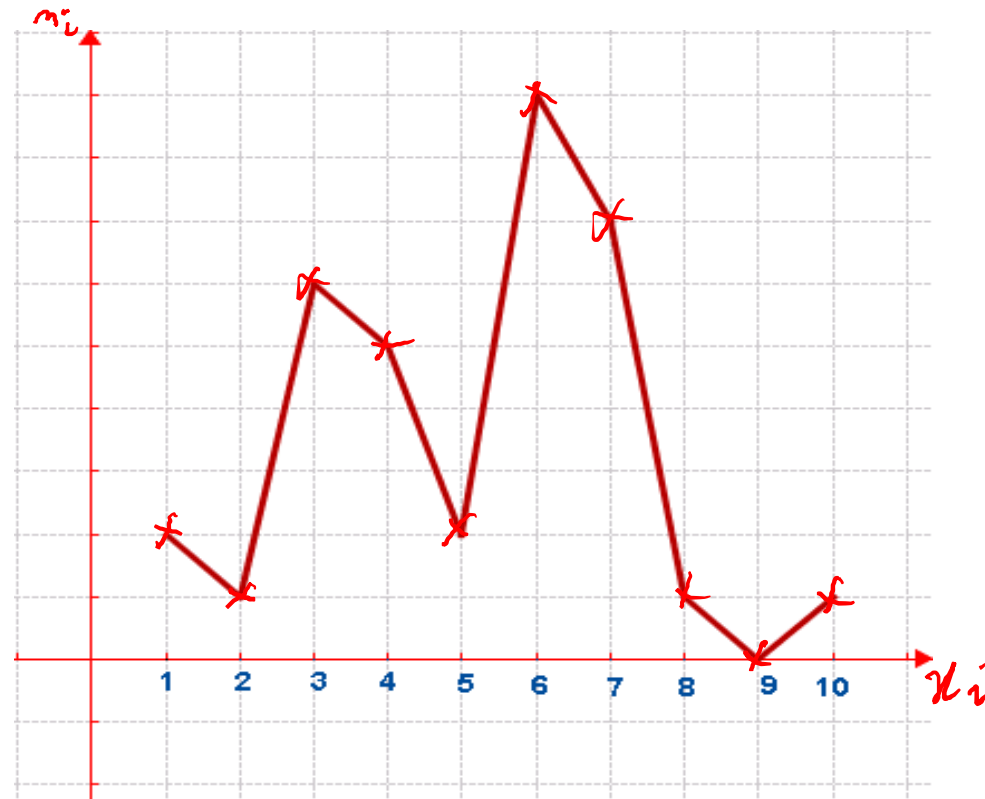
# Représentations graphiques



# Représentations graphiques

## 2. Polygones des effectifs:

Ce polygone s'obtient en joignant par un segment de droite les extrémités des segments voisins des diagrammes en bâtons pour les effectifs. C'est une ligne polygonale joignant les points  $(x_1, n_1), (x_2, n_2), \dots, (x_p, n_p)$ . Dans l'exemple précédent le polygone est :





# Représentations graphiques

## 3. Polygones des fréquences:

Ce polygone s'obtient en joignant par un segment de droite les extrémités des segments voisins des diagrammes en bâtons pour les fréquences. C'est une ligne polygonale joignant les points  $(x_1, f_1)$ ,  $(x_2, f_2)$ , ...,  $(x_p, f_p)$ .

## 4. Polygones des fréquences cumulées:

Polygones des fréquences cumulées est le graphique de la fonction  $F(x)$  définie de la manière suivante: pour tout  $x \in \mathbb{R}$

# Représentations graphiques

$$\left\{ \begin{array}{lll} 0 & \text{si} & x < x_1 \\ F_1 = f_1 & \text{si} & x_1 \leq x < x_2 \\ F_2 = f_1 + f_2 & \text{si} & x_2 \leq x < x_3 \\ \vdots & & \\ F_{i-1} = f_1 + f_2 + \cdots + f_{i-1} & \text{si} & x_{i-1} \leq x < x_i \quad i = 2, \dots, p \\ F_p = f_1 + f_2 + \cdots + f_p = 1 & \text{si} & x \geq x_p \end{array} \right.$$

Cette fonction s'appelle **fonction de répartition** de la variable  $X$

# Représentations graphiques

**Exemple28:** L'assistant social d'un centre public d'aide social s'est intéressé au nombre des personnes à charge parmi les ayant-droit demandeurs d'assistance. Il a sélectionné au hasard 100 dossiers et a relevé, dans chacun d'eux, le nombre de personnes à charge déclarées.

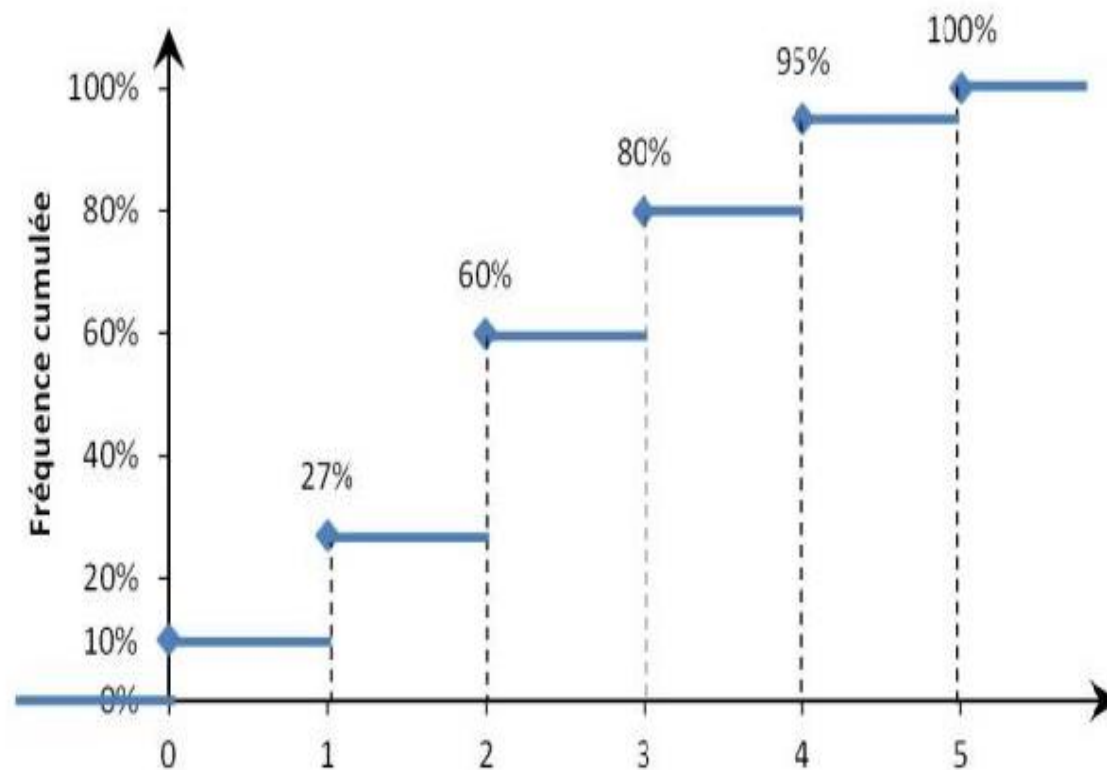
1	2	3	0	1	2	2	2	3	4	1	3	1	2	3	4	1	3	4	1
3	3	5	3	4	0	3	4	3	0	0	4	2	5	2	4	2	2	2	3
2	2	4	2	3	2	1	2	3	1	4	1	2	3	5	0	2	5	2	3
4	2	2	5	1	3	2	4	1	2	2	3	0	1	4	2	1	4	0	2
2	2	4	2	3	0	2	1	0	0	4	2	2	3	1	1	2	3	1	2

# Représentations graphiques

Nombre de personne à chargé	Nombre de dossier Effectif	Proportion de dossier Fréquence	Fréquence cumulé
0	10	0,10	0,10
1	17	0,17	0,27
2	33	0,33	0,60
3	20	0,20	0,80
4	15	0,15	0,95
5	5	0,05	1

# Représentations graphiques

La fonction de répartition ( La courbe cumulative des fréquences associée à cette série est présentée ci-dessous



# Représentations graphiques

## Propriétés de la fonction de répartition

- $F$  est une fonction définie sur  $\mathbb{R}$  à valeurs dans  $[0,1]$
- $F$  est une fonction en escalier.
- $F$  est une fonction croissante.
- $F$  est une fonction continue à droite en tout point.

# Représentations graphiques

## 5. Diagrammes en secteurs:

### Distribution groupée de fréquences

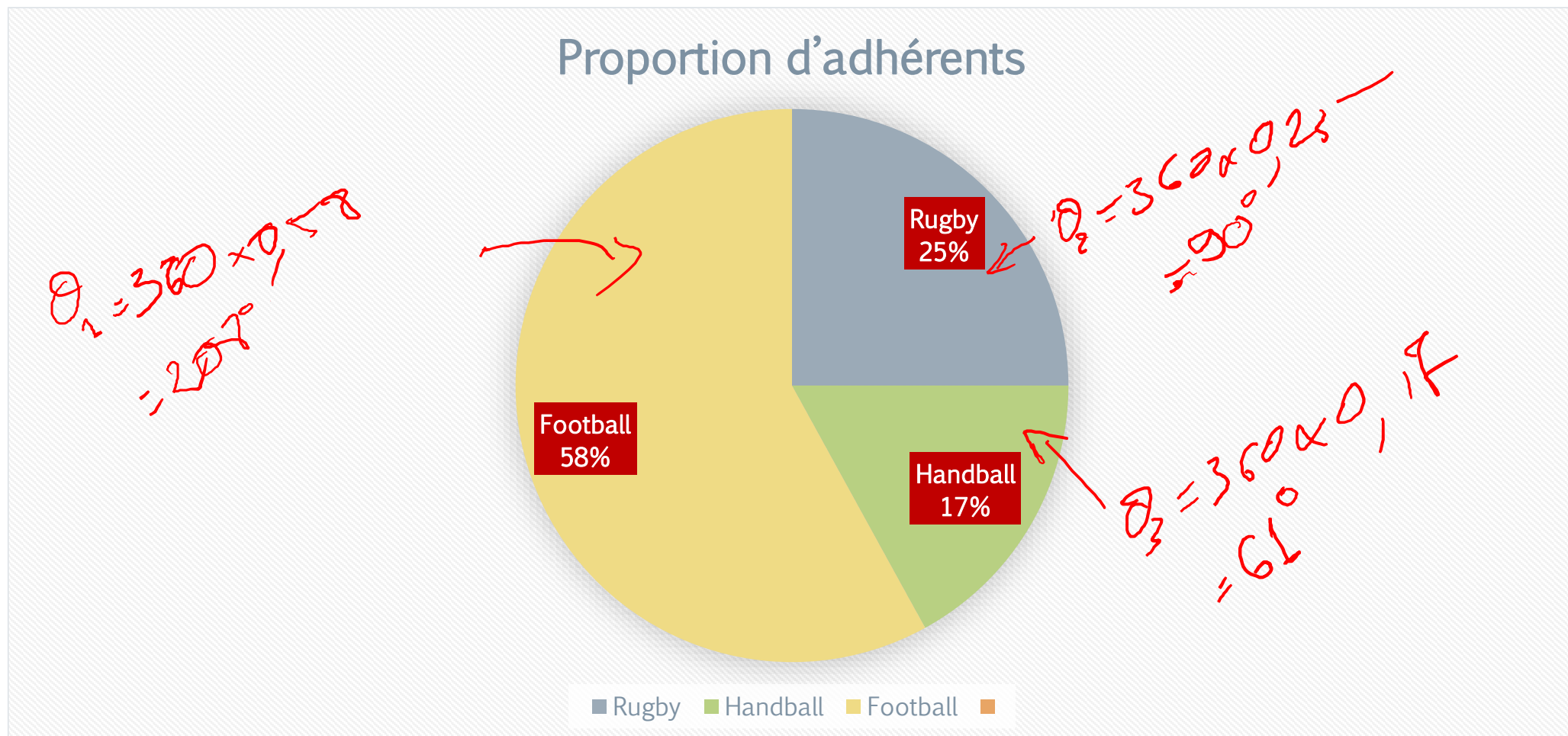
Chaque secteur représente une modalité et l'angle du secteur est proportionnel à l'effectif (ou fréquence) de la modalité.

Chaque angle  $\theta_i = 360 \times f_i$

**Exemple:29** Proportion d'adhérents à un club sportif dans différentes sections:

1. 17% jouent au handball,
2. 25% jouent au rugby,
3. 58% jouent Football.

# Représentations graphiques





# Grouperment de données en classe

Pour une série statistique qui présente un grand nombre de valeurs distinctes on a intérêt à grouper les données. Cela signifie qu'au lieu d'énumérer les valeurs de la variable, on partitionne le domaine de celle-ci en intervalles appelés classes.

**Exemple30:** On désire étudier la taille des étudiants de la faculté pluridisciplinaire. Pour cela, on range les tailles en classes:

[155, 160[, [160, 165[, [165, 170[, [170, 175[, [175, 180[, [180, 185[, [185, 190[.

**Définition:** Pour une classe  $c_i = [a_i, a_{i+1}[$ :

- $a_i$  et  $a_{i+1}$  s'appellent les bornes ou les limites de la classe  $c_i$
- $m_i = \frac{a_i + a_{i+1}}{2}$  s'appelle **le centre** (ou milieu) de la classe  $c_i$ .
- $e_i = a_{i+1} - a_i$  s'appelle **l'amplitude** ou **l'étendue** de la classe  $c_i$ .

$e = x_{\max} - x_{\min}$  est l'étendue de la série

# Grouperment de données en classe

- L'effectif  $n_i$  de la classe  $c_i$  est le nombre d'individus pour lesquels la variable statistique prend une valeur de l'intervalle  $c_i$ .
- La fréquence  $f_i$  de la classe  $c_i = [a_i, a_{i+1}[$  est le rapport  $\frac{n_i}{N}$
- La fréquence cumulée  $F_i$  de la classe  $c_i$  est la somme des fréquences de cette classe et des classes précédentes càd,

$$F_i = f_1 + f_2 + \dots + f_i$$

- La densité ~~(ou fréquence unitaire)~~ de la classe  $c_i$  est le rapport  $d_i = \frac{f_i}{a_{i+1} - a_i}$
- La distribution des fréquences d'une variable est un tableau contenant les classes de cette variable et pour chacune des classes, la fréquence correspondante

# Groupement de données en classe

**Remarque:** Pour avoir une étude un peu homogène de la série, le nombre de classes ne doit être ni trop grand ni trop petit et souvent, les classes ont même amplitude,  $a_{i+1} - a_i = a = \text{constante}$ .

**Exemple31:** Le revenu mensuel d'un groupe de 200 ingénieurs se répartissaient comme suit en l'an 2009

# Groupeement de données en classe

Revenus en DH	Eff : $n_i$	Fr : $f_i$	Fréq. cum. $F_i$	milieu : $m_i$	densité: : $d_i$
[6000,8000[	30	0,15	0,15	7000	$\frac{0,15}{2000}$
[8000 , 10000[	37	0,185	0,335	9000	$\frac{0,185}{2000}$
[10000 , 12000[	40	0,2	0,535	11000	$\frac{0,2}{2000}$
[12000 , 14000[	16	0,08	0,615	13000	$\frac{0,08}{2000}$
[14000 , 16000[	12	0,06	0,675	15000	$\frac{0,06}{2000}$
[16000 , 18000[	10	0,05	0,725	17000	$\frac{0,05}{2000}$
[18000 , 20000[	20	0,1	0,825	19000	$\frac{0,1}{2000}$
[20000 , 22000[	18	0,09	0,915	21000	$\frac{0,09}{2000}$
[22000 , 24000[	17	0,085	1	23000	$\frac{0,085}{2000}$
Total	200	1			

# Groupement de données en classe

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable **une répartition en classes** des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou **intervalle de classe**. En règle générale, on choisit des classes de même **amplitude**. Pour que la distribution en fréquence a un sens, il faut que chaque classe comprenne un nombre suffisant de valeurs ( $n_i$ ). Diverses formules empiriques permettent d'établir le **nombre de classes** pour un échantillon de taille  $n$ . On va donner deux règles pour les calculer

# Groupement de données en classe

La règle de **STURGE** : Nombre de classes =  $1 + (3,3 \log n)$

La règle de **YULE** : Nombre de classes =  $2,5\sqrt[4]{n}$

Amplitude de chaque classe est obtenu ensuite de la manière suivante :

$$\text{Amplitude de la classe} = \frac{(X_{\max} - X_{\min})}{\text{Nombre de classe}}$$

Avec  $X_{\max}$  et  $X_{\min}$  respectivement la plus grande et la plus petite valeur de  $X$  dans la série statistique.

Exemple: Dans le cadre de l'étude de la population de **gélinottes huppées** (*Bonasa umbellus*), Parmi les caractères mesurés figure **la longueur de la rectrice centrale** (plume de la queue). Les résultats observés exprimés en millimètres sur un échantillon de 50 males juvéniles sont notés dans la série ci-dessous

# Groupement de données en classe



**La gâlinette huppée**

153	165	160	150	159	151	163
160	158	149	154	153	163	140
158	150	158	155	163	159	157
162	160	152	164	158	153	162
166	162	165	157	174	158	171
162	155	156	159	162	152	158
164	164	162	158	156	171	164
158						

# Grouperment de données en classe

Dans le cadre de l'étude de la population de **gélinothtes huppées** (*Bonasa umbellus*), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

- Nombre de classes:

Règle de Sturge:  $= 1 + (3,3 \log 50) = 6,60$

Règle de Yule:  $2,5 \sqrt[4]{50} = 6,64$

Les deux valeurs sont très peu différentes.

Amplitude de chaque classe:  $\frac{174-140}{6,6} = 5,15mm$  que l'on arrondit à 5mm par commodité



# Groupeement de données en classe

- **Tableau de distribution des fréquences**

<b>Caractère <math>X</math> :</b> $x_i$ : longueur de la rectrice bornes des classes	[140-145[	[145-150[	[150-155[	[155-160[	[160-165[	[165-170[	[170-175[
Valeur médiane des classes, $x_i'$	142,5	147,5	152,5	157,5	162,5	167,5	172,5
$n_i$ : nombre d'individu par classe de taille $x_i$	1	1	9	17	16	3	3
$f_i$ : fréquence relative	0,02	0,02	0,18	0,34	0,32	0,06	0,06
$f_i \text{ cum.}$ : fréquence relative cumulée	0,02	0,04	0,22	0,56	0,88	0,94	1

# Représentations graphiques

## 6. Histogramme:

Il s'agit d'un diagramme composé de rectangle dont les bases sont les classes de la variable et dont les surfaces sont proportionnelles aux fréquences de ces classes .

