Centre Universitaire Ain Témouchent

Math 04 : Probabilités et Statistiques

Dr. AISSA MAMOUNE Sidi Mohammed

Département des Sciences et Technologie

Institut des Sciences et Technologie

E-mail : aissa_mamoune@yahoo.fr

Intitulé du domaine	**Sciences et Technologie**
Année	**2ème année**
Intitulé de la matière	**Maths 4 : Probabilités et Statistiques**
Annuel ou semestriel	**Semestriel**
Unité d'enseignement	**UEM3 Méthodologie**
Volume horaire global 45 heures	**1h30 /semaine Cours** **1h30 /semaine TD**
Chargé de la matière	**Mr Sidi Mohammed AISSA MAMOUNE**
Nombre de crédits	**4**

Dr Sidi Mohammed AISSA MAMOUNE

Département des Sciences et Technologie - Institut des Sciences et Technologie
Centre Universitaire Ain Témouchent
BP 284 (46000), Tel/Fax : 043 60 34 47

Email : aissa_mamoune@yahoo.fr

Doctorat en Génie Civil	**UABB - FSI**	**2009**	**Algérie**
Certificate of Distance-Learning Methodologies	**Université Missouri Rolla (UMR)**	**2005**	**USA**
Magister en Génie Civil	**UABB - FSI**	**2002**	**Algérie**
Ingénieur d'Etat en Génie Civil	**UABB - FSI**	**1999**	**Algérie**

Objectifs du cours

Le cours a pour but d'initier les étudiants
aux principes de base de la probabilité et statistique.

Support pédagogique

Il est mis à la disposition des étudiants un support pédagogique
sur papier du Cours et des Travaux Dirigés (TD).

Plateforme Elearning (**l'adresse vous sera transmise prochainement**)

Notation et examens

La note finale **MOY** est calculée sur la base de deux (02) notes :

- Epreuve finale **Note1**
- Contrôle continu **Note2**

MOY=(2*Note1 + Note2) / 3

La note du contrôle continu (**Note2**) est calculée sur la base de deux (**02**) notes

- Epreuve **T1**
- Assiduité **T2**

Note2= **T1***0.80 + **T2***0.20

Organisation du Cours

Partie A	**Introduction générale et organisation du cours** *Chapitre1*
Partie B	**Traitement statistique de l'information** *Chapitre2 à Chapitre 5*
Partie C	**Traitement probabiliste de l'information** *Chapitre6 à Chapitre 9*

SOMMAIRE

L’incertain est-il notre quotidien????

Ce qui est sûr, c’est que rien n’est sûr

Probabilité et Statistiques :

Outils au service de l'engineering

Place du cours dans votre futur métier

- Analyse des données
- Prédictions
- Simulations (processus stochastiques)
- Décisions (probabilités d'occurrence et risque)
- …

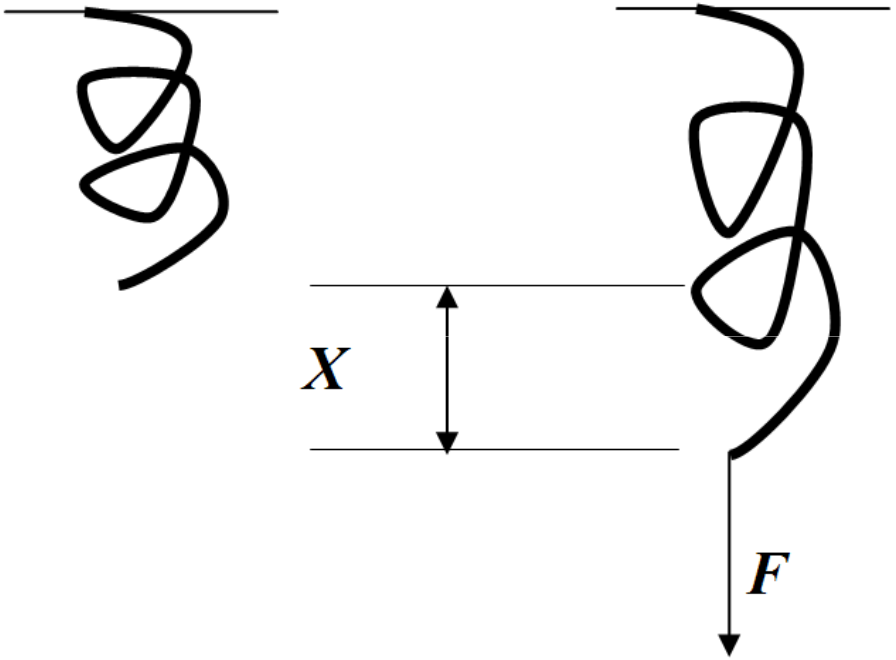
La conception d'un système donné nécessite trois étapes :

1. La compréhension du système : quelle est sa fonction

2. La modélisation du système : quel est le modèle à développer pour décrire ce système avec l'identification de l'input et l'output de ce modèle

3. Le recours aux données : l'utilisation de données nécessaires pour l'utilisation du modèle. Ces données sont l'input du modèle.

Considérons un exemple très simple……



Etape 1

Nul besoin de montrer le fonctionnement d'un ressort dans un système mécanique (suspension de voiture,…)

Etape 2

Essayons maintenant de voir comment modéliser ce ressort.

L'input du modèle est:

- La force agissante, dans notre cas *F*
- Le ou les caractéristiques du ressort, dans notre cas *K*
- L'output du modèle est par exemple la déformation du ressort, dans notre cas *X*.

Une relation toute simple a été mise en place **(modèle mathématique).**

$$F = K.X \qquad X = \frac{F}{K}$$

Etape 3

Donc en se basant sur ce modèle mathématique, il est possible de connaître la déformation du ressort connaissant *F* et *K*

En fait, il est possible de dire que la connaissance de l'output (Déformation du ressort) est acquise.

Est-ce vrai ?

NON

Pourquoi ?

1. Le premier problème qui se pose est tout d'abord: Est ce que le modèle mis en place est "exact" ?

2. Le deuxième problème auquel on est confronté est la validité de l'information de l'Input. En d'autres termes peut-on dire avec certitude que **les valeurs de *F* et *K* sont exactes et connues**

La statistique est un ensemble de méthodes permettant:

- ✓ de recueillir des données “brutes”;
- ✓ de présenter, résumer ces données;
- ✓ de tirer des conclusions sur la population étudiée (sa structure, sa composition), d’aider à la prise de décision; en présence de données dépendant du temps, de faire de la prévision.

Les outils de la statistique et de la probabilité permettent de répondre à la question suivante:

Comment déterminer la valeur de l'Output si l'Input et/ou le modèle mathématique du phénomène étudié ne sont pas connus

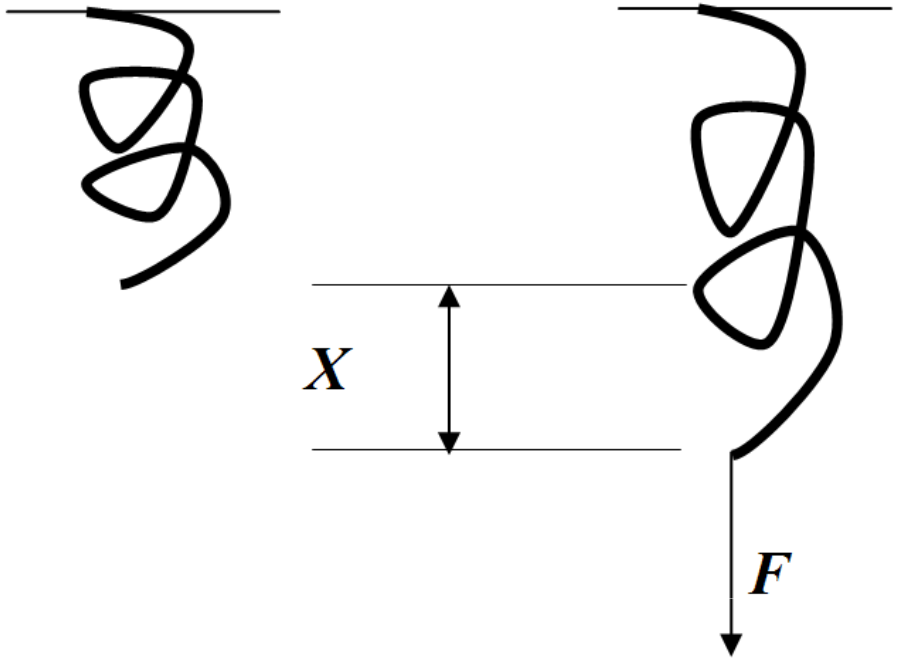
- *Population ou unité statistique et/ou échantillon*
- *Individu*
- *Caractère*
- *Modalités*

-Population: Employés d'une usine

-Individu: Un employé de cette usine

-Caractère: Salaire

-Modalités: 10000DA, 20000DA, 25000DA

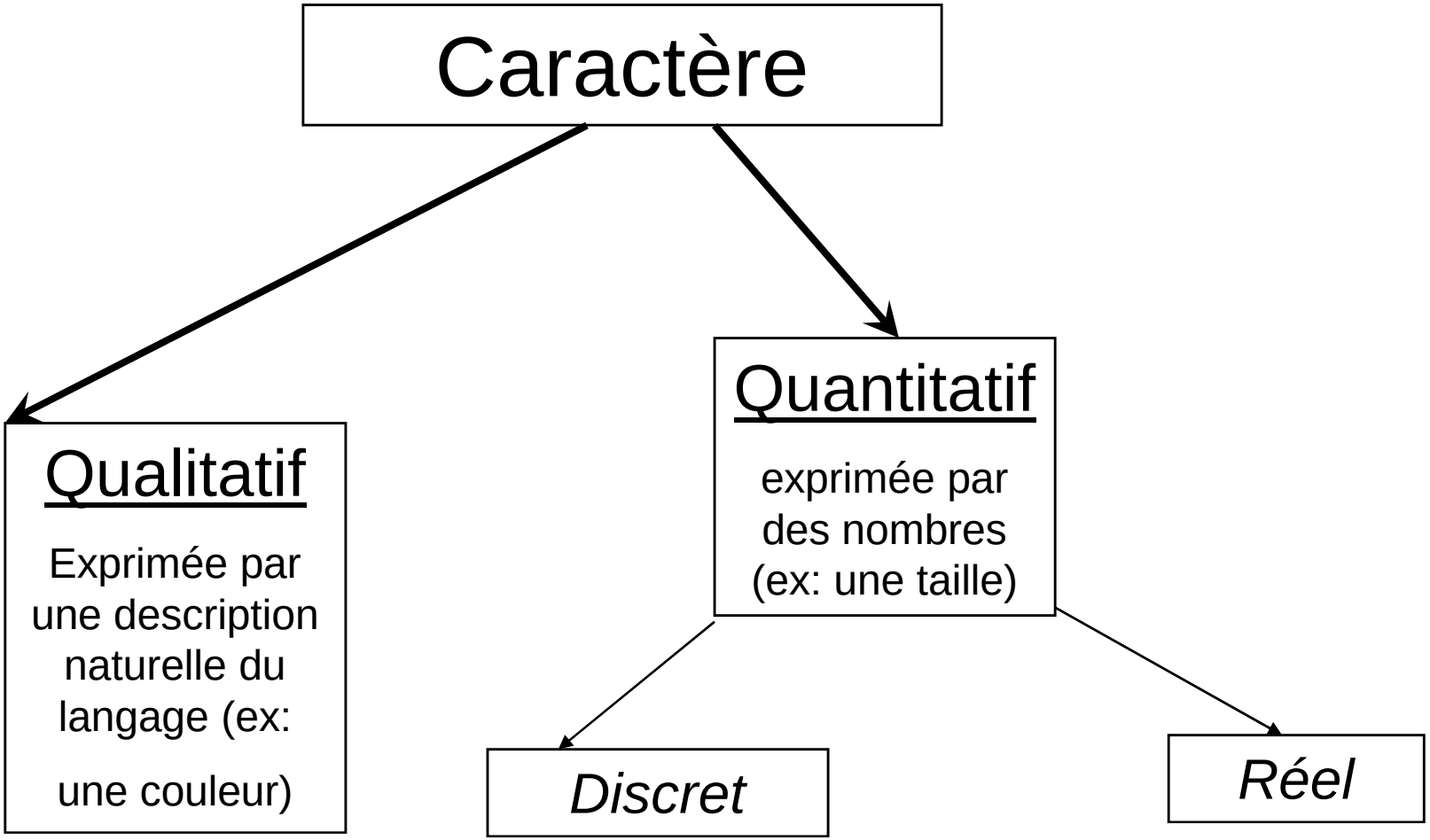


-Population: Ressorts

-Individu: Un ressort parmi ces ressort

-Caractère: Rigidité K

-Modalités: $K \in [10,20] \quad N/m$


Caractère
Qualitatif
Exprimée par une description naturelle du langage (ex:
une couleur)
Quantitatif
exprimée par des nombres (ex: une taille)
Discret
Réel

Exemple:
On souhaite connaitre l'état des maisons
Choix entre les trois types de caractère

-Population: Maisons (100)

-Individu: Une maison parmi ces 100 maisons

-Caractère: L'état de la maison

-Modalités: Petite, moyenne, grande

Caractère qualitatif

-Population: Maisons (100)

-Individu: Une maison parmi ces 100 maisons

-Caractère: Nombre de pièces

-Modalités: 1, 2, 3, 4, 5

Caractère quantitatif discret

-Population: Maisons (100)

-Individu: Une maison parmi ces 100 maisons

-Caractère: Surface (notée S)

-Modalités: $S \in [60, 200]$ m²

Caractère quantitatif continu

Une compagnie achète 10 000 ampoules électriques d'un fabricant qui affirme que ses ampoules fonctionnent durant au moins 1 000 heures **(1 mois et 11 jours, sans arrêt).** Cette compagnie vérifie 15 ampoules et, suite à ces résultats doit décider si elle garde ou non les 10 000 ampoules.

Identifier la population, l'individu, le caractère et les modalités

Une compagnie achète 10 000 ampoules électriques d'un fabricant qui affirme que ses ampoules fonctionnent durant au moins 1 000 heures **(1 mois et 11 jours, sans arrêt).** Cette compagnie vérifie 15 ampoules et, suite à ces résultats doit décider si elle garde ou non les 10 000 ampoules.