Lorsque le caractère étudié est quantitatif et continu, et lorsque les modalités sont regroupées en classes,, on peut représenter la série par un histogramme: l'aire de chaque rectangle est alors proportionnelle à l'effectif (*ou à la fréquence*) associée à chaque classe.

a. **Classe d'amplitude égales:** Si toutes les classes ont la même amplitude  $e_i$  , on porte directement en ordonnée les effectifs  $n_i$  (respectivement les fréquences  $f_i$ )

# Représentations graphiques

b. **Classe d'amplitude inégales:** On calcule les effectifs (*resp les fréquences*) rectifiés et qui sont:

$$n'_i = \frac{\alpha n_i}{e_i} \quad \text{ou} \quad f'_i = \frac{\alpha f_i}{e_i}$$

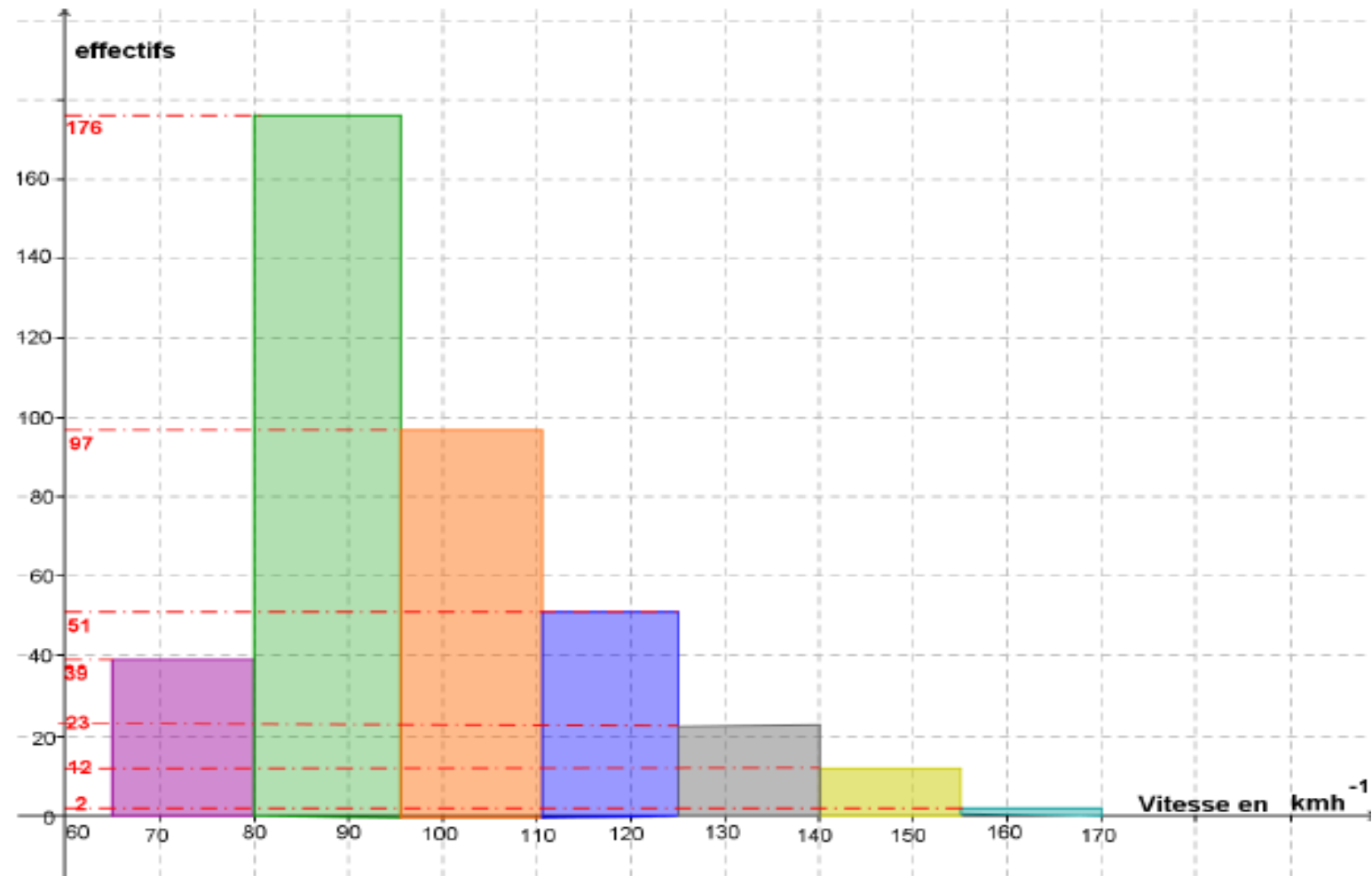
Où  $\alpha$  est choisi une amplitude de référence, en particulier on peut prendre  $\alpha$  la plus petite amplitude et on porte en ordonnée les effectifs rectifiés  $n'_i$  ou les fréquences rectifiées  $f'_i$ .

# Représentations graphiques

**Exemple32:** On étudie la vitesse de 400 véhicules enregistrée par un radar lors d'un contrôle routier

Classes vitesse en Km/h	[65, 80[	[80, 95[	[95, 110[	[110, 125[	[125, 140[	[140, 155[	[155, 165[
effectif	39	176	97	51	23	12	2

# Représentations graphiques



# Représentations graphiques

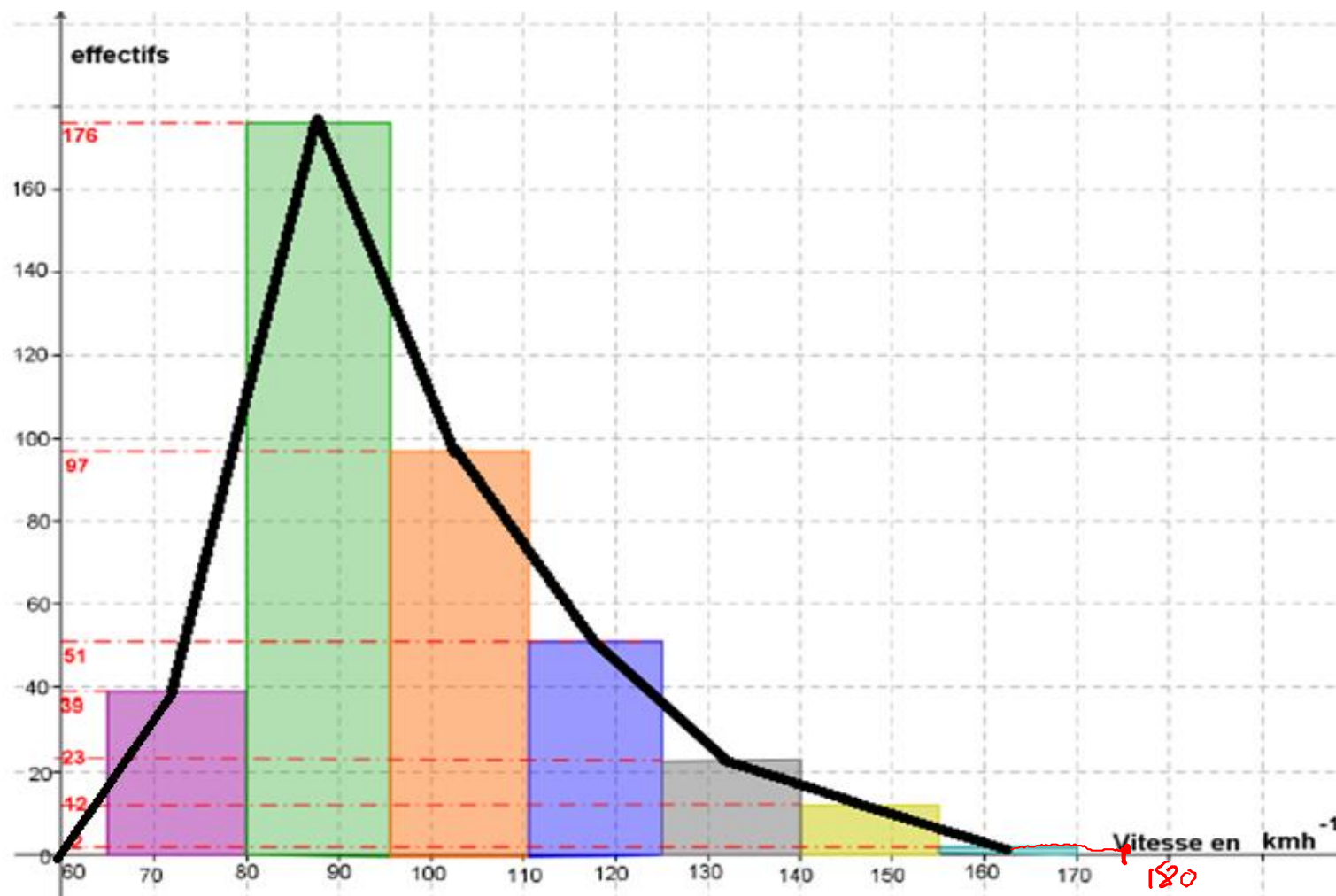
## 7. Polygone des effectifs ou des fréquences:

Le Polygone des effectifs (respectivement des fréquences) de la distribution statistique  $\{([a_i, a_{i+1}[ , n_i) \mid 1 \leq i \leq p \}$  s'obtient en joignant les points  $B_i(m_i, n_i)$

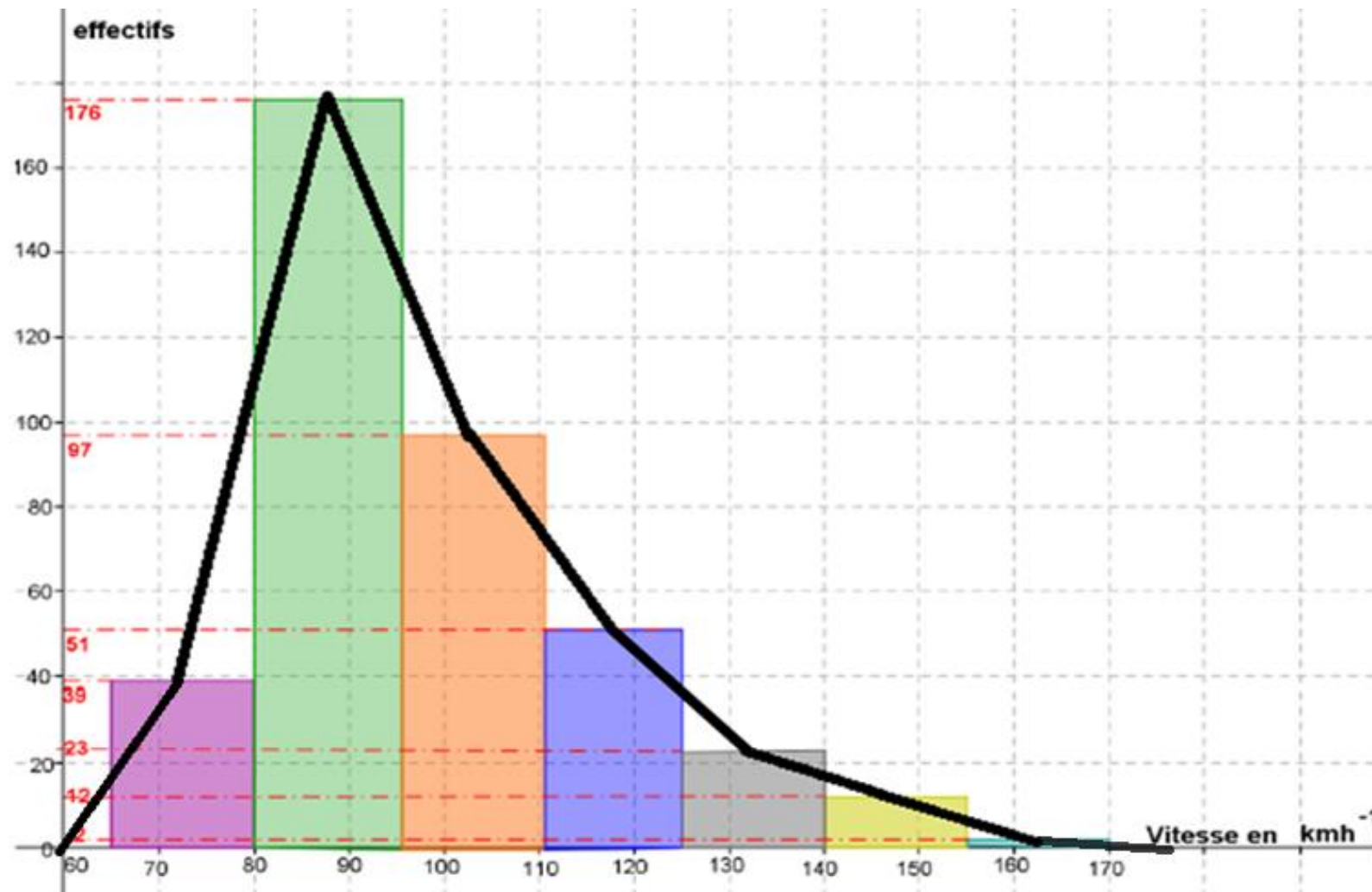
(respectivement des fréquences  $B_i(m_i, f_i)$ ) pour  $1 \leq i \leq p$  où  $m_i = \frac{a_i + a_{i+1}}{2}$ ,

Pour le tracer on ajoute deux classes fictives de hauteur nulles aux extrémités tout en veillant au respect de la conservation des aires (l'amplitude de chacune des

# Représentations graphiques



# Représentations graphiques





# Représentations graphiques

## 8. Fonction de répartition ou courbe cumulative

**Définition:** On appelle fonction de répartition d'une série statistique, la fonction, notée  $F$ , définie pour tout  $x \in \mathbb{R}$  par:

$$\begin{aligned} F(x) &= \text{fréquence} \text{ (des observations } \leq x \text{)} \\ &= \text{proportions} \text{ (des observations } \leq x \text{)} \end{aligned}$$

- La courbe cumulative s'obtient en joignant les points d'abscisses : la borne supérieure de la classe, et d'ordonnée: la fréquence cumulée correspondante. Autrement dit: on joint les points de coordonnées  $(a_i, F_i)$ .
- La courbe cumulative permet de lire pour chaque valeur de , le pourcentage des fréquences  $\leq x$ .

# Représentations graphiques

Remarque:

Pour  $a < b$

$F(a) - F(b)$  = proportion des observations dans  $]a, b]$

Proposition:

$$F(x) = F_{i-1} + \frac{F_i - F_{i-1}}{a_i - a_{i-1}} (x - a_{i-1}), \quad x \in ]a_{i-1}, a_i],$$

Où  $F_i = F(a_i)$  est la fréquence cumulée de  $a_i$ ,

$$F_i = F(a_i) = f_1 + f_2 + \dots + f_i$$

# Représentations graphiques

Propriétés de la fonction de répartition:

- $F$  est une fonction définie sur  $\mathbb{R}$ .
- $F$  est une fonction continue sur  $\mathbb{R}$ .
- $F$  est une fonction croissante.
- $\forall x \in \mathbb{R}, 0 \leq F(x) \leq 1$ .

# Représentations graphiques

**Exemple:** Soucieuse des problèmes de retard rencontrés sur les lignes ferroviaire et des nombreuses réclamations introduites par les navetteurs, la SNCB (société nationale des chemins de fer belges) a pris le parti de faire un relevé systématique du retard moyen (en minute) observé chaque jour sur certaines de ses grandes lignes.

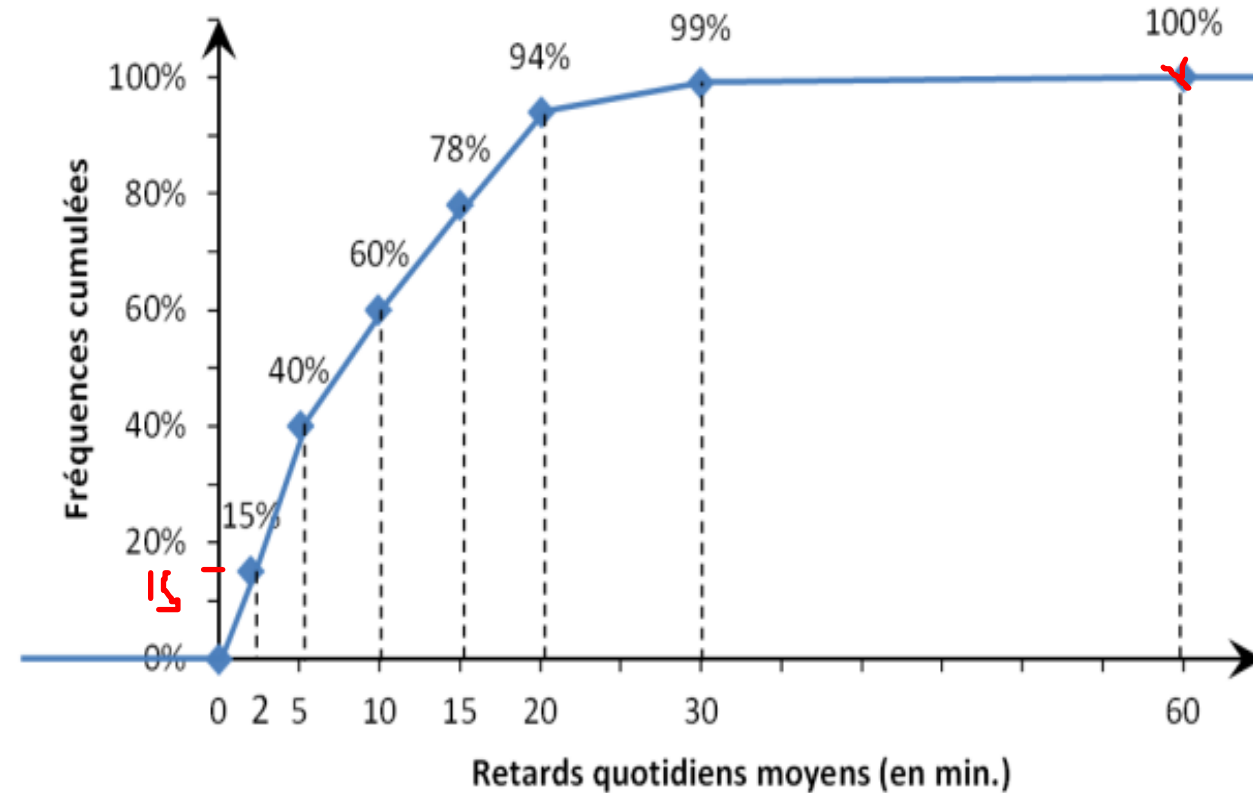
Les résultats relatifs aux 200 derniers jours de l'année 2010 pour la ligne Bruxelles-Liege ont donné lieu au tableau suivant :

# Représentations graphiques

Exemple33:

classe	Effectif $n_i$	Fréquence : $f_i$ en %	Fréquence cumulé en %
[0 , 2]	30	15%	15%
]2 , 5]	50	25%	40%
]5 , 10]	40	20%	60%
]10 , 15]	36	18%	78%
]15 , 20]	32	16%	94%
]20 , 30]	10	5%	99%
]30 , 60]	2	1%	100%
<i>Total</i>	$n=200$	100%	

# Représentations graphiques



*Courbe cumulative des fréquences*

# Représentations graphiques

## Question 1:

Quelle est la part des retards quotidiens moyens qui dépassent 10 minutes ?

## Réponse:

La courbe cumulative nous indique que 60% des retards quotidiens moyens étaient inférieurs ou égaux à 10 minutes. La part des retards quotidiens moyens qui dépassent 10 minutes s'élève donc à 40%.

# Représentations graphiques

## Question2:

Durant les 200 jours d'observation, combien y a-t-il eu de jours pour lesquels le retard quotidien moyen était supérieur à 15 minutes mais inférieur ou égal à 30 minutes ?

**Réponse:** La courbe cumulative nous indique que 78% des retards quotidiens moyens étaient inférieurs ou égaux à 15 minutes et 99% étaient inférieurs ou égaux à 30 minutes. Il s'ensuit que  $99\% - 78\% = 21\%$  des retards quotidiens moyens étaient supérieurs à 15 minutes mais inférieurs ou égaux à 30 minutes, ce qui correspond à  $0,21 \times 200 = 42$  jours d'observation.



# Représentations graphiques

## Exercice1: Données groupées

La série statistique ci-dessous donne le nombre de voitures vendues au cours du dernier mois par chacun des 40 distributeurs de la marque Renault.