Population	:	l'ensemble des *10 000 ampoules achetées*.
Échantillon	:	les *15 ampoules vérifiées*.
Individu	:	une *ampoule* parmi les 15
Caractère	:	*durée* de fonctionnement de l'ampoule
Modalités	:	*Durée* en heures

Variable statistique (VS) continue

Population	:	l'ensemble des *10 000 ampoules achetées*.
Échantillon	:	les *15 ampoules vérifiées*.
Individu	:	une ampoule parmi les 15
Caractère	:	l'*état* de l'ampoule
Modalités	:	Bon ou mauvais

Caractère qualitatif

Un même individu peut il avoir plusieurs caractères????

Oui

Population	:	l'ensemble des *10 000 ampoules achetées*.
Échantillon	:	les *15 ampoules vérifiées*.
Individu	:	une ampoule parmi les 15
Caractère	:	l'*état* de l'ampoule la *durée* de fonctionnement de l'ampoule
Modalités	:	Bon ou mauvais *Durée* en heures

Les notes des étudiants

8,75	6,00	3,75	11,25	11,50	7,00	11,75	10,50	5,00	6,75	7,50
11,75	8,75	16,25	12,00	10,50	15,25	12,50	11,00	9,50	7,75	8,50
12,50	4,00	16,00	7,50	9,00	11,00	13,50	8,00	13,00	10,75	12,75
5,50	9,00	9,25	0,00	8,75	10,75	6,50	7,00	12,00	7,00	9,50
15,25	11,50	10,75	5,25	11,50	9,25	9,75	9,75	14,75	6,00	15,25
10,50	11,00	4,75	13,25	11,50	12,00	12,00	12,00	9,00	4,25	7,00
9,00	9,50	13,25	15,25	8,00	12,25	10,75	7,25	9,50	7,50	10,25
14,75	15,50	10,50	10,25	13,50	9,50	5,00	5,50	9,50	5,50	5,75
6,50	1,25	14,75	16,50	11,75	5,75	4,50	7,50	16,25	6,50	17,75
6,50	2,00	14,00	13,50	11,00	10,75	9,50	0,00	15,50	10,75	4,50
6,50	7,00	8,75	7,75	16,75	11,50	8,50	10,25	12,75	10,00	2,75

Population : l'ensemble des étudiants (121)
Individu : un étudiant parmi les 121.
Caractère : la note de l'examen
Modalités : les valeurs de la note [0,20]

Soit $\boldsymbol{P}$ une population formée de n individus (x_i, $i=1,..n$).

Soit $\boldsymbol{C}$ un caractère ayant k modalités c_1, c_2,…,c_k . Ce caractère peut être qualitative, quantitative (discret ou continu)

Remarquez que $n \neq k$

Pourquoi????

La collecte de l'information relative au caractère C auprès de la population P consiste à observer pour chaque individu de P la modalité qui lui correspond.

Le résultat obtenu est la ***série statistique***

L'individu n°1 identifié par x_1 présente une modalité parmi les k modalités (avec par exemple k=10). Les autres individus vont avoir les modalités suivantes:

$$x_1 = c_3$$
$$x_2 = c_{10}$$
$$x_3 = c_3$$
$$.$$
$$.$$
$$.$$
$$x_n = c_8$$

Le traitement de cette information élémentaire consiste à dénombrer pour chaque modalité c_j, le nombre d'individus de la population P n_j qui présentent cette modalité.

La distribution statistique est donc formée par les couples

$$\left(c_j, n_j\right) \quad j = 1,2,...,k$$

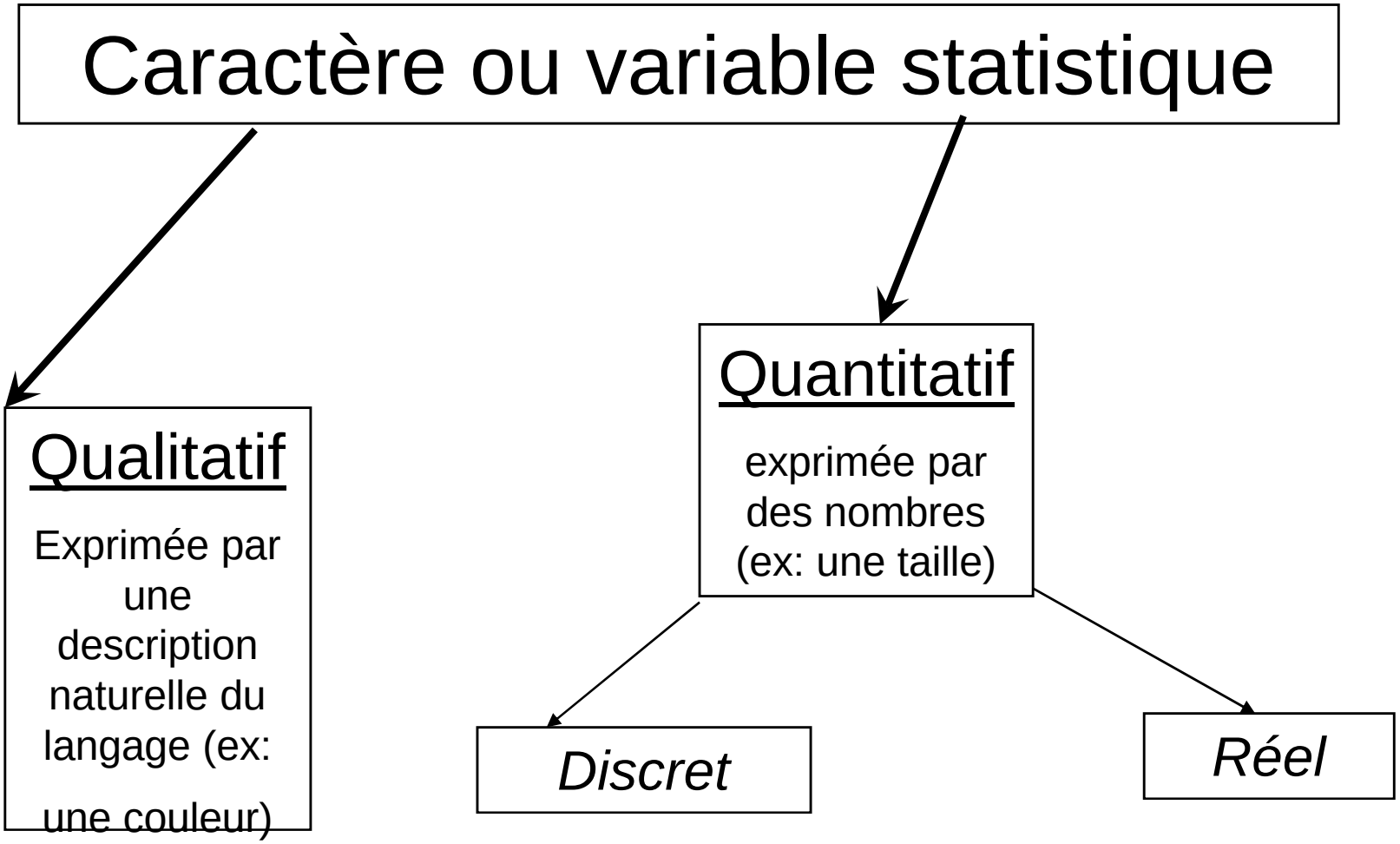
Notion de fréquence

Le nombre n_j est dit effectif ou fréquence absolue de la modalité c_i. Il est clair que :

$$n = \sum_{j=1}^{j=k} n_j$$

Le nombre f_j est dit fréquence de la modalité c_j.