7 – 1 – 5 – 12 – 3 – 2 – 1 – 4 – 7 – 2

6 – 4 – 1 – 8 – 10 – 3 – 3 – 6 – 5 – 2

• 5 – 8 – 2 – 6 – 0 – 7 – 4 – 4 – 6 – 8

5 – 5 – 4 – 7 – 8 – 10 – 6 – 5 – 3 – 3

Variable étudiée = Nombre de voitures vendues au cours du dernier mois.

Effectif total = Nombre de valeurs observées =  $N = 40$

# Représentations graphiques

## Questions:

1. Tracer un tableau statistique pour cette série contenant la fréquence, l'effectif cumulé et la fréquence cumulée
2. Tracer le diagramme en bâton des effectifs et en pointillé le polygone des effectifs
3. Tracer le polygone des fréquences cumulées (Diagramme de la fonction de répartition)

# Représentations graphiques

## Exercice2: Données groupées

Le responsable du stock d'un atelier a noté, au cours de 98 jours de travail, le nombre de boulons d'un certain type utilisés dans cet atelier. Les données brutes sont données ci-dessous

~~72~~-~~51~~-~~56~~ 95 -~~68~~ -~~66~~- ~~77~~ -81- 83- ~~75~~- 41 -~~79~~ -92 -~~78~~ -85 -55 -104 -76- ~~80~~- ~~61~~  
~~65~~- ~~70~~ -83- 92- 88 -59 -~~75~~ -~~75~~- 81- ~~69~~- ~~71~~- 96- 101- 87- ~~65~~- ~~74~~-~~68~~ -~~73~~ -~~78~~ -~~68~~  
~~73~~ -86- 84- ~~51~~- 85 -~~75~~ -~~79~~ -90 -~~68~~ -~~71~~ -~~75~~ -~~74~~ -81- ~~64~~ -88 -~~78~~ -~~77~~ -~~66~~ -91 -75  
~~69~~- ~~73~~ -82 -~~75~~ -~~76~~ -~~71~~ -~~74~~ -96 -~~72~~ -~~74~~ -102- ~~74~~ -80 -82 -86 -~~78~~ -87 -~~61~~ -~~80~~ -~~78~~  
48- ~~68~~- ~~71~~- ~~66~~- 59- 92- ~~77~~- ~~76~~ -81 -~~70~~ -85 -~~77~~- ~~68~~ -82- ~~78~~- 75- 91- ~~77~~.

1. Regrouper ces données en classes de longueur 10 et donner les distributions de fréquences.
1. Tracer le histogramme des effectifs et pointiller le polygone des effectifs.
2. Tracer le polygone des fréquences cumulées (Diagramme de la fonction de

# Etude d'une variable quantitative: Paramètres

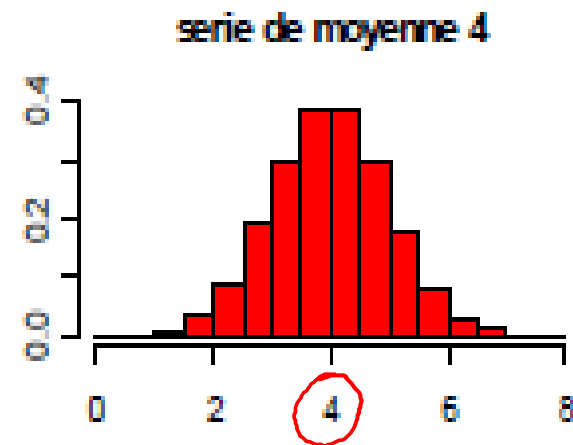
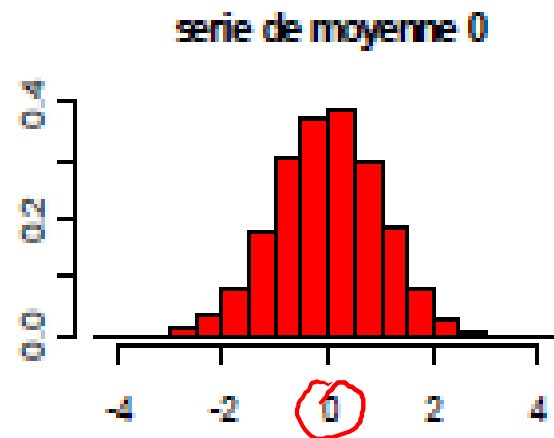
**Objectif** : caractériser la distribution de la série à l'aide de nombres et éventuellement de graphiques résumant de façon suffisamment complète l'ensemble ses valeurs. Ces paramètres faciliteront la comparaison d'échantillons.

Il y a trois types de paramètres :

- Paramètres de tendance centrale.
- Paramètres de dispersion.
- Paramètres de forme.

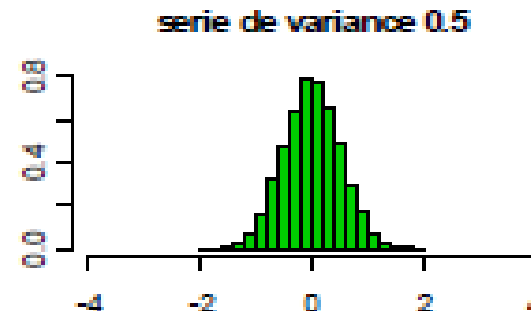
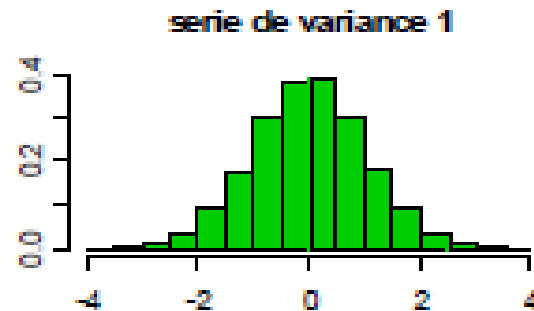
# Etude d'une variable quantitative: Paramètres

Les **paramètres de tendance centrale**: fournissent l'ordre de grandeur des valeurs de la série et la position où se rassemblent ces valeurs.



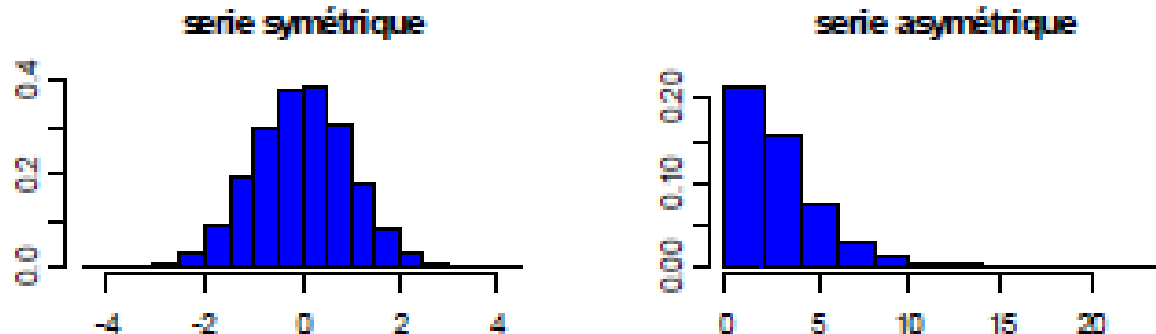
# Etude d'une variable quantitative: Paramètres

**Les paramètres de dispersion:** quantifient les fluctuations des valeurs autour de la valeur centrale. Permettant d'apprécier l'étalement des valeurs de la série (les unes par rapport aux autres ou à la valeur centrale)



# Etude d'une variable quantitative: Paramètres

**Les paramètres de forme:** donnent une idée de la symétrie et de de l'aplatissement d'une distribution. Leur usage est moins fréquent.



# Paramètres de position

## 1. Le Mode:

**Définition:** On appelle mode d'une série statistique, la modalité ou la valeur qui a le plus grand effectif (valeur de fréquence maximale), on le note **Mo**

**Remarque:**

- Une série statistique peut avoir plusieurs modes.
- Dans le cas de distributions groupées, on parle de classe modale (classe de *plus grand effectif*)



# Paramètres de position

Note de TP ( <i>physique</i> )	Effectif
8	3
9	2
10	4
12	6
13	1
14	3
16	1

Le mode=12

# Paramètres de position

## b. Cas d'une variable continue

- Dans le cas d'amplitude égale

$$Mo = a_{i-1} + (a_i - a_{i-1}) \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$$

- $a_{i-1}$  borne inférieur de la classe modale
- $a_i$  borne supérieur de la classe modale
- $n_i$  l'effectif de la classe modale
- $n_{i-1}$  l'effectif de la classe avant modale
- $n_{i+1}$  l'effectif de la classe post modale

**Remarque:** Dans le cas d'amplitudes inégales, il faut corriger les effectifs pour déterminer le mode

# Paramètres de position

$\pi$

**Exemple35:** Le revenu mensuel d'un groupe de 220 ingénieurs se répartissaient comme suit:

Revenus en DH	Nombre d'ingénieurs
[6000, 8000[	40
[8000, 10000[	37
[10000, 12000[	50
[12000, 14000[	16
[14000, 16000[	12
[16000, 18000[	10
[18000, 20000[	20
[20000, 22000[	18
[22000, 24000[	17

Classe modale = [10000, 12000[

$$Mo = 10000 + 2000 \frac{50 - 37}{(50 - 16) + (50 - 37)} = 10553,191 \text{ DH}$$

# Paramètres de position

## 2. La moyenne arithmétique:

Par la suite, on suppose que la variable statistique est une variable quantitative.

**Définition:** La moyenne arithmétique notée  $\mu$  ou  $\bar{x}$ , d'une série quantitative est donnée par:

$$\begin{aligned}\mu &= \frac{n_1 x_1 + n_2 x_2 + \dots + n_p x_p}{n_1 + n_2 + \dots + n_p} \\ &= \frac{1}{N} \sum_{i=1}^p n_i x_i = \sum_{i=1}^p f_i x_i\end{aligned}$$

Où  $x_i$  sont les valeurs observées (pour les distributions non groupées) ou les milieux des classes (pour les distributions groupées).

Remarque: La moyenne arithmétique est un paramètre très affecté par les valeurs extrêmes (attention aux points aberrants).

# Paramètres de position

**Exemple 36:** Dans l'exemple 34, la moyenne vaut

$$\mu = \frac{1}{20} (8 \times 3 + 9 \times 2 + 10 \times 4 + 12 \times 6 + 13 \times 1 + 14 \times 3 + 16 \times 1) = \frac{225}{20} = 11,25$$

**Exemple 37:** Calculons la moyenne dans l'exemple 35. Nous allons commencer par déterminer les milieux  $m_i$  des classes  $x_i$ .

$m_i$	7000	9000	11000	13000	15000	17000	19000	21000	23000
$n_i$	40	37	50	16	12	10	20	18	17

La moyenne est donnée par:

$$\mu = \frac{1000}{220} (7 \times 40 + 9 \times 37 + 11 \times 5 + 16 \times 13 + \dots + 23 \times 17) \approx 13045 \text{ dh.}$$

# Paramètres de position

## Propriétés

**Propriétés 1:** lorsqu'on ajoute un terme constant aux valeurs de la variable, la moyenne arithmétique de la série est augmenté de ce terme constant,

autrement:

Soit la série  $(x_i, n_i)$  dont la moyenne est  $\mu$ ; définissons une nouvelle série  $(x_i + a, n_i)$  avec  $a$  une constante et notons  $\mu'$  la moyenne de cette série alors

$$\mu' = \mu + a$$

**Propriétés 2:** En multipliant les valeurs d'une série statistique par un terme constant, la moyenne arithmétique de cette série est aussi multipliée par ce terme constant,

autrement:

Soit la série  $(x_i, n_i)$  dont la moyenne est  $\mu$ ; la nouvelle série  $(a \cdot x_i, n_i)$  aura pour moyenne arithmétique  $\mu'$  telle que

$$\mu' = a \cdot \mu$$

# Paramètres de position

En général:

La transformation linéaire d'une variable statistique  $x$  est une variable  $y$  de la forme  $y = ax + b$ . Alors la moyenne de  $y$  noté  $\mu'$  est donnée par

$$\mu' = a\mu + b$$

Avec  $\mu$  est la moyenne arithmétique de la variable  $x$

# Paramètres de position

**Propriétés 3:** La moyenne arithmétique d'une série regroupant plusieurs séries peut être obtenu en calculant la moyenne des moyennes arithmétiques de ces séries.

**Autrement:**

soient  $\Omega_1$  et  $\Omega_2$  deux population de tailles respectives  $N_1$  et  $N_2$  avec

$\Omega_1 \cap \Omega_2 = \emptyset$ . Si une variable  $X$  définie sur  $\Omega_1$  admet  $\mu_1$  comme moyenne et la même variable statistique  $X$  définie sur  $\Omega_2$  admet  $\mu_2$  comme moyenne, alors la moyenne  $\mu$  de la variable statistique  $X$  sur  $\Omega = \Omega_1 \cup \Omega_2$  est donnée par

$$\mu = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$$



# Paramètres de position

## 3. La Médiane:

**Définition:** La médiane  $Me$  d'une série statistique, rangée en ordre croissant ou décroissant, est un nombre qui partage l'effectif total  $N$  de cette série en deux parties égales.

$$\begin{array}{ccc} 50\% & & 50\% \\ \hline \dots\dots Me \dots\dots \end{array}$$

## Détermination pratique de La Médiane

### a. Cas d'une variable discrète

- › **Cas de fréquence unitaire** Lorsque les données sont présentées individuellement, chaque donnée a la même fréquence unitaire d'apparition, leur **effectif** est égal à 1

Considérons une population de  $n$  observations, les valeurs de la variable étant rangées par ordre de grandeur croissant :  $x_1 < x_2 < x_3 < \dots < x_N$

- Si  $N$  est impair càd  $N=2k + 1$ , la médiane sera alors le nombre  $Me=x_{k+1}$ .
- Si  $N$  est pair càd  $N=2k$ , la médiane sera alors le nombre  $Me = \frac{x_k+x_{k+1}}{2}$ .

# Paramètres de position

**Exemple38:** Considérons la série statistique

$$2 - 11 - 5 - 1 - 7 - 6 - 8 - 9 - 4.$$

On range la série statistique par ordre croissant,

$$\begin{array}{c} \downarrow \\ 1 - 2 - 4 - 5 - 6 - 7 - 8 - 9 - 11 \end{array}$$

Puisque l'effectif total  $N = 9 = 2 \times 4 + 1$

Alors, la médiane  $Me = x_5 = 6$ .

**Exemple:** Considérons la série statistique

$$1 - 2 - 5 - 8 - 11 - 13$$

la série est rangé par ordre croissant, l'effectif total  $N = 6 = 2 \times 3$  Alors, la médiane

$$Me = \frac{x_3 + x_4}{2} = \frac{5 + 8}{2} = 6,5$$

# Paramètres de position

## ii. Cas de fréquence non unitaire

- On cherche la valeur  $\frac{N}{2}$  alors la valeur de la variable  $x_i$  correspond à l'effectif cumulé  $\frac{N}{2}$  représente la médiane.
- Si la valeur  $\frac{N}{2}$  ne figure pas sur la colonne des effectifs cumulés croissant, la valeur qui lui est juste supérieure est donc la médiane.

# Paramètres de position

**Exemple 39:** Répartition du nombre d'enfant par ménage

Nombre d'enfants $x_i$	Nombre de ménage $n_i$	Effectif cumulés
0	100	100
1	120	220
2	210	430
3 ↵	230	660 ↵
4	150	810
5	110	920
6	80	1000
Total	1000	

$$\frac{N}{2} = 500 \text{ alors } Me=3$$

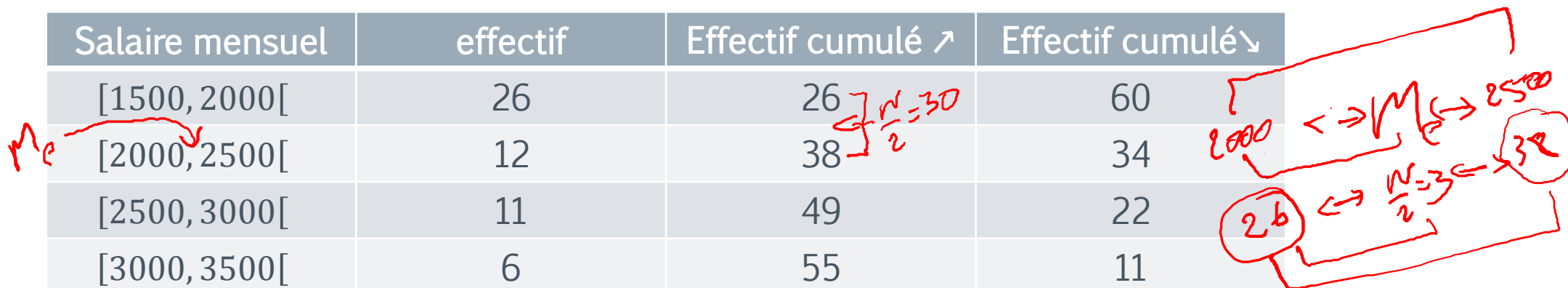
# Paramètres de position

## b. Cas d'une variable continue

- **Détermination algébrique:** Pour calculer la médiane on utilise la méthode d'interpolation linéaire, qui consiste premièrement à déterminer l'intervalle médiane  $c$  à  $d$  l'intervalle qui contient la valeur  $\frac{N}{2}$ .

**Exemple 40:** Le tableau suivant donne la répartition du personnel d'une entreprise selon leur salaire mensuel en DH

Salaire mensuel	effectif	Effectif cumulé ↗	Effectif cumulé ↘
[1500, 2000[	26	26	60
[2000, 2500[	12	38	34
[2500, 3000[	11	49	22
[3000, 3500[	6	55	11
3500 et plus	5	60	5
Total	60		



# Paramètres de position

Le calcul de **Me** s'effectue en trois étapes :

- **1<sup>ère</sup> étape:** Détermination du rang de la médiane: Le rang de Me est  $\frac{60}{2} = 30$ , donc Me correspond au salaire du 30<sup>ème</sup> individu.
- **2<sup>ème</sup> étape:** Détermination de la classe Me: En consultant la colonne des effectifs cumulés croissant, on voit que Me appartient à la classe  $[2000, 2500[$  et on applique le schéma suivant:

$$\begin{array}{c} 26 \rightarrow 30 \rightarrow 38 \\ 2000 \rightarrow Me \rightarrow 2500 \end{array}$$

$$\frac{Me - 2000}{2500 - 2000} = \frac{30 - 26}{38 - 26}$$

D'où  $Me = 2166,67 \text{ DH}$

# Paramètres de position

Généralement:

$$Me = a_{i-1} + (a_i - a_{i-1}) \frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}$$

$Me \in [a_{i-1}, a_i[$  et  $N_i$  représente l'effectif cumulé associé à la classe  $[a_{i-1}, a_i[$   
classe de la médiane  $Me$  (càd classe contenant  $Me$ )

# Paramètres de position

## Interprétation:

- 50% des salariés touchent un salaire mensuel inférieur à 2166,67 *DH*.
- 50% des salariés touchent un salaire mensuel supérieur à 2166,67 *DH*

## Détermination graphique de la médiane:

La détermination graphique de la médiane consiste à tracer la courbe des effectifs cumulés croissantes (ou des fréquences cumulés croissant) et la courbe effectifs cumulés décroissantes (ou des fréquences cumulés



# Paramètres de position

## Remarque:

On peut se limiter à tracer uniquement la courbe cumulative croissante ou la courbe cumulative décroissante, dans ce cas, la Me est l'abscisse du point d'intersection de cette courbe avec la droite parallèle à l'axe des  $(ox)$ , d'équation  $y = \frac{N}{2}$

## Remarque:

La médiane est plus robuste que la moyenne (pas influencée par les valeurs extrêmes) mais elle est influencée par le nombre d'observations.