$$f_j = \frac{n_j}{n} \qquad \sum_{j=1}^{j=k} f_j = 1$$



k modalités $c_1, c_2, \ldots, c_k$

Jours du mois de Janvier **Population**

Climat **Caractère**

Ensoleillé, pluvieux, orageux **Modalité**

$n=31$.
$k=3$ modalités

c_1=*Ensoleillé*
c_2=*Pluvieux*
c_3=Orageux

Remarquez que $n \neq k$

Série statistique

Individu	Modalité
1	Ensoleillé
2	Ensoleillé
3	Ensoleillé
4	Pluvieux
5	Pluvieux
6	Pluvieux
7	Orageux
8	Orageux
9	Pluvieux
10	Pluvieux
11	Ensoleillé
12	Ensoleillé
13	Pluvieux
14	Pluvieux
15	Pluvieux
16	Pluvieux
17	Orageux
18	Orageux
19	Pluvieux
20	Pluvieux
21	Pluvieux
22	Ensoleillé
23	Ensoleillé
24	Ensoleillé
25	Pluvieux
26	Pluvieux
27	Pluvieux
28	Orageux
29	Orageux
30	Orageux
31	Ensoleillé

Modalité	Total	Fréquence de la modalité
Total Ensoleillé	9	0.29
Total Orageux	7	0.23
Total Pluvieux	15	0.48

Tableau statistique

Les 10 séismes recensés durant une année au niveau d'une région **Population**

Intensité **Caractère**

Intensité I à XII **Modalité**

$n=10$.
$k=12$ modalités

$c_1=1$
$c_2=2$

$c_{12}=12$

Série statistique

Séisme n°	Intensité (Modalité)
1	2
2	4
3	6
4	8
5	2
6	1
7	6
8	7
9	5
10	5

Tableau statistique

Intensité (Modalité)	Effectif	Fréquence de la modalité
1	1	0.1
2	2	0.2
3	0	0
4	1	0.1
5	2	0.2
6	2	0.2
7	1	0.1
8	1	0.1
9	0	0
10	0	0
11	0	0
12	0	0

Les 10 séismes recensés durant une année au niveau d'une région

	Population
Magnitude de Richter	**Caractère**
Valeur de la Magnitude	**Modalité**

<u>Les valeurs de la magnitude sont réels</u>

$n=10$.

$k=????$

Comment construire le tableau statistique???

Série statistique

Séisme n°	Magnitude (Modalité)
1	2.2
2	3.1
3	5.5
4	7.5
5	2.2
6	1.5
7	4.3
8	6.8
9	4.9
10	4.7

Du moment que le caractère est quantitatif continu alors l'idée consiste à établir des classes et ce pour mettre en place le tableau statistique

Pas=1

Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,2.5[	3	0.3
Classe 2	[2.5,3.5[	1	0.1
Classe 3	[3.5,4.5[	1	0.1
Classe 4	[4.5,5.5[	2	0.2
Classe 5	[5.5,6.5[	1	0.1
Classe 6	[6.5,7.5[	1	0.1
Classe 7	[7.5,8.5[	1	0.1

Est-ce qu'il y a un seul tableau statistique ?

NON

Séisme n°	Magnitude (Modalité)
1	2.2
2	3.1
3	5.5
4	7.5
5	2.2
6	1.5
7	4.3
8	6.8
9	4.9
10	4.7

Pas=1.5

Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,3[	3	0.3
Classe 2	[3,4.5[	2	0.2
Classe 3	[4.5,6[	3	0.3
Classe 4	[6,7.5[	1	0.1
Classe 5	[7.5,9[	1	0.1

Pas=3.0

Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,4.5[	5	0.5
Classe 2	[4.5,7.5[	4	0.4
Classe 3	[7.5,9.5[	1	0.1

Donc

- Pour un pas de 1 on a 7 classes
- Pour un pas de 1.5 on a 5 classes
- Pour un pas de 3 on a 3 classes

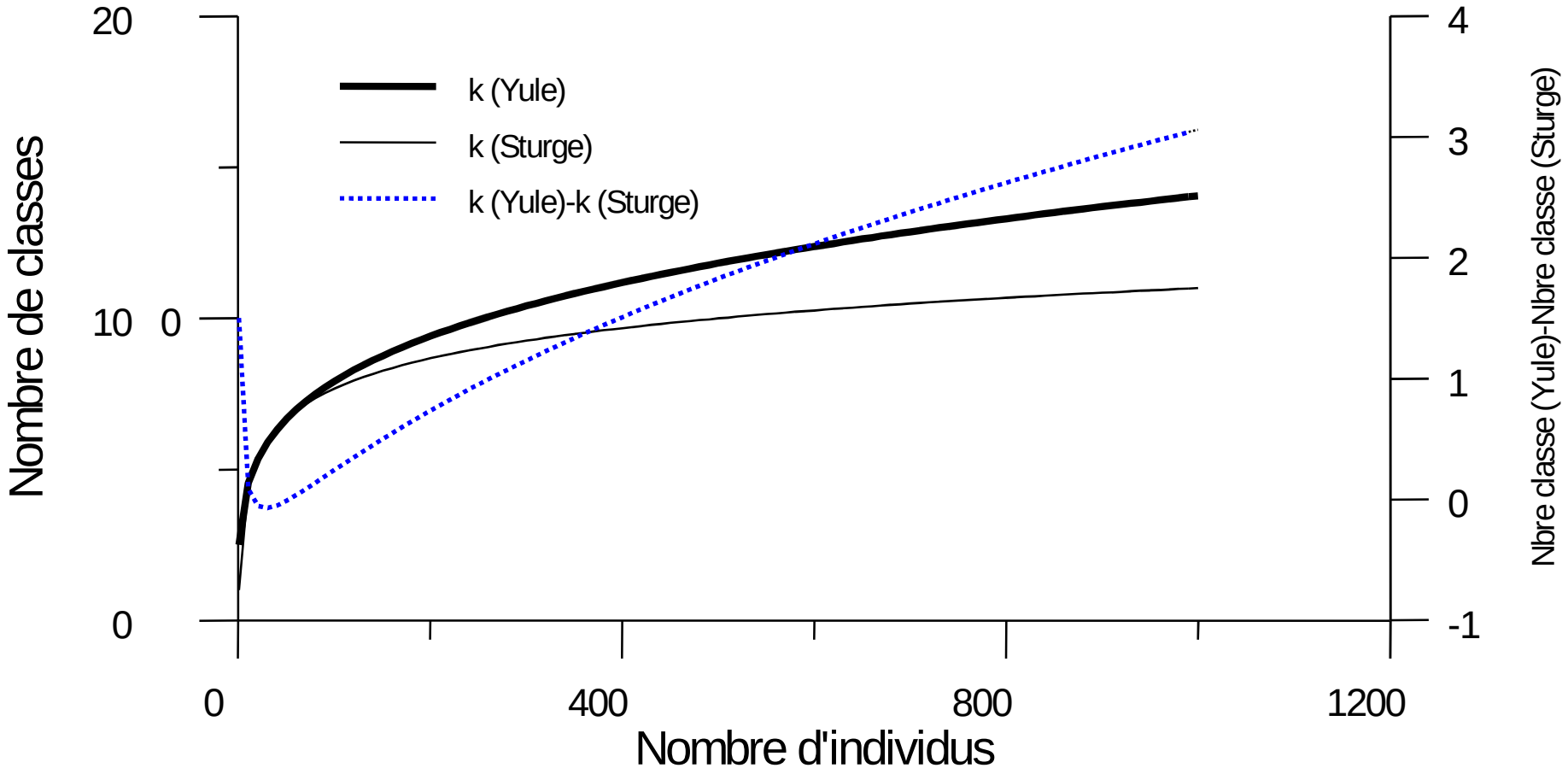
<u>Questions</u> :