# Paramètres de position

## Symétrie

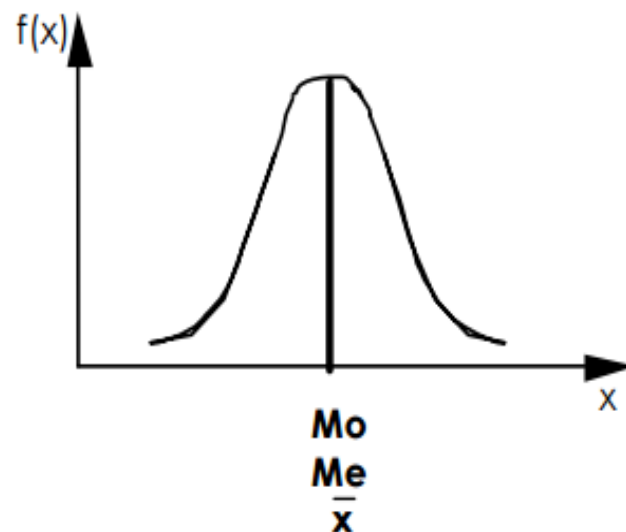
**Définition:** Une série a une distribution symétrique si ses valeurs sont également dispersées de part et d'autre de la valeur centrale, c'est-à-dire si le graphe de la distribution histogramme ou diagramme en bâton en fréquences admet une axe de symétrie.

# Paramètres de position

## Position relative du mode, de la médiane et de la moyenne

Renseigne sur une caractéristique de forme de la distribution, à savoir l'asymétrie.

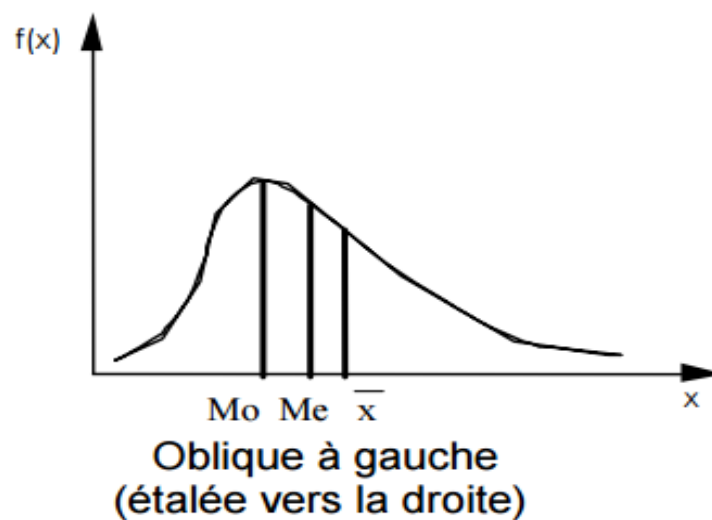
- La distribution est **symétrique** quand le mode, la médiane et la moyenne sont confondus.



**$Mo = Me = \bar{x}$**   
courbe normale  
ou courbe de Gauss  
ou courbe en cloche

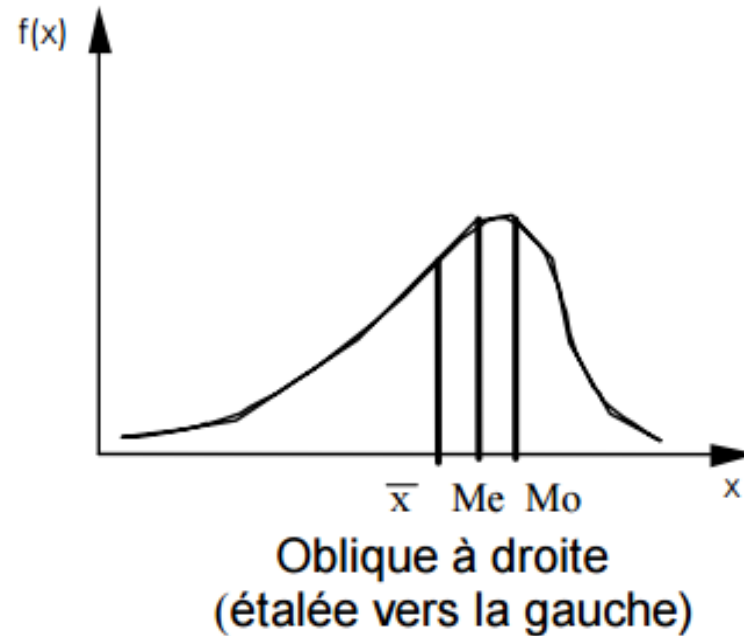
# Paramètres de position

- Distribution **asymétrique** ou **dissymétriques**: On a deux cas  
Si **Mode** < **médiane** alors on dit que la distribution est étalée à droite



# Paramètres de position

Si **Mode** > **médiane** alors on dit que la distribution est étalée à gauche



# Paramètres de position

La moyenne et la médiane plus généralement ne sont pas suffisantes pour donner des interprétations, il faut introduire d'autre paramètre qui permettent de mesurer la dispersion d'une série autour de sa moyenne. Les caractéristiques de dispersion, vont nous donner des indications sur Les données les unes par rapport aux autres : sont-elles près de leur centre ? Proches les unes des autres ? Sont-elles dispersées ?

Les principaux éléments de mesure de dispersion sont l'étendue, L'écart interquartile, l'écart type, la variance, le coefficient de variation.

# Paramètres de dispersion

$\pi$

## 4. L'étendue

**Définition:** On appelle étendue d'une série statistique, la différence entre les deux valeurs extrêmes de la série càd la plus grande valeur moins la plus petite valeur.

Dans le cas d'une variable continue l'étendue est la différence entre la borne supérieure de la dernière classe et la borne inférieure de la première classe .

## 5. Quantiles:

**Définition:** On appelle quantile d'ordre  $\alpha$ , noté  $q_\alpha$  , la valeur qui partage la population en deux effectifs. Une partie de la population possède un caractère inférieur à  $q_\alpha$  et une partie un caractère supérieur à  $q_\alpha$ .

A partir de cette définition nous pouvons définir les quartiles, les déciles et les centiles.

# Paramètres de dispersion

a. **Quartiles** : Dans une série statistique, il y a trois quartiles

- $Q_1$  (quartile inférieur): 25% d'observations inférieures et 75% d'observations supérieures
- $Q_2$  (quartile central): c'est la médiane
- $Q_3$  (quartile supérieure): 75% d'observations inférieures et 25% d'observations supérieures

Les 3 quartiles  $Q_1$ ,  $Q_2$  et  $Q_3$  divisent la distribution statistique en quatre classes :

$$[x_1, Q_1[ \quad [Q_1, Q_2[ \quad [Q_2, Q_3[ \quad \text{et} \quad [Q_3, x_k[$$



# Paramètres de dispersion

Ayant toute le même nombre d'observation:  $\frac{N}{4}$  (ou 25% des effectifs)

## Remarque:

- Les quartiles se calculent de façon analogue a celle de la médiane.
- L'intervalle  $[Q_1 \ Q_3]$  s'appelle intervalle interquartile.
- La différence  $Q_3 - Q_1$  s'appelle écart interquartile.

## b. Déciles :

Ce sont les valeurs du caractère qui partagent la série en dix sous ensembles égaux comprenant chacun  $\frac{1}{10}$  ème des effectifs. Ils sont au nombre de 9.

- $D_1$  laisse 10% des observations avant et 90% après.
- $D_2$  laisse 20% des observations avant et 80% après.
- $D_5 = Me$  c'est la médiane.
- $D_9$  laisse 90% des observations avant et 10% après.

# Paramètres de dispersion

L'intervalle  $[D_1, D_9]$  s'appelle intervalle interdécile, il contient 80% des observations.

## c. Centiles

Les centiles partagent la distribution en 100 parties de même effectifs. Ils sont notés  $C_1, C_2, \dots, C_{99}$  avec  $C_{50} = Me$ .

## Détermination des quartiles

### i. Cas variable discrète:

**Exemple:** on fait une étude statistique sur les 50 notes attribuées par un jury à un examen, voici les résultats obtenus en classant ces notes par ordre croissant (variable discrète ).

# Paramètres de dispersion

Note	Effectif	Effectif cumulé
0	1	1
1	2	3
2	2	5
3	3	8
4	2	10
5	3	13
6	2	15
7	3	18
8	4	22
9	3	25

$Q_5$  (5)  $\rightarrow$  12,5 =  $\frac{N}{4}$   
 $Q_2$  (9)  $\rightarrow$  25 =  $\frac{N}{2}$

Note	Effectif	Effectif cumulé
10	2	27
11	3	30
12	4	34
13	4	38
14	3	41
15	1	42
16	2	44
17	1	45
18	2	47
19	2	49
20	1	50

$Q_3$  (13)  $\rightarrow$  38 =  $\frac{3N}{4}$

# Paramètres de dispersion

## *Détermination des quartiles*

$\frac{N}{4} = 12,5$  n'est pas un entier donc le premier quartile est le terme de rang 13  
soit  $Q_1 = 5$

$\frac{3N}{4} = 37,5$  n'est pas un entier donc le troisième quartile est le terme de rang 38  
soit  $Q_3 = 13$

Le premier quartile  $Q_1 = 5$

Le second quartile ou la médiane  $Q_2 = 9$

Le troisième quartile  $Q_3 = 13$

# Paramètres de dispersion

## ii. Cas variable continue

**Exemple:** Le tableau suivant donne la répartition de 80 entreprise d'une certaine région selon le montant des investissement réalisés pendant une période donnée

Taches d'investissement 1000 DH	Nombre d'entreprise $n_i$	Effectif cumulé $\nearrow N_i$
[200, 300[	10	10
[300, 400[	18	28
[400, 500[	30	58
[500, 600[	12	70
[600, 700[	6	76
[700, 800[	4	80
Total	80	

Handwritten notes on the table:

- Red arrows pointing to the first and third rows of the data (excluding the header) are labeled  $Q_1$  and  $Q_3$  respectively.
- Red circles are drawn around the cumulative frequencies 28 and 70.
- Handwritten calculations:  $\frac{N}{n} = 20$  (pointing to 28) and  $\frac{3n}{4} = 60$  (pointing to 70).

# Paramètres de dispersion

Pour calculer  $Q_1$ : On suit trois étapes:

- Rang de  $Q_1$ : d'abord on calcule la valeur de  $\frac{N}{4}$ , dans ce cas on trouve  $\frac{N}{4} = 20$
- Classe contenant  $Q_1$ : Le tableau nous donne la classe contenant  $Q_1$ , ici  $[300, 400[$
- Interpolation linéaire

$$\begin{array}{l} 300 \rightarrow 10 \\ Q_1 \rightarrow 20 \\ 400 \rightarrow 28 \end{array}$$

$$\frac{Q_1 - 300}{400 - 300} = \frac{20 - 10}{28 - 10}$$

Donc  $Q_1 = 355,55 \times 10^3 DH$

# Paramètres de dispersion

$\pi$

Interprétation :

25% des entreprises investissent un montant inférieur à  $355,556.10^3 DH$ , alors que 75% des entreprises investissent un montant supérieur à  $355,556.10^3 DH$ .

Pour la médiane, on a :  $Me = 400 + (500 - 400) \times \frac{40-28}{58-28}$

$$Me = 440.10^3 DH$$

De la même façon on détermine  $Q_3$ :

- Rang de  $Q_3$  est  $\frac{3N}{4} = 60$
- Classe contenant  $Q_3$  est  $[500; 600[$
- Interpolation linéaire :

$500 \rightarrow 58$   
 $Q_3 \rightarrow 60$   
 $600 \rightarrow 70$

$$\frac{Q_3 - 500}{600 - 500} = \frac{60 - 58}{70 - 58}$$

# Paramètres de dispersion

$$Q_3 = 516,67.10^3 DH$$

## Interprétation:

75% des entreprises investissent un montant inférieur à  $516.670.10^3$  DHS, alors que 25% des entreprises investissent un montant supérieur à  $516.670.10^3$  DHS.

## Généralement:

$$Q_\alpha = a_{i-1} + (a_i - a_{i-1}) \frac{\frac{\alpha N}{4} - N_{i-1}}{N_i - N_{i-1}}$$

Avec  $\alpha = 1$  ou  $\alpha = 3$  et  $Q_\alpha \in [a_{i-1}, a_i[$  et  $N_i$  représente l'effectif cumulé associé à la classe  $[a_{i-1}, a_i[$  classe du quartile  $Q_\alpha$  (càd classe contenant  $Q_\alpha$ ).



# Paramètres de dispersion

**L'intervalle interquartile:** c'est l'intervalle  $[Q_1 \ Q_3]$  d'amplitude  $Q_3 - Q_1$  qui s'appelle écart interquartile on la note  $E_q = Q_3 - Q_1$

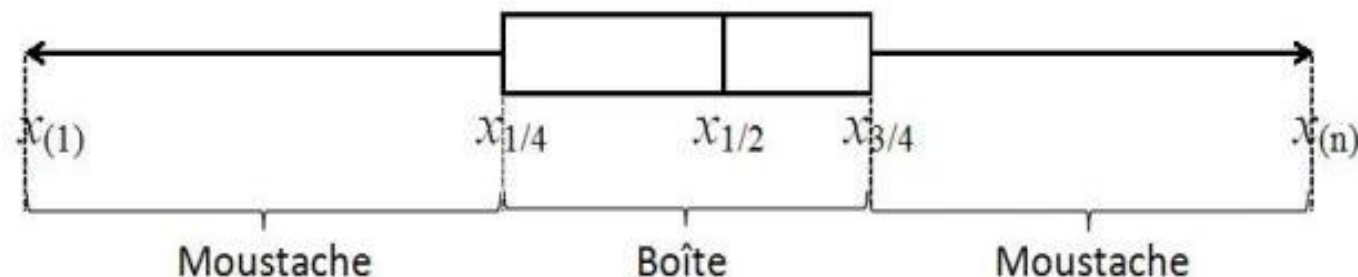
**Interprétation :**

L' écart interquartile est faible lorsque les observation de la série sont groupées autour de la médiane. Inversement, plus elles sont dispersées et plus l' écart interquartile a une valeur élevée.

# Paramètres de dispersion

## Boîte à moustaches (boxplot)

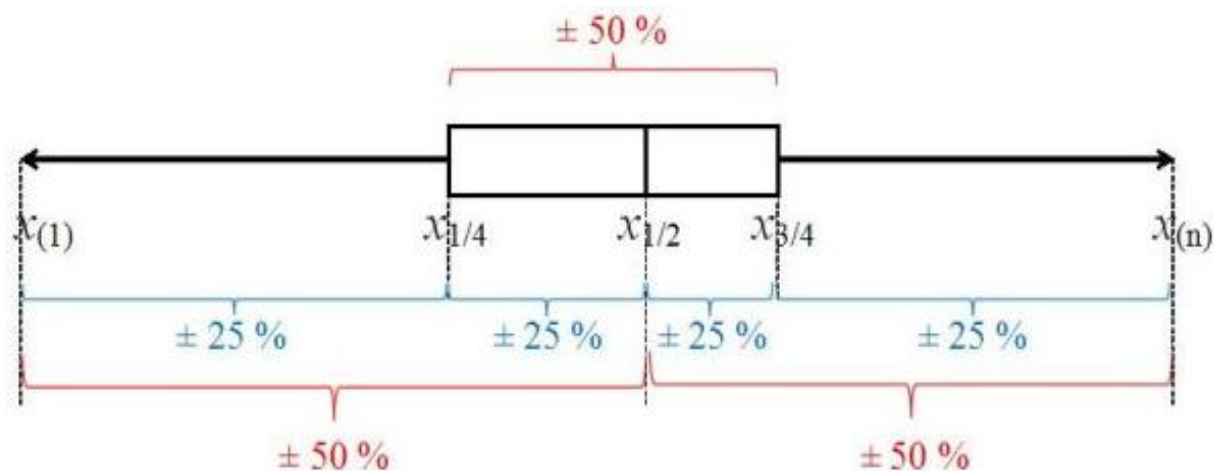
Il est possible de résumer, sous la forme d'un graphique, l'information fournie par l'étendue, ainsi que par les trois quartiles et les intervalles qui les séparent. Ce graphique porte le nom de boîte à moustaches, ou encore **boîte à pattes** ou **diagramme en boîte** (boxplot en anglais).



$x_{1/4}=Q_1$  le premier quartile,  $x_{1/2}=Me$  la médiane et  $x_{3/4}=Q_3$  le troisième quartile

Une boîte à moustaches nous indique de façon simple et visuelle quelques traits marquants de la série observée :

# Paramètres de dispersion



*Informations fournies par la version de base*

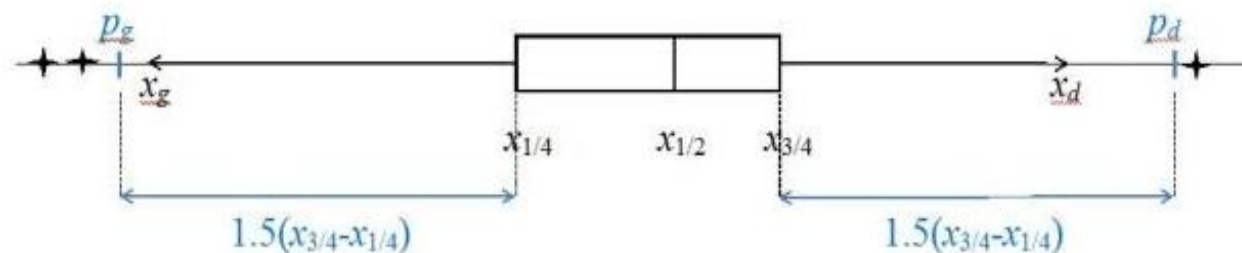
# Paramètres de dispersion

- la médiane nous renseigne sur le milieu de la série ;
- les largeurs des deux parties de la boîte rendent compte de la dispersion des valeurs situées au centre de la série (la boîte contient 50% (environ) de l'ensemble des observations : 25% à gauche de la médiane et 25% à sa droite) ;
- la longueur des moustaches renseigne sur la dispersion des valeurs situées au début de la série ordonnée (les valeurs les plus petites correspondant à 25% des observations) ou à la fin de celle-ci (les valeurs les plus grandes correspondant aussi à 25% des observations) ;
- de façon générale, la boîte et les moustaches seront d'autant plus étendues que la dispersion de la série statistique est grande.

# Paramètres de dispersion

## Remarque:

Quand la série observée contient l'une ou l'autre valeur extrême (très petite ou très grande), les moustaches risquent de devenir très longues, ce qui nuit à leur interprétation. La solution à ce problème consiste à construire plutôt **la version modifiée** de la boîte à moustaches.



*Version modifiée de la boîte à moustaches*

# Paramètres de dispersion

La version modifiée de la boîte à moustaches se construit en 4 étapes :

1. construction de la **boîte**, comme dans la version de base ;
2. calcul des **valeurs pivots gauche** ( $p_g$ ) et **droite** ( $p_d$ )
3. détermination des **valeurs adjacentes gauche** ( $x_g$ ) et **droite** ( $x_d$ ) ces valeurs adjacentes correspondent aux extrémités des moustaches gauche et droite ;
4. détermination des **valeurs extérieures** éventuelles.

## Les valeurs pivots

**Définition:** Les valeurs pivots sont définies par les relations suivantes :

$$\begin{cases} p_g = Q_1 - 1,5(Q_3 - Q_1) & \text{pivot gauche} \\ p_d = Q_3 + 1,5(Q_3 - Q_1) & \text{pivot droit} \end{cases}$$

# Paramètres de dispersion

Elles sont situées de part et d'autre de la boîte, à une distance valant 1.5 fois l'écart interquartile.