- Est-ce que le choix du **pas** a une importance dans le traitement des données ?
- Quel **pas** choisir ? Cette question peut être posée autrement
- Quel est le nombre de classes que l'on doit choisir

Soit k le nombre de classes

STURGE $$k = 1 + 3.3 \log_{10}(n)$$

YULE $$k = 2.5 \sqrt[4]{n}$$



$$k = 1 + 3.3\log_{10}(10) = 4.3$$

$$k = 5$$

Chapitre 3

Distribution statistique à un caractère

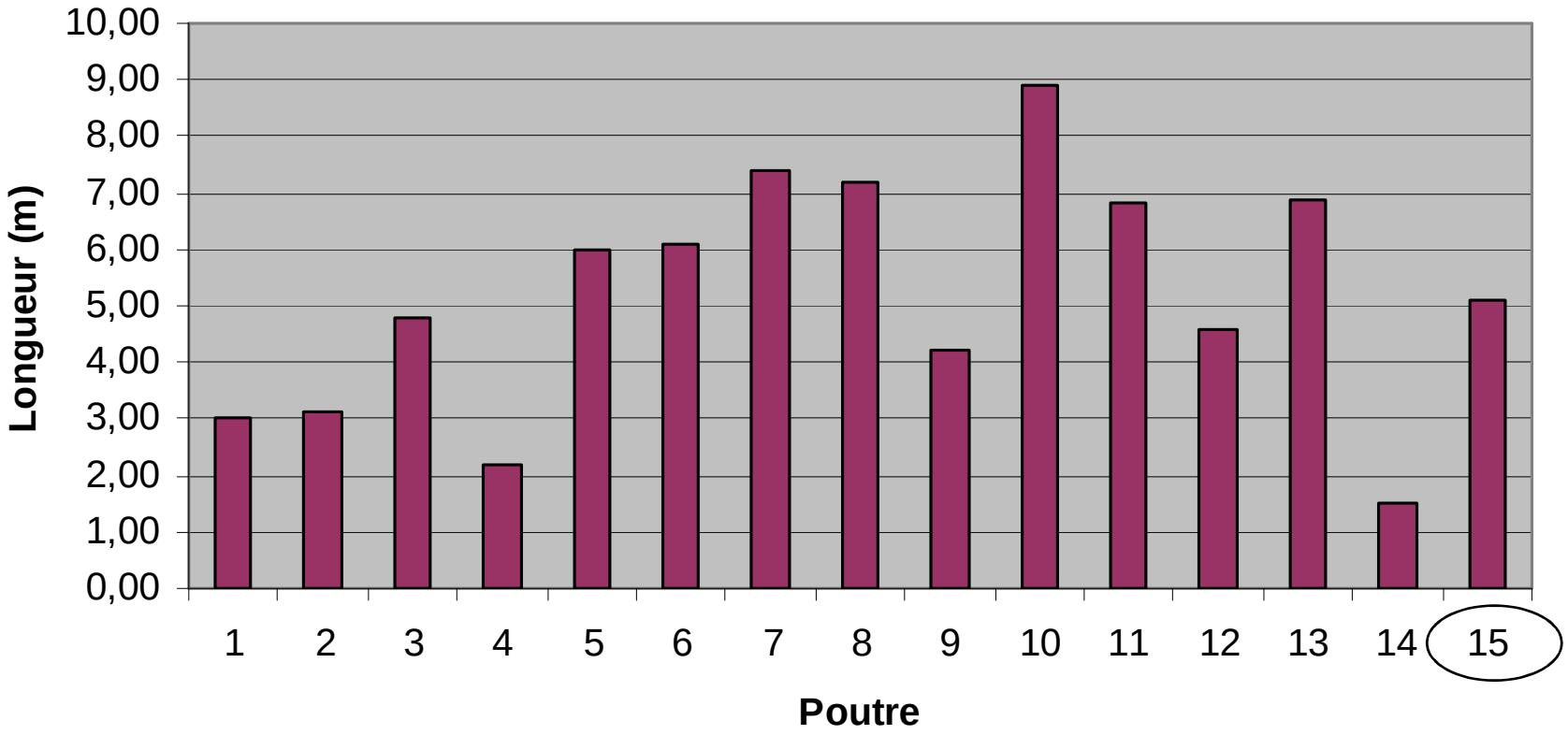
3.1 Description graphique

3.1.1 Variable quantitative continue

La série statistique

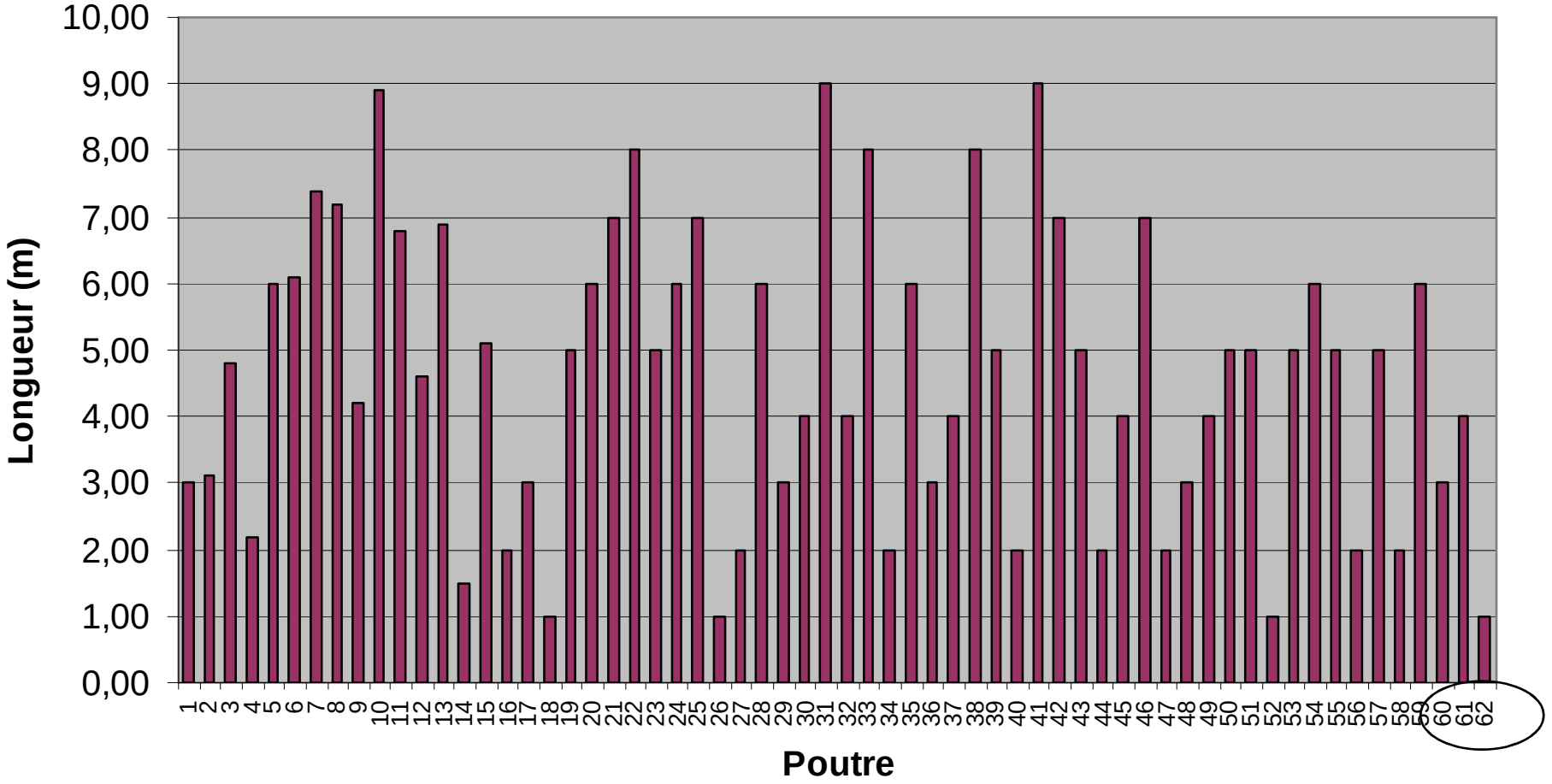
N° Individu	Longueur (m)
1	3.0
2	3.1
3	4.8
4	2.2
5	6.0
6	6.1
7	7.4
8	7.2
9	4.2
10	8.9
11	6.8
12	4.6
13	6.9
14	1.5
15	5.1

Histogramme



Exploitation difficile

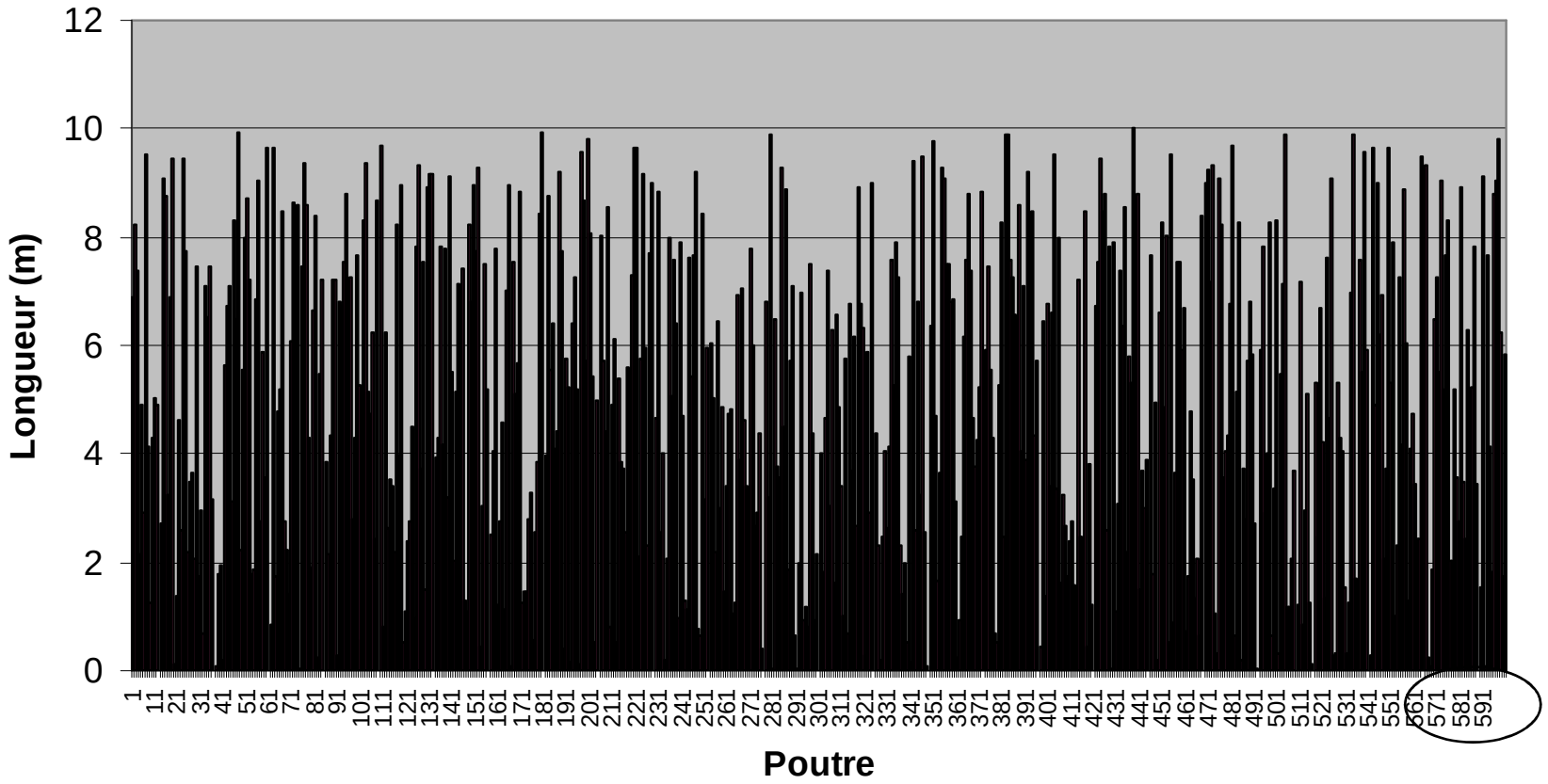
Histogramme



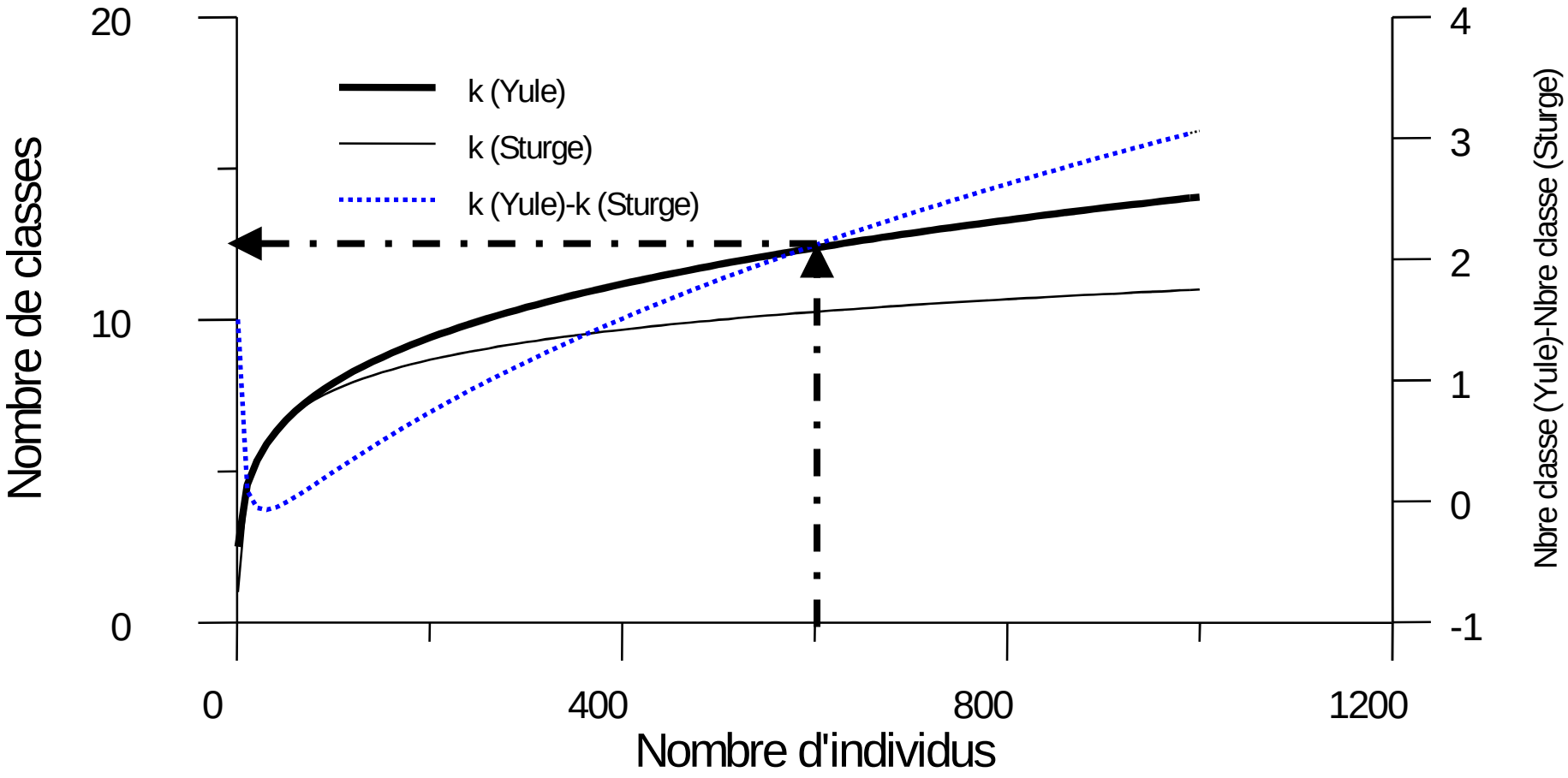
L'Exploitation devient plus difficile si le nombre d'individu augmente