## Remarque:

La définition des valeurs pivots résulte d'une constatation : la plupart des séries statistiques qui ne contiennent pas de valeurs extrêmes ou aberrantes, ont leurs observations situées dans l'intervalle  $[p_g, p_d]$ .

## Remarque:

$p_g$  et  $p_d$  ne coïncident généralement pas avec des valeurs observées. Il s'agit juste de valeurs **calculées** dans le but de déterminer, dans un deuxième temps, les valeurs adjacentes.

# Paramètres de dispersion

## Les valeurs adjacentes (extrémités des moustaches)

Les valeurs adjacentes, contrairement aux valeurs pivots, doivent être des valeurs **observées** de la série statistique. Elles correspondront aux extrémités des moustaches gauche et droite du diagramme en boîte.

**Définition:** On définit les valeurs adjacentes par rapport aux valeurs pivots  $p_g$  et  $p_d$  comme suit

- la valeur adjacente **gauche**, notée  $x_g$ , est la **plus petite valeur observée supérieure ou égale à  $p_g$** .
- la valeur adjacente **droite**, notée  $x_d$  est la **plus grande valeur observée inférieure ou égale à  $p_d$**



# Paramètres de dispersion

## Les valeurs extérieures

Si toutes les observations  $x_i$  sont comprises entre le pivot gauche  $p_g$  et le pivot droit  $p_d$ , alors  $x_g = x_{(1)}$  et  $x_d = x_{(n)}$ . Dans le cas contraire, on isole les valeurs observées situées en dehors de l'intervalle  $[p_g, p_d]$  pour en examiner les caractéristiques.

### Définition:

Toutes les observations situées en dehors de  $[p_g, p_d]$  sont dites **extérieures**. Elles sont représentées par des symboles appropriés (étoiles, points, triangles, ...) de manière à être mises en évidence.

**Remarque:** Lorsque toutes les observations  $x_i$  sont comprises entre le pivot gauche  $p_g$  et le pivot droit  $p_d$ ,  $x_g = x_{(1)}$  et  $x_d = x_{(n)}$  et il n'y a pas de valeur extérieure. Dans ce cas, la version modifiée de la boîte à moustaches coïncide avec la version de base.

# Paramètres de dispersion

## Utilité de la boîte à moustache pour comparer des séries

La boîte à moustache permet de comparer des séries du point de vue de leur dispersion

mais aussi de leur caractéristique de tendance centrale (puisque la médiane est repérée).

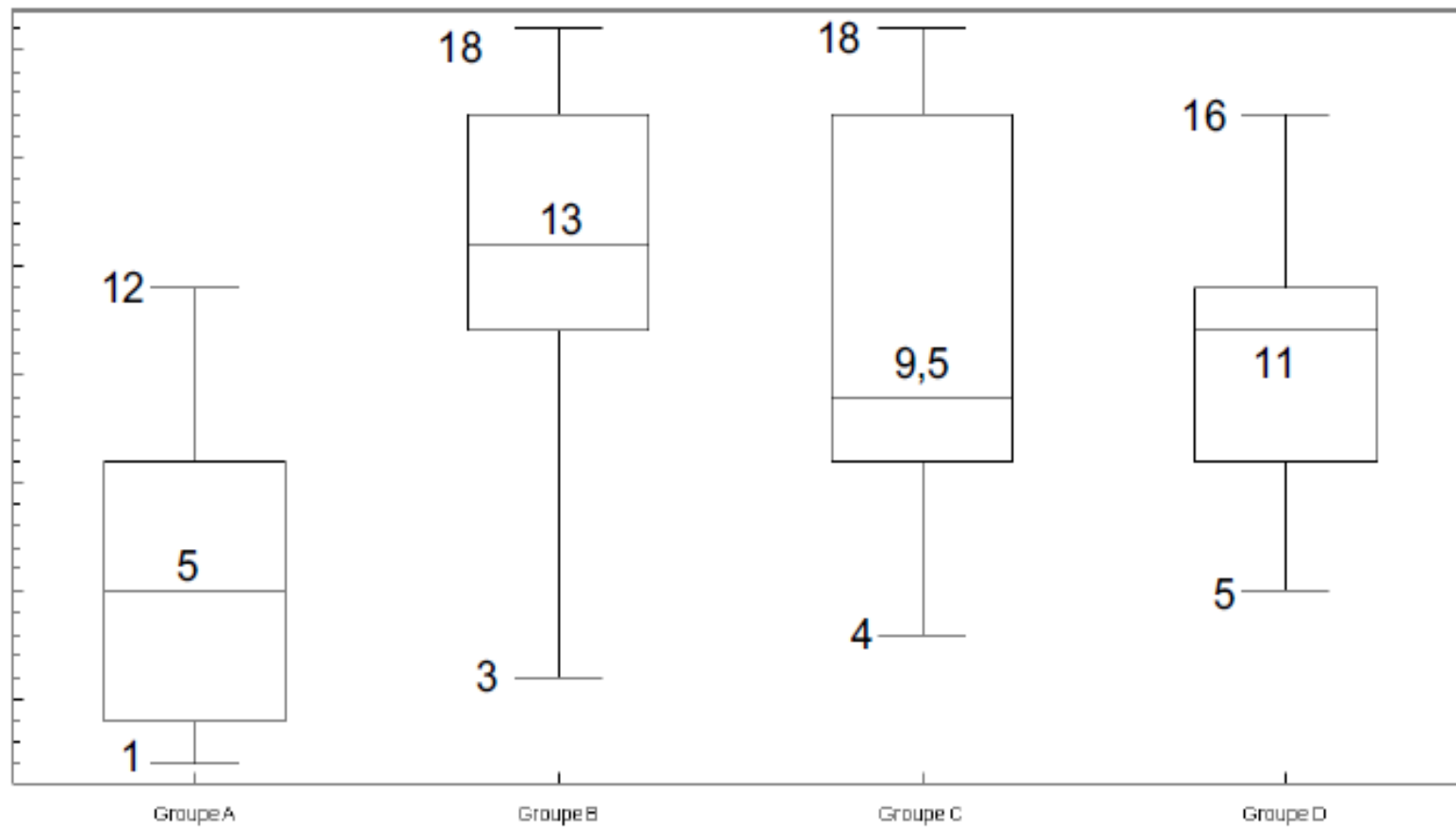
**Exemple** : soient les notes sur 20 de 4 groupes d'étudiants :

- Groupe A {1, 2, 2, 12, 5, 5, 9, 5, 7, 11, 7, 8, 2}
- Groupe B {16, 13, 15, 13, 11, 13, 16, 3, 18, 11}
- Groupe C {8, 8, 8, 7, 4, 16, 13, 16, 18, 11}
- Groupe D {12, 10, 6, 8, 5, 16, 12, 15, 10, 15, 12, 10}

## Paramètres de dispersion

La comparaison des graphiques boîtes à moustaches de chaque groupe permet d'avoir une bonne idée de la dispersion des notes, tout en visualisant la note médiane (qui est souvent jugée préférable à la note moyenne).

# Paramètres de dispersion



# Paramètres de dispersion

## Utilité de la boîte à moustache pour déterminer la forme d'une distribution

Suivant la position de la médiane au sein de la boîte, on peut en déduire des informations sur la forme de la distribution.

- 1) Si la médiane est proche du centre de la boîte, c'est que la distribution est symétrique.
- 2) Si la médiane est à gauche du centre de la boîte, c'est que la distribution est étalée à droite.
- 3) Si la médiane est à droite du centre de la boîte, c'est que la distribution est étalée à gauche.

De même, en comparant la longueur respective de chaque moustache, on peut en déduire des informations sur la forme de la distribution.

## Paramètres de dispersion

- 1) Si les moustaches sont à peu près de la même longueur, c'est que la distribution est symétrique.
- 2) Si la moustache de droite est plus longue que la moustache de gauche, c'est que la distribution est étalée à droite.
- 3) Si la moustache de gauche est plus longue que la moustache de droite, c'est que la distribution est étalée à gauche.

# Paramètres de dispersion

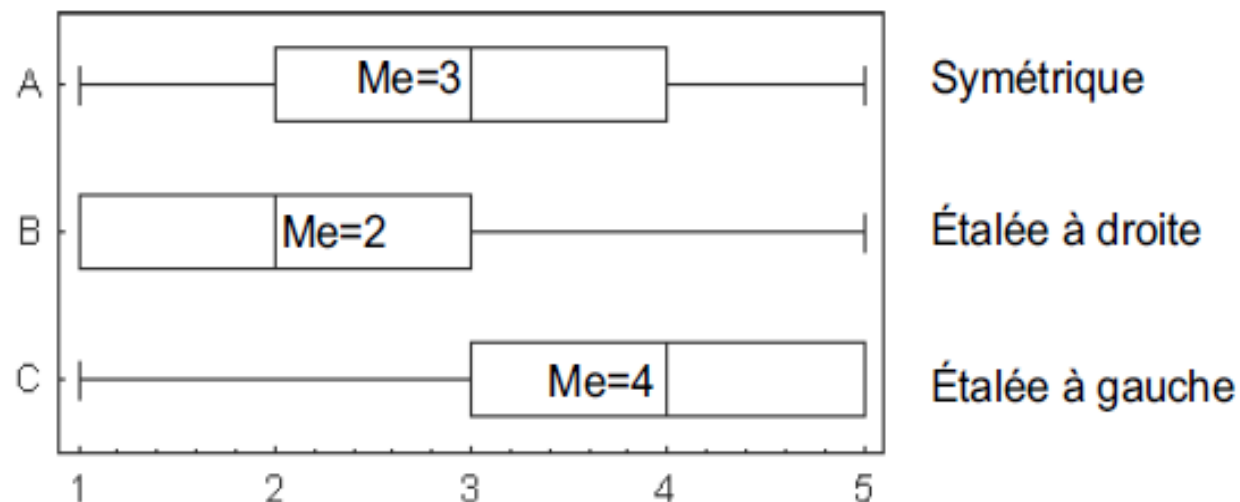
Soit les trois séries utilisées, dont les distributions

$A = \{1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5\}$

$B = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5\}$

$C = \{1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5\}$

Les boîtes à moustaches correspondantes



## Paramètres de dispersion

La variance, l'écart-type et le coefficient de variation sont les indicateurs les plus fréquemment utilisés pour mesurer la dispersion d'une série. Ces indicateurs renseignent sur la **dispersion des données autour de la moyenne**.

Plus les données sont concentrées autour de la moyenne, plus les valeurs de ces trois indicateurs sont faibles. Inversement, plus les données sont dispersées autour de la moyenne, plus ces trois indicateurs sont élevés.



# Paramètres de dispersion

## 6. La variance

› **Définition:** La variance notée  $Var$  est la somme pondérée des carrés des écarts des valeurs de la série à la moyenne d'une série statistique quantitative

$$\begin{aligned} Var &= \frac{1}{N} \sum_{i=1}^p n_i (x_i - \mu)^2 \\ &= \sum_{i=1}^p \frac{n_i}{N} (x_i - \mu)^2 \\ &= \sum_{i=1}^p f_i (x_i - \mu)^2, \end{aligned}$$

Car  $\frac{n_i}{N} = f_i$

Où les  $x_i$  représentent les valeurs observées par les distributions non groupées et les centres des classes pour les distributions groupées.

# Paramètres de dispersion

Propriété:

$$\text{Var}X \geq 0$$

Formule pratique pour la variance

Proposition

$$\begin{aligned}\text{Var}X &= \frac{1}{N} \sum_{i=1}^p n_i x_i^2 - \mu^2 \\ &= \sum_{i=1}^p f_i x_i^2 - \mu^2.\end{aligned}$$

Preuve: exercice.

# Paramètres de dispersion

- La variance est nulle si, et seulement si,  $X$  possède une seule valeur.
- Soit la série  $X$  dont la variance est  $Var(X)$ ; définissons une nouvelle série  $Y$  tel que  $Y = aX + b$  avec  $a$  et  $b$  deux constantes, alors la variance de  $Y$  notée  $Var(Y)$

$$Var(Y) = a^2 Var(X)$$

# Paramètres de dispersion

## 7. L'écart-type

**Définition** On appelle écart-type, le nombre,

$$\sigma = \sqrt{Var}$$

La signification de l' écart-type et de la variance est très simple: Plus les valeurs observées sont homogènes, plus ces deux quantités sont très petites. Inversement, plus les valeurs observées sont hétérogènes plus ces deux quantités sont très grandes.

Autrement Une série peu dispersée (ayant des valeurs regroupées autour de la valeur moyenne) aura un écart-type plutôt faible.

# Paramètres de dispersion

**Exemple:** La variance dans l'exemple 34 est donnée par:

$$Var = \frac{1}{29} (8^2 \times 3 + 9^2 \times 2 + \dots + 16^2 \times 1) - (11,25)^2$$

$$= \frac{2631}{20} - 126,56 \cong 131,56 - 126,56 = 4,99$$

L'écart-type vaut  $\sigma \cong 2,233$

La variance dans l'exemple 35 est donnée par:

$$Var = \frac{10^6}{220} (7^2 \times 40 + 9^2 \times 37 + \dots + 23^2 \times 17) - 13045^2$$

$$= \frac{10^6}{220} \times 43452 - 170172025 \cong 27337066$$

L'écart-type vaut  $\sigma \cong 5228,5$

# Paramètres de dispersion

## Propriétés

**Propriétés 1:** Si la série  $(x_i, n_i)$  a pour écart-type  $\sigma$ ; alors l'écart-type de la série  $(x_i + a, n_i)$  est aussi  $\sigma$ ,  $a$  étant un terme constant.

**Propriétés 2:** Si la série  $(x_i, n_i)$  a pour écart-type  $\sigma$ ; alors l'écart-type de la série  $(a.x_i, n_i)$  est  $a.\sigma$ ,  $a$  étant un terme constant

# Paramètres de dispersion

Même si l'écart-type est le plus utilisé et le plus performant des paramètres de dispersion. Il amplifie le poids des valeurs extrêmes ou anormales de la série étudiée.

## Les paramètres de dispersion relative

La moyenne arithmétique, l'écart-type, la médiane d'une série statistique quantitative  $X$  se mesurent avec la même unité que  $X$ . Si on veut comparer des dispersion de séries statistiques se mesurant avec la même unité, mais ayant des ordres de grandeurs différents, par exemple le débit d'un grand fleuve, et celui d'une petite rivière, on est obligé d'utiliser des nombres sans dimension, pour que l'unité de mesure disparaisse. La série doit prendre des valeurs positives et être exprimée en fonction d'une origine non arbitraire ( par exemple, pour des températures il faudrait les rapporter au zéro absolu)

# Paramètres de dispersion

**Les paramètres de dispersion relative:** Ces paramètres relativisent la mesure de la dispersion par rapport à une valeur centrale de la série. Ils s'expriment habituellement en pourcentage.

## a. Le coefficient de variation de Pearson

**Définition:** Le coefficient de variation de Pearson permet de relativiser l'écart-type par rapport à la moyenne arithmétique et rend plus aisée la comparaison entre séries de nature différente. Plus le coefficient de variation noté  $Cv$  est élevé, plus la dispersion relative est forte.

$$Cv = \frac{\sigma}{\mu} (en \%) \times 100$$



# Paramètres de dispersion

## Avantages

L'écart-type seul ne permet le plus souvent pas de juger de la dispersion des valeurs autour de la moyenne. Si par exemple une distribution a une moyenne de 10 et un écart-type de 1 (CV de 10 %), elle sera beaucoup plus dispersée qu'une distribution de moyenne 1000 et d'écart-type 10 (CV de 1 %).

Ce nombre est sans unité, c'est une des raisons pour lesquelles il est parfois préféré à l'écart type qui lui ne l'est pas. En effet, pour comparer deux séries de données d'unités différentes, l'utilisation du coefficient de variation est plus judicieuse.

# Paramètres de dispersion

## Inconvénient

Quand la moyenne est proche de zéro, le coefficient de variation va tendre vers l'infini et sera par conséquent très sensible aux légères variations de la moyenne.

Contrairement à l'écart type, le coefficient de variation ne peut être utilisé directement pour construire un intervalle de confiance autour de la moyenne.

**Exemple:** Le tableau suivant présente le Débit moyen de certains fleuves:

# Paramètres de dispersion

Mois	La Seine à Paris	Le Rhône à Genève	Le Rhône à Beaucaire	L'Hérault au moulin de Bertrand
Janvier	510	163	2296	32
Février	545	175	2050	26
Mars	445	177	2280	46
Avril	232	180	1673	31
Mai	229	223	1968	25
Juin	157	370	1558	10
Juillet	112	413	1230	5
Aout	94	379	1148	4
Septembre	99	269	1427	15
Octobre	124	197	1066	32
Novembre	244	18	1591	40
Décembre	309	166	1378	27

# Paramètres de dispersion

Le calcul de coefficient de variation est présenté dans le tableau suivant

	La Seine à Paris	Le Rhône à Genève	Le Rhône à Beaucaire	L'Hérault au moulin de Bertrand
Moyenne	265,83	241	1638,75	24,42
Ecart-type	155,25	89,40	406,15	12,80
Coefficient de variation	0,58	0,37	0,25	0,52

Malgré des ordres de grandeur différents, on peut comparer les dispersions de ces séries, et constater par exemple que la Seine à Paris est encore moins raisonnable que L'Hérault

# Paramètres de dispersion

Propriétés:

**Propriétés 1:** Si la série  $(x_i, n_i)$  a de moyenne  $\mu$  et d'écart-type  $\sigma$ ; et la série  $(a.x_i, n_i)$  a de moyenne  $\mu'$  et d'écart-type  $\sigma'$ , *et*  $a$  étant un terme constant alors les deux séries ont même coefficient de variation:

$$Cv = \frac{\sigma'}{\mu'} = \frac{a\sigma}{a\mu} = \frac{\sigma}{\mu}$$

**Propriétés 2:** Par contre les séries  $(x_i, n_i)$  et  $(x_i + a, n_i)$  présentent des coefficients de variation différents

# Paramètres de dispersion

## b. Le coefficient interquartile relatif

**Définition:** Etant donné une série statistique quantitative  $X$  admettant  $Me$  pour médiane, pour  $Q_1$  premier quartile et  $Q_3$  pour troisième quartile, on a coefficient interquartile le rapport  $\frac{Q_3 - Q_1}{Me}$

**Exemple:** Le tableau suivant donne le nombre de voyageurs au départ des principales gares du Languedoc-Roussillon en 1992 et 1995

# Paramètres de dispersion

Gares	1992	1995	Gares	1992	1995
Montpellier	1393	1646	Narbonne	331	331
Nîmes	868	869	Lunel	240	297
Béziers	483	458	Carcassonne	247	219
Perpignan	481	403	Agde	194	184
Sète	339	348	Alèse	142	116

# Paramètres de dispersion

On peut calculer les quartiles correspondants à ces deux séries, ainsi que leurs moyennes arithmétiques et leurs écart-types, le tableau suivant nous donne les paramètres des séries précédentes

Paramètres	1992	1995
$Q_1$	247	219
$Me$	335	339,5
$Q_3$	483	458
Moyenne	471,8	487,1
Ecart-type	365,2	433,4

On en déduit le coefficient interquartile pour 1992:  $\frac{483-247}{335} = 0,70448$

et le coefficient interquartile pour 1995:  $\frac{458-219}{\frac{335}{339,5}} = 0,70398$



# Paramètres de dispersion

Il semblait que la dispersion relative ait diminué.

Calculons les coefficient de variation de ces deux séries on obtient:  $\frac{365,2}{471,8} = 0,77406$

Pour 1992 et  $\frac{433,4}{339,5} = 0,88976$  pour 1995, Il semblait que la dispersion relative ait augmenté.

Ces résultats divergents s'expliquent par le fait que le coefficient de variation fait intervenir l'écart-type qui est très sensible aux valeurs extrêmes. Or le trafic a augmenté de 18% à Montpellier, qui correspond à la plus grande valeur des deux séries, alors que le trafic a stagné ou diminué dans les autres gares.

l'écart-type augmente considérablement de 1992 à 1995 alors que le coefficient interquartile a légèrement diminué .

# Paramètres de dispersion

## Les moments

**Définition:** On appelle moment d'ordre  $k$  par rapport à une origine  $x_0$ , notée  $m_k$ , la valeur:

$$m_k = \frac{1}{N} \sum_{n=1}^{n=p} n_i (x_i - x_0)^k$$

### a. Moments simples

Si  $x_0 = 0$ , les moments sont dits des moments simples

$$m_k = \frac{1}{N} \sum_{i=1}^{n=p} n_i x_i^k$$

# Paramètres de dispersion

## b. Moments centrés

Lorsque  $x_0 = \bar{x}$  on définit les moments centrés d'ordre  $k$  ils sont notés  $\mu_k$

$$\mu_k = \frac{1}{N} \sum_{i=1}^{i=p} n_i (x_i - \bar{x})^k$$

Cas particulier

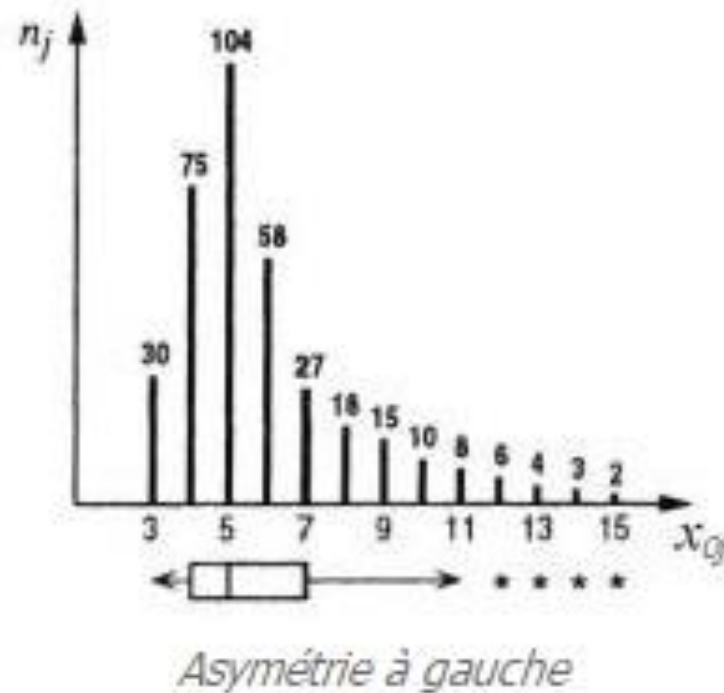
- Si  $k = 0 \Rightarrow \mu_0 = 1$
- Si  $k = 1 \Rightarrow \mu_1 = 0$
- Si  $k = 2 \Rightarrow \mu_2 = \text{var}(X)$

# Caractéristique de forme et de concentration

Souvent, l'analyse du diagramme en bâtons – ou de l'histogramme dans le cas d'une distribution permet de se rendre compte du caractère symétrique ou non d'une distribution. L'examen de la boîte à moustaches permet aussi de se faire une idée sur cette question selon que la boîte et les moustaches sont symétriques ou, au contraire, de plus petite amplitude à gauche (asymétrie à gauche) ou à droite (asymétrie à droite).

Ainsi, par exemple, le diagramme en bâtons et la boîte à moustaches ci-dessous permettent de se rendre compte aisément que la distribution observée présente une asymétrie gauche, c'est-à-dire que les petites valeurs observées sont plus fréquentes que les valeurs plus élevées.

# Caractéristique de forme et de concentration



# Caractéristique de forme et de concentration

Mais il est également possible de caractériser l'asymétrie et d'en quantifier l'importance via l'un ou l'autre **coefficient d'asymétrie**

## Le coefficient de Fisher

Le coefficient d'asymétrie de Fisher, noté  $g_1$ , se définit comme étant le rapport entre le moment centré d'ordre 3 ( $\mu_3$ ) et le cube de l'écart-type:

$$g_1 = \frac{\mu_3}{\sigma^3}$$

- Si  $g_1 = 0$ , la série est symétrique.
- Si  $g_1 < 0$ , la série est étalée vers la gauche (la série est oblique à droite)
- Si  $g_1 > 0$ , la série est étalée vers la droite (la série est oblique à gauche)

# Caractéristique de forme et de concentration

**Exemple:** Considérons la distribution de taille  $N=360$  ci-dessous avec les tailles mesurées tantôt en mètre tableau de gauche tantôt en centimètre tableau de droite

# Caractéristique de forme et de concentration

Tailles en mètres

$x_{oj}$	$n_j$	$n_j x_{oj}$	$n_j(x_{oj} - \bar{x})^2$	$n_j(x_{oj} - \bar{x})^3$
1,10	25	27,50	0,5625	-0,084375
1,15	75	86,25	0,75	-0,075
1,20	103	123,60	0,2575	-0,012875
1,25	60	75,00	0	0
1,30	26	33,80	0,065	0,00325
1,35	18	24,30	0,18	0,018
1,40	15	21,00	0,3375	0,050625
1,45	12	17,40	0,48	0,096
1,50	8	12,00	0,5	0,125
1,55	6	9,30	0,54	0,162
1,60	4	6,40	0,49	0,1715
1,65	4	6,60	0,64	0,256
1,70	3	5,10	0,6075	0,273375
1,75	1	1,75	0,25	0,125
360		450	5,66	1,1085

$$\bar{x} = 450/360 = 1,25 \text{ (m)}$$

$$s^2 = 5,66/360 = 0,0157 \text{ (m}^2\text{)}$$

$$s = 0,1254 \text{ (m)}$$

$$m_3 = 1,1085/360 = 0,0031 \text{ (m}^3\text{)}$$

$$g_1 = 1,5619$$

Tailles en centimètres

$x_{oj}$	$n_j$	$n_j x_{oj}$	$n_j(x_{oj} - \bar{x})^2$	$n_j(x_{oj} - \bar{x})^3$
110	25	2750	5625	-84375
115	75	8625	7500	-75000
120	103	12360	2575	-12875
125	60	7500	0	0
130	26	3380	650	3250
135	18	2430	1800	18000
140	15	2100	3375	50625
145	12	1740	4800	96000
150	8	1200	5000	125000
155	6	930	5400	162000
160	4	640	4900	171500
165	4	660	6400	256000
170	3	510	6075	273375
175	1	175	2500	125000
360		45000	56600	1108500

$$\bar{x} = 45000/360 = 125 \text{ (cm)}$$

$$s^2 = 56600/360 = 157,2222 \text{ (cm}^2\text{)}$$

$$s = 12,5388 \text{ (cm)}$$

$$m_3 = 1108500/360 = 3079,1667 \text{ (cm}^3\text{)}$$

$$g_1 = 1,5619$$



# Caractéristique de forme et de concentration

Que les tailles soient mesurées en mètres ou en centimètres, le coefficient de Fisher a toujours la même valeur positive:  $g_1 = 1,5619$  et puisque  $g_1 > 0$

la série est étalée vers la droite (la série est oblique à gauche) c'est ce qu'on appelle asymétrie à gauche.

## Les coefficients empiriques

Il existe d'autres coefficients d'asymétrie plus rapides à calculer que  $g_1$ , mais dont les propriétés résultent de constatations empiriques.

## Le coefficient empirique de Pearson

**Définition:** Le coefficient empirique de Pearson  $P_1$  se fonde sur l'écart entre la moyenne  $\bar{x}$  et le mode  $Mo$  de la distribution observée. Cet écart est divisé par l'écart-type  $\sigma$  de telle sorte que  $P_1$  soit un nombre sans unité :

$$P_1 = \frac{\bar{x} - Mo}{\sigma}$$

# Caractéristique de forme et de concentration

**Remarque:**  $P_1$  possède des propriétés semblables à celles de  $g_1$ . En effet

- pour une distribution symétrique  $\bar{x} = Mo$  et donc  $P_1 = 0$
- pour une distribution dissymétrique à gauche :  $\bar{x} > Mo$  et donc  $P_1 > 0$
- pour une distribution dissymétrique à droite :  $\bar{x} < Mo$  et donc  $P_1 < 0$

## Le coefficient empirique de Yule

**Définition:** Le coefficient empirique de Yule et Kendall se définit à partir des trois quartiles de la distribution observée :

$$Y = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1}$$

# Caractéristique de forme et de concentration

**Remarque:**  $Y$  possède des propriétés semblables à celles de  $g_1$ . En effet, outre le fait d'être lui aussi un nombre sans dimension, on vérifie aisément que:

- pour une distribution symétrique :  $Y = 0$ .
- pour une distribution dissymétrique à gauche :  $Y > 0$ .
- pour une distribution dissymétrique à droite :  $Y < 0$ .

On peut également vérifier que  $-1 < Y < 1$

# Statistique descriptive à deux variable

# Introduction

Dans certaines enquêtes, on pose plusieurs questions. Par exemple lors d'un recensement général de la population, on note pour chaque personne interrogé, son sexe, son âge, son caractère marital (célibataire, marié, veuf, divorcé), le nombre de pièces de son habitation, les éléments du confort du logement etc. Chaque question correspond à un caractère statistique de la population étudiée. Chaque caractère peut être étudié en soi, mais on s'intéresse également aux liens qui peuvent exister entre les variables statistiques

# Introduction

On souhaite réaliser une étude statistique sur une population de nouveau nés.  
L'étude pourra alors porter sur:

- Un caractère: la masse en kg

ou

- Un caractère: la taille en cm

ou

- deux caractères: la masse et la taille à la fois

Dans le dernier cas, les résultats seront présentés sous la forme d'un tableau à double entrée

# Introduction

Enfants	1	2	3	4	5	6	7	8	9	10
Masse en kg	2,4	2,6	2,7	3	3,2	3,3	3,6	3,7	3,8	4
Taille en cm	45	47	48	50	51	52	53	54	54	56

**Définition:** Une série statistique à deux variables est une série statistique pour laquelle deux caractères sont relevés pour chaque individu.

L'étude de deux caractères statistiques à la fois a pour but de déterminer s'il existe un lien entre les deux.

# Distribution *Conjointe*

Soient  $X$  et  $Y$  deux séries statistiques quantitatives discrètes ou continues définies sur la même population  $\Omega$  de taille  $N$  (resp qualitatifs ou l'une quantitatif et l'autre qualitatif)  $X(\Omega) = \{x_1, x_2, \dots, x_p\}$ ;  $Y(\Omega) = \{y_1, y_2, \dots, y_q\}$ .

Le couple  $(X, Y)$  s'appelle série statistique double (ou *distribution conjointe*).

Nous disposons cette fois de  $pq$  couples  $(x_i, y_i)$  de valeurs observées,  
 $i = 1, 2, \dots, p$  ;  $j = 1, 2, \dots, q$ .

La 1<sup>ère</sup> valeur de chaque couple se rapporte à la variable statistique  $X$ ;

La 2<sup>ème</sup> valeur de chaque couple se rapporte à la variable statistique  $Y$ ;

## Remarques:

- Si  $X$  est qualitatif,  $x_i$  représente la modalité numéro  $i$  de ce caractère.
- Si  $X$  est quantitatif  $x_i$  représente la  $i^{\text{ème}}$  valeur de ce caractère ou le centre de  $i^{\text{ème}}$  classe. De même pour  $Y$



## Distribution *Conjointe*

**Exemple :** supposons que l'on ait les données suivantes sur le sexe et le statut d'activité de 20 personnes. Les données sont présentées par paire. La première concerne le sexe avec deux modalités et la deuxième ~~La première~~ concerne le statut d'activité, avec trois modalités (actif occupé [AO], chômeur [C], inactif [I]).

{F ; AO} ; {M ; I} ; {F ; C} ; {F ; C} ; {M ; AO} ; {M ; AO} ; {M ; C} ; {F ; I} ; {F ; I} ; {F ; I} ; {M ; C} ;  
{F ; AO} ; {F ; AO} ; {F ; AO} ; {M ; AO} ; {M ; C} ; {M ; AO} ; {F ; I} ; {F , C} ; {M , AO}

Regroupons ces données dans un tableau de contingence

Sexe \ Statut	Actifs occupés	Chômeurs	Inactifs
Masculin	5	3	1
Féminin	4	3	4

# Distribution *Conjointe*

**Définition:** L'**effectif partiel** d'un couple  $(x_i, y_j)$  est le nombre  $n_{ij}$  de couple observés égaux à  $(x_i, y_j)$ .

La **fréquence partielle** du couple  $(x_i, y_j)$  est le rapport

$$f_{ij} = \frac{n_{ij}}{N}; \quad i = 1, 2, \dots, p; \quad j = 1, 2, \dots, q.$$

La distribution des effectifs partiels est un tableau à double entrée appelé **tableau de contingence**

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_q$	Total
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1q}$	$n_{1.}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2q}$	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$x_p$	$n_{p1}$	$n_{p2}$	$\dots$	$n_{pq}$	$n_{p.}$
Total	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.q}$	$\sum_{i=1}^p n_{i.} = \sum_{j=1}^q n_{.j} = N$

## Distribution *Conjointe*

L'effectif  $n_{ij}$  représente le nombre d'individus qui ont à la fois la modalité/valeur  $x_i$  et la modalité/valeur  $y_j$ . On a ensuite les symboles suivants :

- $n_{22}$  effectif des individus qui ont la modalité/valeur 2 de  $X$  et la modalité 2 de  $Y$ .

Par convention, on note toujours la modalité/valeur de  $X$  ( $i$ ) avant celle de  $Y$  ( $j$ ).

- $n_{2q}$  : effectif des individus qui ont la modalité/valeur 2 de  $X$  et la modalité  $q$  de  $Y$
- $n_{pq}$  : effectif des individus qui ont la modalité/valeur  $p$  de  $X$  et la modalité  $q$  de  $Y$

$n_{i.}$  effectif des individus qui ont la modalité/valeur  $i$  (le « . » à la place du  $j$  signifie que l'on ne tient pas compte de  $Y$ ).

**Exemple** :  $n_{1.}$  désigne tout l'effectif des individus qui ont la modalité/valeur 1 de  $X$ .

## Distribution *Conjointe*

$n_{.j}$  effectif des individus qui ont la modalité  $j$  (le « . » à la place du  $i$  signifie que l'on ne tient pas compte de  $X$ ). **Exemple** :  $n_{.1}$  désigne **tout** l'effectif des individus qui ont la modalité/valeur 1 de  $Y$ .  $n_{..}$  : effectif total on le note  $N$ . Dès lors:

$$n_{i.} = \sum_{j=1}^q n_{ij} = n_{i1} + n_{i2} + \cdots + n_{iq}$$

$$n_{.j} = \sum_{i=1}^p n_{ij} = n_{1j} + n_{2j} + \cdots + n_{pj}$$

$$n_{..} = \sum_{i=1}^p n_{i.} = \sum_{i=1}^p \left( \sum_{j=1}^q n_{ij} \right) = \sum_{j=1}^q n_{.j} = \sum_{j=1}^q \left( \sum_{i=1}^p n_{ij} \right)$$

# Distributions marginales

Les  $p$  couples  $(x_i, n_{i.})$  forment la *distribution marginale* de la *variable*  $X$ .

Les  $q$  couples  $(y_j, n_{.j})$  forment la *distribution marginale* de la *variable*  $Y$ .

Les distributions marginales peuvent aussi être données sous forme de fréquences

$$f_{i.} = \frac{n_{i.}}{N}, \quad i = 1, 2, \dots, p \qquad f_{.j} = \frac{n_{.j}}{N}, \quad j = 1, 2, \dots, q$$

et Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale, de dispersion, de forme...

# Distributions marginales:

L'**effectif marginal** de la valeur  $x_i$  (*resp*  $y_j$ ) est la somme des effectifs partiels des couples contenant  $x_i$  (*resp*  $y_j$ ) c'à d la quantité

$$n_{i.} = \sum_{j=1}^q n_{ij}, \quad i = 1, 2, \dots, p, \quad \left( \text{resp } n_{.j} = \sum_{i=1}^p n_{ij}, \quad j = 1, 2, \dots, q, \right)$$

La **fréquence marginale** de la valeur  $x_i$  (*resp*  $y_j$ ) est la somme des fréquences partielles des couples contenant  $x_i$  (*resp*  $y_j$ ) c'à d la quantité

$$f_{i.} = \sum_{j=1}^q f_{ij} = \sum_{j=1}^q \frac{n_{ij}}{N} = \frac{n_{i.}}{N}, \quad i = 1, 2, \dots, p$$

$$\left( \text{resp } f_{.j} = \sum_{i=1}^p f_{ij} = \sum_{i=1}^p \frac{n_{ij}}{N} = \frac{n_{.j}}{N}, \quad j = 1, 2, \dots, q \right)$$

# Distributions marginales:

La distribution des effectifs marginaux ou des fréquences marginales de la 1<sup>ère</sup> (*resp* la 2<sup>ème</sup>) variable est La distribution à une dimension où chaque valeur  $x_i$  (*resp*  $y_i$ ) est associé son effectif marginal ou sa fréquence marginale.

Remarque

$$\sum_{i=1}^p \sum_{j=1}^q n_{ij} = \sum_{j=1}^q \sum_{i=1}^p n_{ij} = N,$$

$$\sum_{i=1}^p \sum_{j=1}^q f_{ij} = \sum_{j=1}^q \sum_{i=1}^p f_{ij} = \sum_{i=1}^p f_{i\cdot} = \sum_{j=1}^q f_{\cdot j} = 1$$

# Exemples

**Exemple 1:** l'étude statistique suivante porte une population de 100 ménages deux caractères sont étudiés:

$X$ : le nombre d'enfants;  $X(\Omega) = \{0, 1, 2, 3, 4, 5\}$ ,

$Y$ : le nombre de pièces de l'appartement occupé;  $Y(\Omega) = \{1, 2, 3, 4\}$ .

Dans cet exemple, nous avons  $p = 6$ ,  $q = 4$  et  $\text{card}(\Omega) = N = 100$

$X \setminus Y$	1	2	3	4	Total
0	6	3	1	0	$n_{1.} = 10$
1	4	11	3	1	$n_{2.} = 19$
2	1	10	16	3	$n_{3.} = 30$
3	0	5	13	5	$n_{4.} = 23$
4	0	1	4	8	$n_{5.} = 13$
5	0	0	1	4	$n_{6.} = 5$
Total	$n_{.1} = 11$	$n_{.2} = 30$	$n_{.3} = 38$	$n_{.4} = 21$	100



# Exemples

Par exemple,

L'effectif partiel  $n_{12}$  de (0,2) est égal à 3, la fréquence partielle  $f_{12}$  de (0,2) est égal à  $\frac{3}{100} = 0,03$ .

L'effectif partiel  $n_{33}$  de (2,3) est égal à 16, la fréquence partielle  $f_{33}$  de (2,3) est égal à  $\frac{16}{100} = 0,16$ .

On vérifie que

$$\sum_{i=1}^6 n_{i.} = \sum_{j=1}^4 n_{.j} = N = 100$$

La dernière colonne donne la distribution des effectifs marginaux du nombre d'enfants (càd de la variable  $X$ )

# Exemples

La dernière ligne donne la distribution des effectifs marginaux du nombre pièces (càd de la variable  $Y$ ).

En divisant ces chiffres par 100, on obtient la distribution des fréquences marginales des variables  $X$  et  $Y$ .