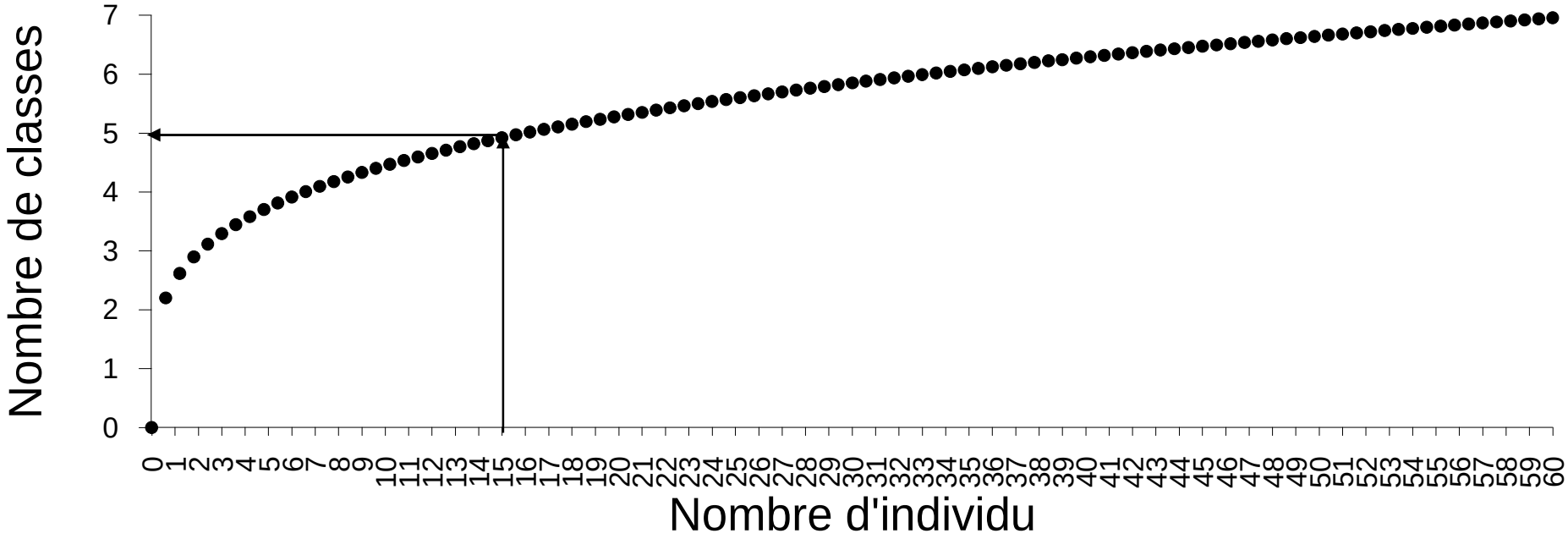
Dans le présent cas de 15 à 62

Histogramme



Avec un nombre d'individu qui avoisine les 600 individus, l'exploitation devient pratiquement **impossible**


20
10
0
Nombre de classes
k (Yule)
k (Sturge)
k (Yule)-k (Sturge)
4
3
2
1
0
-1
Nbre classe (Yule)-Nbre classe (Sturge)
0
400
800
1200
Nombre d'individus


Nombre de classes
0
1
2
3
4
5
6
7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
Nombre d'individu

Établissement du tableau (ou distribution) statistique

N° Individu	Longueur (m)
1	3.0
2	3.1
3	4.8
4	2.2
5	6.0
6	6.1
7	7.4
8	7.2
9	4.2
10	8.9
11	6.8
12	4.6
13	6.9
14	1.5
15	5.1