Distribution marginale de  $X$

$X$	0	1	2	3	4	5
Effectif	10	19	30	23	13	5
Fréquence	0,10	0,19	0,30	0,23	0,13	0,05

# Exemples

Distribution marginale de  $Y$

$Y$	1	2	3	4
Effectif	11	30	38	21
Fréquence	0,11	0,3	0,38	0,21

# Groupement des données

On peut, comme dans le cas d'une variable statistique à une dimension, grouper les valeurs d'une ou des deux variables statistiques en classes. Les effectifs partiels et les fréquences partielles définis précédemment se rapportent alors aux classes plutôt qu'aux valeurs, comme pour les variables statistiques à une dimension.

**Exemple 2:** On a mesuré le poids et la taille de 1000 individus. Le tableau ci-dessous donne la distribution des effectifs partiels (après groupement des

# Groupement des données

$X \backslash Y$	[65, 70[	[70, 75[	[75, 80[	[80, 85[	[85, 90[	Total
[160, 165[	50	48	5	4	3	$n_{1.} = 110$
[165, 170[	80	140	50	25	5	$n_{2.} = 300$
[170, 175[	60	150	90	25	2	$n_{3.} = 327$
[175, 180[	5	40	90	40	5	$n_{4.} = 180$
[180, 185[	13	10	25	20	15	$n_{5.} = 83$
Total	$n_{.1} = 208$	$n_{.2} = 388$	$n_{.3} = 260$	$n_{.4} = 114$	$n_{.5} = 30$	$N = 1000$

La dernière colonne donne la distribution des effectifs marginaux de la variable  $X = \text{taille}$ .  
La dernière ligne donne la distribution des effectifs marginaux de la variable  $Y = \text{poids}$ .

# Groupement des données

**Définitions:** A partir de la distribution conjointe de  $(X, Y)$  on peut déduire la distribution de  $X$  seul et  $Y$  seul.

la distribution marginale de  $X$  est donnée par:

$$\{(x_i, n_{i.})\} \quad \forall i = 1, 2, \dots, p$$

la distribution marginale de  $Y$  est donnée par:

$$\{(y_i, n_{.j})\} \quad \forall j = 1, 2, \dots, q$$

**Exemple:** dans l'exemple 2: Les tableaux statistiques des distributions  $X$  et  $Y$  sont

# Grouperment des données

$X$	Effectifs
[160, 165[	110
[165, 170[	300
[170, 175[	327
[175, 180[	180
[180, 185[	83
Total	1000

$Y$	Effectifs
[60, 70[	208
[70, 75[	388
[75, 80[	260
[80, 85[	114
[85, 90[	30
Total	1000

# Effectifs cumulés

**Définition:** On appelle **effectif partiel cumulé** du couple de valeurs  $(x_i, y_j)$  le nombre d'individus  $N_{ij}$  tel que:

$$N_{ij} = \sum_{s \leq i, k \leq j} n_{sk}$$

**Exemple:** Dans l'exemple 2

$$N_{22} = \sum_{s \leq 2, k \leq 2} n_{sk} = n_{11} + n_{12} + n_{21} + n_{22} = 50 + 48 + 80 + 140 = 318$$

**Définition:** On appelle **effectif marginal cumulé** le nombre  $N_{.j} = \sum_{t=1}^j n_{.t}$  <sup>ou</sup> ~~ou~~  $N_{i.} = \sum_{t=1}^i n_{t.}$

**Exemple:** Dans l'exemple 2:  $N_{3.}$  est la somme de tous les effectifs marginales des couples  $(x_i, y_j)$  pour lesquels  $i \leq 3$

$$N_{3.} = \sum_{t=1}^{t=3} n_{t.} = n_{1.} + n_{2.} + n_{3.} = 110 + 300 + 327 = 737$$



# fréquences cumulées

**Définition:** On appelle **fréquence partielle cumulée** du couple de valeurs  $(x_i, y_j)$  le

nombre d'individus  $F_{ij}$  tel que:

$$F_{ij} = \frac{N_{ij}}{N}$$

$$F_{ij} = \sum_{s \leq i, k \leq j} f_{sk}$$

**Exemple:** Dans l'exemple 2

$$F_{22} = \sum_{s \leq 2, k \leq 2} f_{sk} = f_{11} + f_{12} + f_{21} + f_{22} = 0,050 + 0,048 + 0,080 + 0,140 = 0,318$$

**Définition:** On appelle **fréquence marginale cumulée** le nombre  $F_{.j} = \sum_{t=1}^j f_{.t}$  où  $F_{i.} = \sum_{t=1}^i f_{it}$ .

**Exemple:** Dans l'exemple 2:  $F_{3.}$  est la somme de tous les ~~effectifs~~ marginales des couples  $(x_i, y_j)$  pour lesquels  $i \leq 3$   
fréquences

$$F_{3.} = \sum_{t=1}^{t=3} f_{.t} = f_{1.} + f_{2.} + f_{3.} = 0,110 + 0,300 + 0,327 = 0,737$$

# Valeurs typiques d'une distribution à deux variables

## Valeurs typiques des distributions marginales

la distribution des effectifs marginaux sont des distributions à une dimension. On peut donc leur associer les valeurs typiques examinées au chapitre précédent.

Soient  $X$  et  $Y$  deux variables statistiques définies sur la même population  $\Omega$  de taille  $N$ .

$$X(\Omega) = \{x_1, x_2, \dots, x_p\}; \quad Y(\Omega) = \{y_1, y_2, \dots, y_q\}.$$

## Moyennes marginales de $X$ et $Y$

**Définition:** La **moyenne marginale** de  $X$  est le nombre  $\bar{X}$  défini par:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^p n_{i.} x_i = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} x_i$$

# Valeurs typiques d'une distribution à deux variables

La **moyenne marginale** de  $Y$  est le nombre  $\bar{Y}$  défini par:

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^q n_{.j} y_j = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} y_j$$

Variances marginales de  $X$  et  $Y$

**Définition:** La **variance marginale** de  $X$  est la quantité définie par:

$$Var(X) = \frac{1}{N} \sum_{i=1}^p n_{i.} (x_i - \bar{X})^2 = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (x_i - \bar{X})^2$$

$\sqrt{Var(X)}$  est l'écart-type marginal de  $X$

# Valeurs typiques d'une distribution à deux variables

La **variance marginale** de  $Y$  est la quantité définie par:

$$\text{Var}(Y) = \frac{1}{N} \sum_{j=1}^q n_{.j} (y_j - \bar{Y})^2 = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (y_j - \bar{Y})^2$$

$\sqrt{\text{Var}(Y)}$  est l'écart-type marginal de  $Y$

**Proposition (*Formule de Koenig*)**

$$\text{Var}(X) = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} x_i^2 - (\bar{X})^2 = \frac{1}{N} \sum_{i=1}^p n_{i.} x_i^2 - (\bar{X})^2,$$

$$\text{Var}(Y) = \frac{1}{N} \sum_{j=1}^q \sum_{i=1}^p n_{ij} y_j^2 - (\bar{Y})^2 = \frac{1}{N} \sum_{j=1}^q n_{.j} y_j^2 - (\bar{Y})^2$$

# Valeurs typiques d'une distribution à deux variables

## Remarque

- Lorsque les données sont groupées en classe  $x_i$  et  $y_i$  représentent les centres de ces classes.
- Pour calculer ses valeurs typiques, on ajoute aux tableaux de contingence des lignes ou des colonnes comportant les quantités  $n_{i.}x_i$ ,  $n_{.j}y_j$ ,  $n_{i.}x_i^2$  et  $n_{.j}y_j^2$ .

**Exemple 3:** Moyennes et variances marginales de la taille et du poids dans l'exemple 2

# Valeurs typiques d'une distribution à deux variables

$X$	Centre des classes $x_i$	Effectif marginal $n_{i.}$	$n_{i.} \cdot x_i$	$n_{i.} \cdot x_i^2$
[160, 165[	162,5	110	17875	2904687,5
[165, 170[	167,5	300	50250	8416875
[170, 175[	172,5	327	56407,5	9730293,75
[175, 180[	177,5	180	31950	5671125
[180, 185[	182,5	83	15147,5	2764418,75
Total		1000	171630	29487400

# Valeurs typiques d'une distribution à deux variables

$Y$	Centre des Classes $\frac{y_j}{n_{.j}}$	Effectifs marginal $n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
[65, 70[	67,5	208	14040	947700
[70, 75[	72,5	388	28130	2039425
[75, 80[	77,5	260	20150	1561625
[80, 85[	82,5	114	9405	775912,5
[85, 90[	87,5	30	2625	229687,5
Total		1000	74350	5554350

# Valeurs typiques d'une distribution à deux variables

D'où,

$$\bar{X} = \frac{1}{N} \sum_{i=1}^5 n_{i.} x_i = \frac{171630}{1000} = 171,63; \quad \bar{Y} = \frac{1}{N} \sum_{j=1}^5 n_{.j} y_j = \frac{74350}{1000} = 74,35;$$

$$Var(X) = \frac{1}{N} \sum_{i=1}^5 n_{i.} x_i^2 - (\bar{X})^2 = \frac{29487400}{1000} - (171,63)^2 = 29487,4 - 29456,9 = 30,54$$

$$Var(Y) = \frac{1}{N} \sum_{j=1}^5 n_{.j} y_j^2 - (\bar{Y})^2 = \frac{5554350}{1000} - (74,35)^2 = 26,4$$

$$\sqrt{Var(X)} = 5,53 \quad ; \quad \sqrt{Var(Y)} = 5,14.$$



# Distribution conditionnelle

**Définition:** Etant donné une variable  $(X; Y)$  de dimension deux, on appelle variable  $X$  Conditionné à  $Y = y_j$ , la variable qui prend toutes les valeurs  $x_i, \forall i = 1, \dots, p$

avec effectif  $n_{ij}$ . On la note par  $X/Y = y_j$

$(X/Y = y_j) = \{(x_i, n_{ij})\} \forall i = 1, \dots, p$ ; Le tableau statistique de cette distribution est

$X/Y = y_j$	Effectif	fréquence
$x_1$	$n_{1j}$	$f_{1/j} = \frac{n_{1j}}{n_{.j}}$
$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{ij}$	$f_{i/j} = \frac{n_{ij}}{n_{.j}}$
$\vdots$	$\vdots$	$\vdots$
$x_p$	$n_{pj}$	$f_{p/j} = \frac{n_{pj}}{n_{.j}}$
Total	$n_{.j}$	1

# Distribution conditionnelle

$$\text{Où } n_{.j} = \sum_{i=1}^{i=p} n_{ij} \quad f_{i/j} = \frac{n_{ij}}{n_{.j}} \quad \text{et} \quad \sum_{i=1}^{i=p} f_{i/j} = \frac{n_{.j}}{n_{.j}} = 1.$$

et  $f_{i/j}$ : est la fréquence conditionnelle de  $X = x_i$  sous condition que  $Y = y_j$  càd la proportion d'individus présentant la modalité  $x_i$  parmi les individus qui présentent uniquement la modalité  $y_j$ .

De la même manière, la variable  $Y$  Conditionné à  $X = x_i$ , est la variable qui prend toutes les valeurs,  $y_j \quad \forall j = 1, \dots, q$  avec effectif  $n_{ij}$ . On la note par  $Y/X = x_i$

$$(Y/X = x_i) = \{(y_j, n_{ij})\} \quad \forall j = 1, \dots, q.$$

Le tableau statistique de cette distribution est:

# Distribution conditionnelle

$\pi$

$Y/X = x_i$	Effectif	fréquence
$y_1$	$n_{i1}$	$f_{1/i} = \frac{n_{i1}}{n_{i.}}$
$\vdots$	$\vdots$	$\vdots$
$y_j$	$n_{ij}$	$f_{j/i} = \frac{n_{ij}}{n_{i.}}$
$\vdots$	$\vdots$	$\vdots$
$y_q$	$n_{iq}$	$f_{q/i} = \frac{n_{iq}}{n_{i.}}$
Total	$n_{i.}$	1

Où  $n_{i.} = \sum_{j=1}^{j=q} n_{ij}$   $f_{j/i} = \frac{n_{ij}}{n_{i.}}$  et  $\sum_{j=1}^{j=q} f_{j/i} = \frac{n_{i.}}{n_{i.}} = 1.$

et  $f_{j/i}$  est la fréquence conditionnelle de  $Y = y_j$  sous condition que  $X = x_i$  càd la proportion d'individus présentant la modalité  $y_j$  parmi les individus qui présentent uniquement la modalité  $x_i$ .

# Distribution conditionnelle

Proposition:

$$f_{ij} = f_{i/.j} f_{.j} = f_{j/i} f_{i.}$$

Moyenne conditionnelle

Définition

- On appelle moyenne conditionnelle de  $X$  sous la condition  $Y = y_j$  et on note  $\overline{X}_j$  ou  $\overline{X}/y_j$  tel que:

$$\overline{X}_j = \frac{1}{n_{.j}} \sum_{i=1}^{i=p} n_{ij} x_i$$

- On appelle moyenne conditionnelle de  $Y$  sous la condition  $X = x_i$  et on note  $\overline{Y}_i$  ou  $\overline{Y}/x_i$  tel que:

$$\overline{Y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{j=q} n_{ij} y_j$$

# Indépendance statistique

**Définition:** On dit que les caractères  $X$  et  $Y$  sont indépendants si

$$f_{ij} = f_{i.}f_{.j} \quad \forall i = 1, 2, \dots, p, \quad \forall j = 1, 2, \dots, q$$

**Remarque:** Si  $X$  et  $Y$  sont des caractères indépendants, leur distribution conjointe s'obtient à partir de leurs distributions marginales

$$f_{ij} = f_{i.}f_{.j} \quad \forall i = 1, 2, \dots, p, \quad \forall j = 1, 2, \dots, q$$

$$n_{ij} = \frac{n_{i.}n_{.j}}{N} \quad \forall i = 1, 2, \dots, p, \quad \forall j = 1, 2, \dots, q$$

Par contre, si  $X$  et  $Y$  ne sont pas des caractères indépendants, il est impossible de trouver une distribution conjointe .

# Indépendance statistique

## Etude de la dépendance linéaire entre les deux caractères

L'étude des séries doubles a pour objectif d'étudier l'existence ou non d'une dépendance entre les deux caractères  $X$  et  $Y$ . On cherche alors une fonction mathématique qui permet de mesurer cette dépendance. Par la suite, nous allons nous intéresser uniquement au cas de l'ajustement affine.

# Nuage de points

$\pi$

**Définition:** Dans un repère orthogonal, l'ensemble des points  $M_{ij}$ , de coordonnées  $x_i$  et  $y_i$ , d'une série statistique à deux variables s'appelle nuage de points de la série statistique. Ce nuage comporte  $N$  points.

**Exemple 4:** le tableau suivant donne l'évolution du nombre d'adhérents d'un club de tennis de 2010 à 2015.

Année	2010	2011	2012	2013	2014	2015
Rang $x_i$	1	2	3	4	5	6
Nombre d'adhérents $y_i$	70	90	115	140	170	220

Le but est d'étudier cette série statistique à deux variables (le rang et le

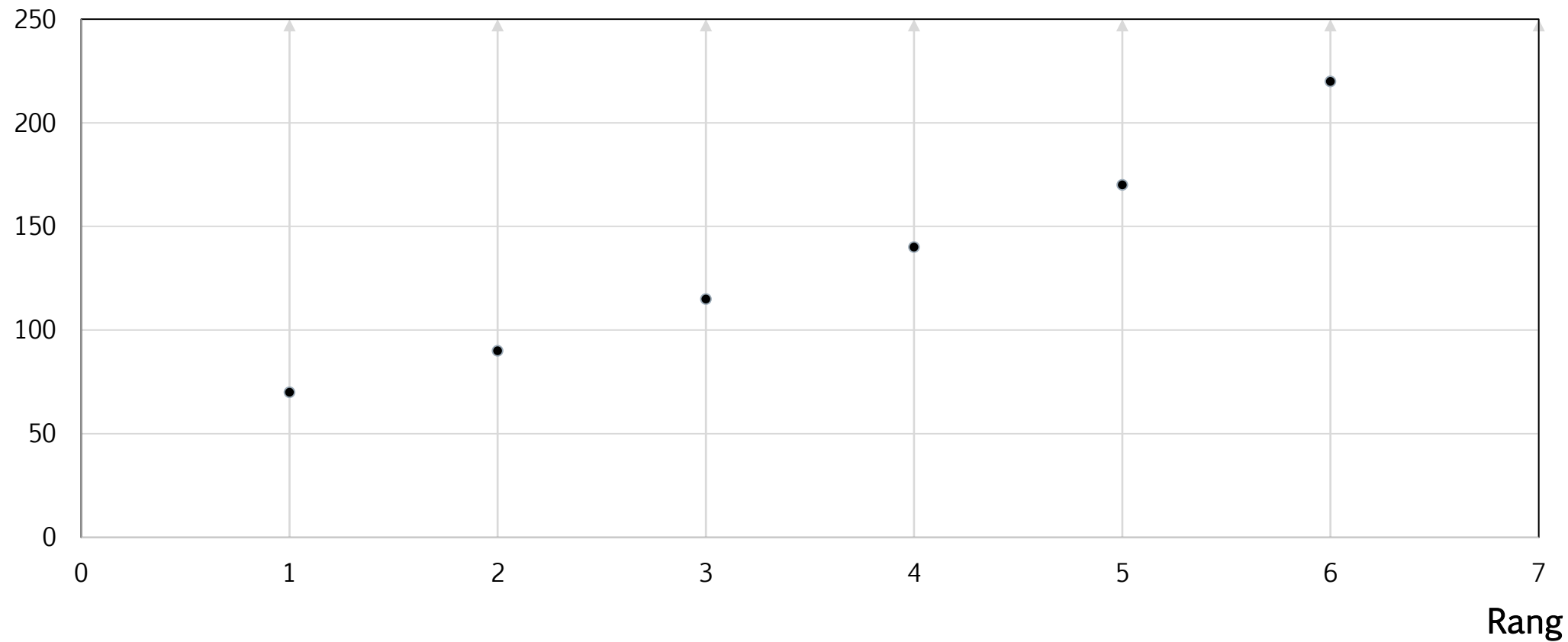
# Nuage de points

Dans cet exemple si on place le rang en abscisses, et le nombre d'adhérents en ordonnées, on peut représenter par un point chaque valeur. On obtient ainsi une succession de points, dont les coordonnées sont  $(1, 70)$ ,  $(2, 90)$ ,  $\dots$   $(6, 220)$ , forment un nuage de points.



# Nuage de points

Nombre d'adhérents



# Nuage de points

## 1. Point moyen

**Définition:** le point moyen d'un nuage de points est le point  $G$  de coordonnées  $\bar{x}$  et  $\bar{y}$  où  $\bar{x}$  (resp  $\bar{y}$ ) représente la moyenne des  $x_i$  (resp des  $y_j$ ).

**Exemple 5:** Le point moyen dans l'exemple 3 est  $G(\bar{x}, \bar{y}) = (171,63 ; 74,35)$

## 2. Covariance des variables $X$ et $Y$ (ou de la série)

**Définition:** On appelle covariance de  $X$  et  $Y$ , la quantité

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^q n_{ij} (x_i - \bar{X})(y_j - \bar{Y})$$

$$= \sum_{i=1}^p \sum_{j=1}^q \frac{n_{ij}}{N} (x_i - \bar{X})(y_j - \bar{Y})$$

$$= \sum_{i=1}^p \sum_{j=1}^q f_{ij} (x_i - \bar{X})(y_j - \bar{Y})$$

# Nuage de points

Proposition: (*formule de Koenig*)

$$\begin{aligned} \text{Cov}(X, Y) &= \frac{1}{N} \sum_{i=1}^P \sum_{j=1}^q n_{ij} x_i y_j - \bar{X} \times \bar{Y} \\ &= \sum_{i=1}^p \sum_{j=1}^q f_{ij} x_i y_j - \bar{X} \times \bar{Y} \end{aligned}$$

Remarque : Si

$$n_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

Alors

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{x_1 + x_2 + \cdots x_N}{N}$$

# Nuage de points

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i = \frac{y_1 + y_2 + \cdots y_N}{N}$$

$$Var(X) = \frac{1}{N} \sum_{i=1}^N x_i^2 - (\bar{x})^2 = \frac{x_1^2 + x_2^2 + \cdots x_N^2}{N} - (\bar{x})^2$$

$$Var(Y) = \frac{1}{N} \sum_{i=1}^N y_i^2 - (\bar{y})^2 = \frac{y_1^2 + y_2^2 + \cdots y_N^2}{N} - (\bar{y})^2$$

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{x} \times \bar{y} = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_N y_N}{N} - \bar{x} \times \bar{y}$$

# Nuage de points

**Exemple 6:** Lors d'une période de sécheresse, un agriculteur relève la quantité d'eau totale (en  $m^3$ ) utilisée par son exploitation depuis le premier jour et donne le résultat suivant:

Nombre de jours écoulé $x_i$	1	3	5	8	10
Volume utilisé (en $m^3$ ) $y_i$	2,25	4,3	8	17,5	27

Nous avons

$$\bar{x} = \frac{1 + 3 + 5 + 8 + 10}{5} = \frac{17}{5} = 5,4$$

$$\bar{y} = \frac{2,25 + 4,3 + 8 + 17,25 + 27}{5} = \frac{59,05}{5} = 11,81$$

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{x} \times \bar{y}$$

## Nuage de points

$$\begin{aligned} Cov(X, Y) &= \frac{1}{5} [2,25 \times 1 + 4,3 \times 3 + 8 \times 5 + 17,5 \times 8 + 27 \times 10] - 5,4 \times 11,81 \\ &= 93,03 - 63,75 = 29,25 \end{aligned}$$

### Remarque

i.  $Cov(X, Y) = Cov(Y, X)$

ii.  $Cov(X, Y) \leq \sqrt{Var(X)} \sqrt{Var(Y)}$

iii. Si  $x'_i = ax_i + b \quad \forall i = 1, \dots, p$  et  $y'_j = cy_j + d \quad \forall j = 1, \dots, q$ , avec  $a, b, c$  et  $d$  sont des réels, alors  $Cov(X', Y') = ac \times Cov(X, Y)$ .

iv. Si  $X$  et  $Y$  sont des caractères indépendants, alors  $Cov(X, Y) = 0$ . La réciproque est fausse, càd  $Cov(X, Y) = 0$  n'entraîne pas que  $X$  et  $Y$  sont indépendants

# Ajustement linéaire d'un nuage de points

Les points  $(x_i, y_i)$  forment un *nuage* dont on cherche une *approximation* dans un but de

- *simplification*. Mais qui dit simplification dit *déformation* : nous voudrions qu'elle soit minimale ;

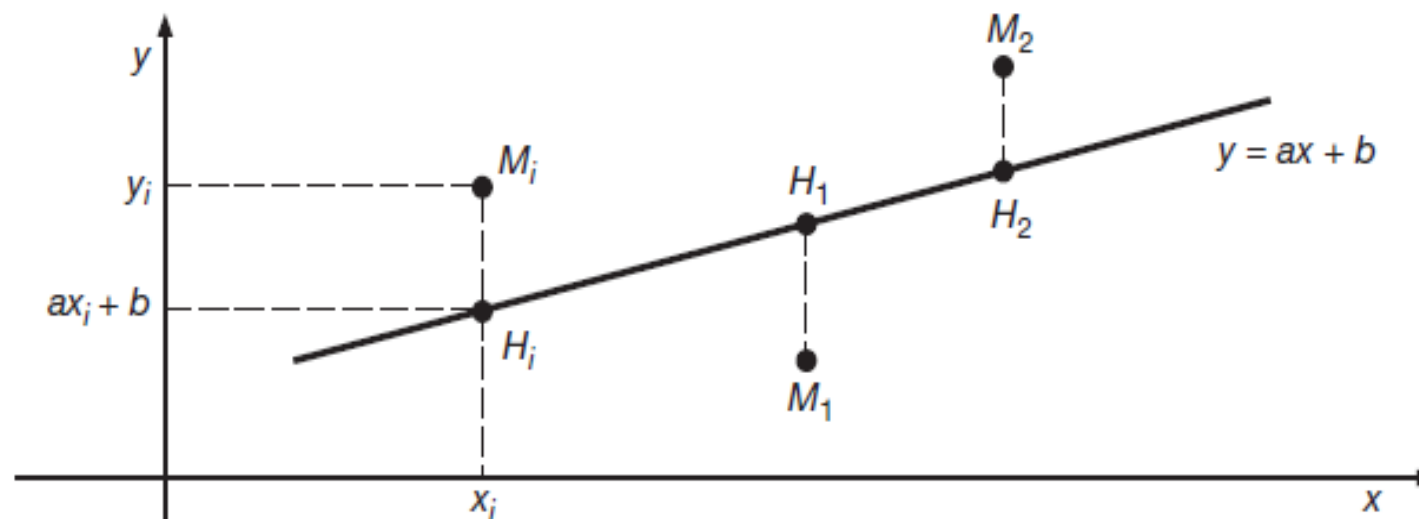
encore faut-il préciser ce que l'on entend par là. Disons tout de suite que le choix du critère sera *arbitraire* même si l'on tente de le justifier par des considérations plus ou moins «intuitives ». On peut vouloir par exemple :

- préserver *au mieux* les distances entre points ;
- préserver *au mieux* les angles des droites joignant les points...

# Ajustement linéaire d'un nuage de points

Il n'existe pas de moyen de satisfaire à toutes ces exigences à la fois. Il nous faut donc choisir.


Nous allons chercher la **meilleure droite au sens des moindres carrés**,  
c'est-à-dire telle que :  $\sum_{i=1}^n |M_i H_i|^2$  soit minimum



*Interprétation géométrique de la droite des moindres carrés*



# Ajustement linéaire d'un nuage de points

Les *distances* sont comptées *parallèlement* à l'un des axes des coordonnées ; nous avons choisi ici l'axe des ordonnées  figure

Il s'agit de déterminer la droite  $\mathbb{D}$  d'équation  $\{y = ax + b\}$  telle que :

$$F(a, b) = \sum_{i=1}^n \left( y_i - (ax_i + b) \right)^2 \text{ soit minimum}$$

Nos *inconnues* sont  $a$  et  $b$ .

Commençons par chercher le minimum de  $F(a, b)$  relativement à  $b$  lorsque  $a$  est fixé. On peut écrire  $F(a, b)$  comme un trinôme du second degré en  $b$  :

$$\begin{aligned} F(a, b) &= \sum_{i=1}^n \left( (y_i - ax_i) - b \right)^2 = \sum_{i=1}^n \left( (y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2 \right) \\ &= \sum_{i=1}^n (y_i - ax_i)^2 - 2b \sum_{i=1}^n (y_i - ax_i) + nb^2 \end{aligned}$$

Quand  $a$  est fixé, le dernier membre constitue une fonction de  $b$  qui atteint

# Ajustement linéaire d'un nuage de points

son minimum pour  $b = \hat{b}$  tel que  $\frac{\partial F}{\partial b}(a, \hat{b}) = 0$ , soit :

$$\frac{\partial F}{\partial b}(a, \hat{b}) = -2 \left( \sum_{i=1}^n (y_i - ax_i) - n\hat{b} \right) = 0$$

$$\Rightarrow \hat{b} = \frac{1}{n} \sum_{i=1}^n (y_i - ax_i) = \bar{y} - a\bar{x}$$

- 1<sup>re</sup> *conséquence* : la droite des moindres carrés passe par le point de coordonnées  $(\bar{x}, \bar{y})$  qu'on appelle parfois *le centre de gravité* ou *point moyen du nuage*.

# Ajustement linéaire d'un nuage de points

Notre problème est maintenant de trouver le minimum de  $F(a, \hat{b})$  relativement à  $a$  :

$$\begin{aligned} F(a, \hat{b}) &= \sum_{i=1}^n \left( (y_i - \bar{y}) - a(x_i - \bar{x}) \right)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2a \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + a^2 \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

ce qui peut encore s'écrire :

$$F(a, \hat{b}) = n \left( a^2 \text{var}(X) - 2a \text{cov}(X, Y) + \text{var}(Y) \right)$$

Le coefficient de  $a^2$  étant positif ou nul, ce trinôme du second degré en  $a$  atteint son *minimum* relativement à  $a$  pour  $a = \hat{a}$  avec :

$$\hat{a} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$$

# Ajustement linéaire d'un nuage de points

Ainsi le couple  $(\hat{a}, \hat{b})$  avec  $\hat{b} = \bar{y} - \hat{a}\bar{x}$  réalise le minimum de la fonction  $F$

- 2<sup>e</sup> conséquence : la droite des moindres carrés a pour équation  $y = \hat{a}x + \hat{b}$  soit

$$y - \bar{y} = \frac{\text{cov}(X, Y)}{\text{var}(X)} \cdot (x - \bar{x})$$

On posera pour tout  $i$  variant de 1 à  $n$  :  $\hat{y}_i = \hat{a}x_i + \hat{b}$ ,  $\hat{y}_i$  est la *valeur estimée* de  $Y$  par la droite des moindres carrés lorsque  $X = x_i$

# Ajustement linéaire d'un nuage de points

## 2) Droite des moindres carrés $\mathbb{D}'$

Dans toute l'étude précédente, on a fait jouer des rôles non symétriques à  $X$  et à  $Y$ . On a procédé comme si la variable  $X$  pouvait être mesurée, et qu'on cherchait à *prévoir* la variable  $Y$ .

Inversement, la droite  $\mathbb{D}'$  des moindres carrés pour laquelle les distances sont comptées parallèlement à l'axe des abscisses a pour équation :

$$x - \bar{x} = \frac{\text{cov}(X, Y)}{\text{var}(Y)} \cdot (y - \bar{y}) \quad \Rightarrow \quad y - \bar{y} = \frac{\text{var}(Y)}{\text{cov}(X, Y)} \cdot (x - \bar{x})$$

Mais, dans certains cas, comme celui où la variable  $X$  désigne le temps, seule la droite  $\mathbb{D}$  a un sens.

# Ajustement linéaire d'un nuage de points

Il est toujours possible de tracer la droite des moindres carrés précédente quelle que soit la forme du nuage. L'approximation du nuage par cette droite est-elle *légitime* ? Quel sens, quelle signification donner à cette droite ?

C'est là une autre question, et fort importante. On pourra dire qu'il est d'autant plus légitime de remplacer le nuage par la droite trouvée que la dispersion du nuage de points par rapport à la droite des moindres carrés :

# Droite de régression

La **droite de régression** est une droite qui passe par le **point moyen**. C'est aussi la droite qui **minimise la somme des carrés des écarts des observations**. Une fois connue, l'équation de cette droite permet de résumer la série et de faire des prévisions.

# Droite de regression

## Définition :

- La droite de regression ( $D$ ) de  $Y$  en  $X$  a pour équation

$$y - \bar{y} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} (x - \bar{x}),$$

$a = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$  s'appelle **coefficient de regression** (ou coefficient directeur) de  $y$  en  $x$

La droite de regression ( $D'$ ) de  $X$  en  $Y$  a pour équation

$$x - \bar{x} = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)} (y - \bar{y}),$$

$a' = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)}$  s'appelle **coefficient de regression** (ou coefficient directeur) de  $x$  en  $y$



# Droite de regression

**Exemple7:** Reprenons l'exemple 6, nous avons

$$\bar{x} = 5,4 ; \bar{y} = 11,8 ; Cov(X, Y) = 29,25$$

Pour déterminer les droites de regression, on doit calculer  $Var(X)$  et  $Var(Y)$

$$\begin{aligned} Var(X) &= \frac{x_1^2 + x_2^2 + \dots + x_5^2}{5} - (\bar{x})^2 \\ &= \frac{1^2 + 3^2 + 5^2 + 8^2 + 10^2}{5} - (5,4)^2 \\ &= 39,8 - 29,16 = 10,64 \end{aligned}$$

## Droite de regression

$$\begin{aligned} Var(Y) &= \frac{y_1^2 + y_2^2 + \dots + y_5^2}{5} - (\bar{y})^2 \\ &= \frac{2,25^2 + 4,3^2 + 8^2 + 17,5^2 + 27^2}{5} - (11,8)^2 \\ &= 224,561 - 139,476 = 85,1. \end{aligned}$$

La droite de regression de  $Y$  en  $X$  est donnée par:

$$y = \frac{Cov(X, Y)}{Var(X)} (x - \bar{x}) + \bar{y} = \frac{29,25}{10,64} (x - 5,4) + 11,8 = 2,74x - 3,04$$

Le coefficient de regression de  $y$  en  $x$  est  $a = 2,74$ .

# Droite de regression

La droite de regression de  $X$  en  $Y$  est donnée par:

$$x = \frac{Cov(X, Y)}{Var(Y)} (y - \bar{y}) + \bar{x} = \frac{29,25}{85,1} (y - 11,8) + 5,4 = 0,34y - 1,34$$

Le coefficient de regression de  $x$  en  $y$  est  $a' = 0,34$ .

## Remarque:

- i. Pour une valeur donnée de  $x_0$ , l'ajustement permet de prévoir approximativement la valeur correspondante  $y_0$
- ii. Les deux droites de regression ( $D$ ) et ( $D'$ ) passent toutes les deux par le point moyen  $G(\bar{x}, \bar{y})$ .

# Droite de regression

La droite de régression sert d'abord à **vérifier l'existence d'une relation linéaire** et la nature de celle-ci. Ainsi, dans notre exemple, le coefficient directeur de la droite

$a = 2,74$  est positif ce qui dénote une relation positive :  $x$  et  $y$  varient dans le même sens.

La droite de régression sert ensuite à **faire des prévisions**. Ainsi, nous pouvons utiliser l'équation de la droite de régression pour calculer des valeurs de  $Y$  associées à une valeur de  $X$  que l'on se donne.

# Coefficient de corrélation

**Définition:** On appelle **coefficient de corrélation** de la série statistique double, le nombre  $r$  défini par:

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Ce coefficient mesure la plus ou moins grande dépendance entre les deux caractères  $X$  et  $Y$ . On le désigne souvent par la lettre " $r$ " et il varie entre -1 et +1 :

**Remarque:**

- i.*  $r$  est un nombre sans dimension (sans unité).
- ii.*  $sign(r) = sign(Cov(X, Y))$ .
- iii.*  $r^2 = aa'$  (coefficient de détermination).
- iv.*  $r^2 \approx 1 \Rightarrow$  les points se trouvent à peu près sur une droite. Il y a une très forte
- v.* corrélation linéaire entre  $X$  et  $Y$ . les deux droites ( $D$ ) et ( $D'$ ) sont presque confondues

# Coefficient de corrélation

vi.  $r^2 \approx 0 \Rightarrow$  les points sont très dispersés. Il y a une faible corrélation (pas de liaison). Les deux droites  $(D)$  et  $(D')$  sont presque perpendiculaire.

**Exemple:** Déterminons le coefficient de corrélation dans l'exemple 6, nous avons

$$Var(X) = 10,64; \quad Var(Y) = 85,1 \text{ et } Cov(X, Y) = 29,25.$$

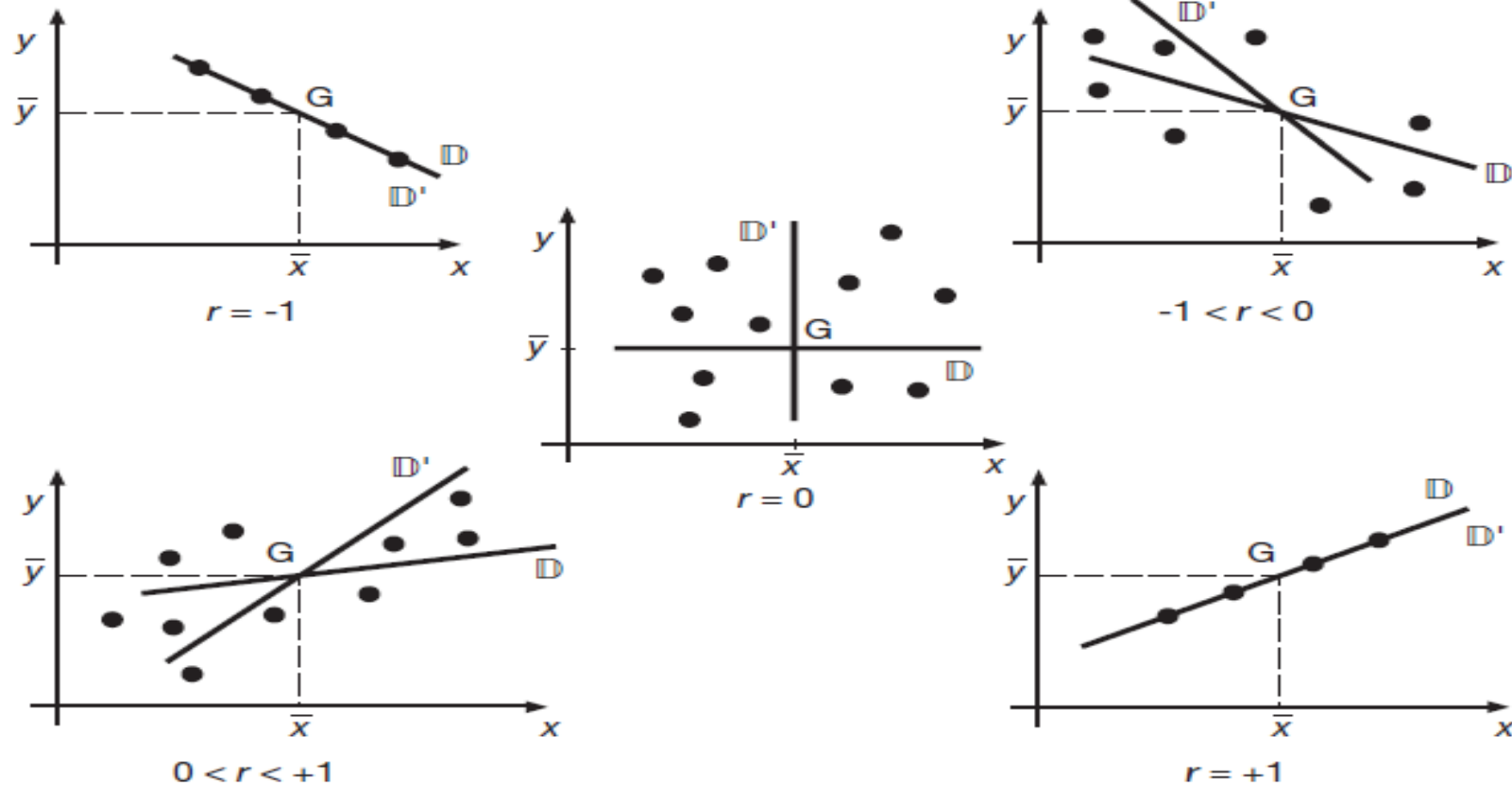
Le coefficient de corrélation est donc

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{29,25}{\sqrt{10,64 \times 85,1}} = 0,97.$$

Le coefficient de détermination est alors  $r^2 = 0,945 \approx 1$

On conclut qu'il y a une très forte corrélation linéaire entre  $X$  et  $Y$ . les deux droites  $(D)$  et  $(D')$  sont presque confondues.

# Coefficient de corrélation



*Positions respectives des droites des moindres carrés selon les valeurs de  $r$*