$$n = 15$$

$$k = 1 + 3.3\log_{10}(15)$$

$$= 4.88$$

$$k = 5$$

$$V_{\max} = 8.9\ \text{m}$$

$$V_{\min} = 1.50\ \text{m}$$

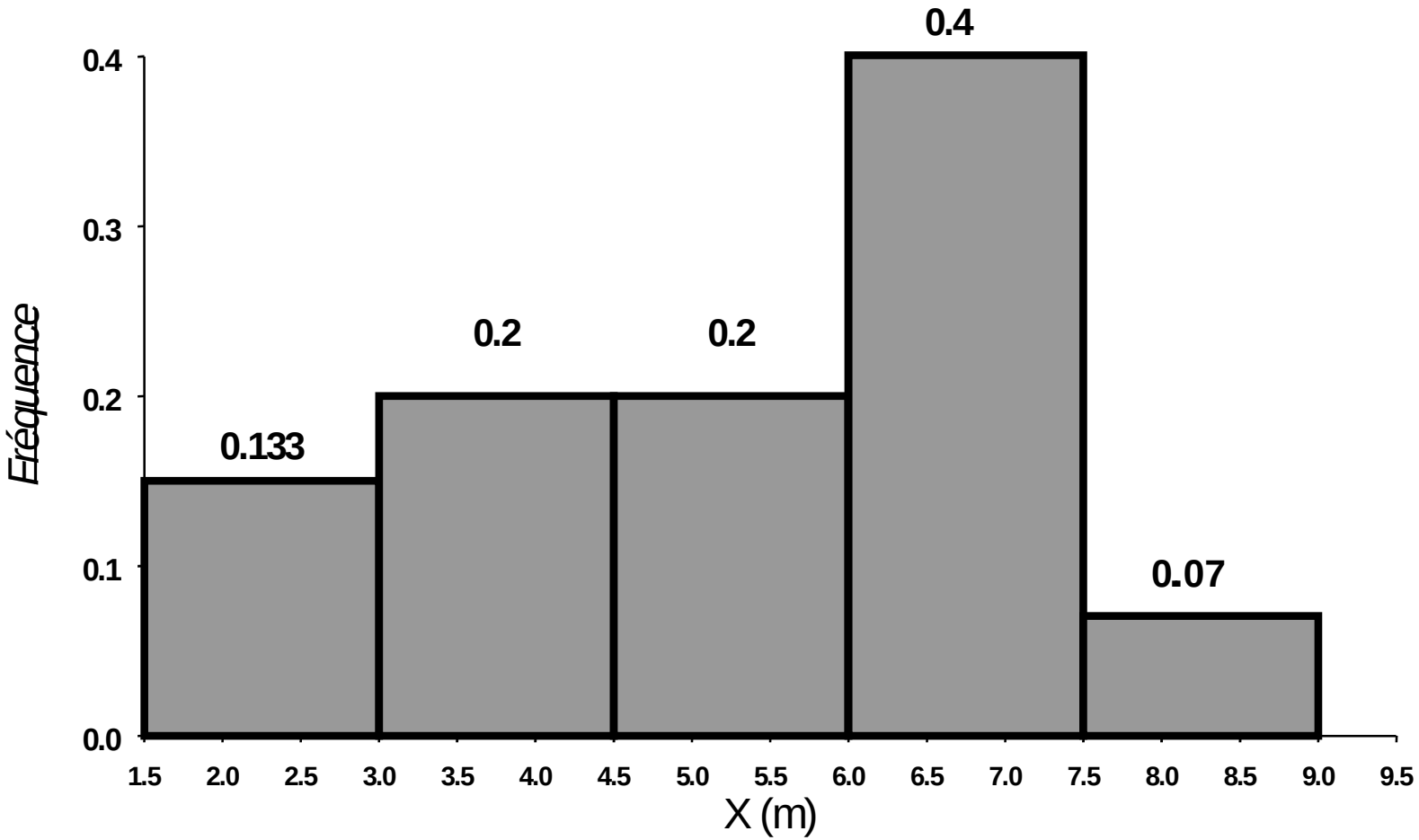
N° Individu	Longueur (m)
1	3.0
2	3.1
3	4.8
4	2.2
5	6.0
6	6.1
7	7.4
8	7.2
9	4.2
10	8.9
11	6.8
12	4.6
13	6.9
14	1.5
15	5.1

$$Pas = \frac{V_{\max} - V_{\min}}{k}$$

$$Pas = \frac{8.9 - 1.5}{5} = 1.48\,\text{m}$$

$$Pas = 1.50\,\text{m}$$

N° Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5, 3.0 [	2	0.13
Classe 2	[3.0, 4.5 [	3	0.2
Classe 3	[4.5, 6.0 [	3	0.2
Classe 4	[6.0, 7.5 [	6	0.4
Classe 5	[7.5, 9.0 [	1	0.07


0.4
0.133
0.2
0.2
0.4
0.07
Fréquence
0.0
0.1
0.2
0.3
0.4
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
X (m)

Pas=1

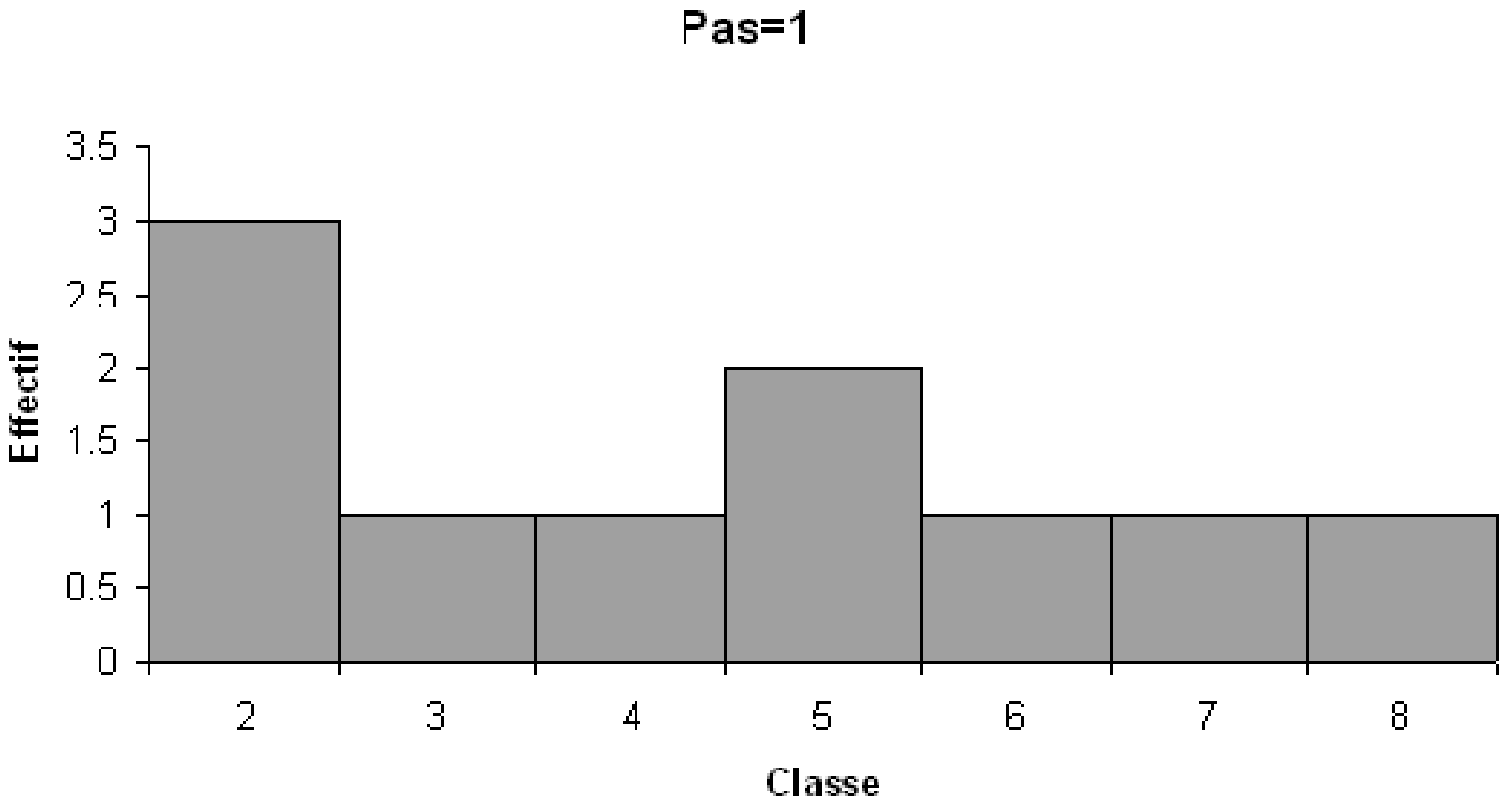
Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,2.5[	3	0.3
Classe 2	[2.5,3.5[	1	0.1
Classe 3	[3.5,4.5[	1	0.1
Classe 4	[4.5,5.5[	2	0.2
Classe 5	[5.5,6.5[	1	0.1
Classe 6	[6.5,7.5[	1	0.1
Classe 7	[7.5,8.5[	1	0.1

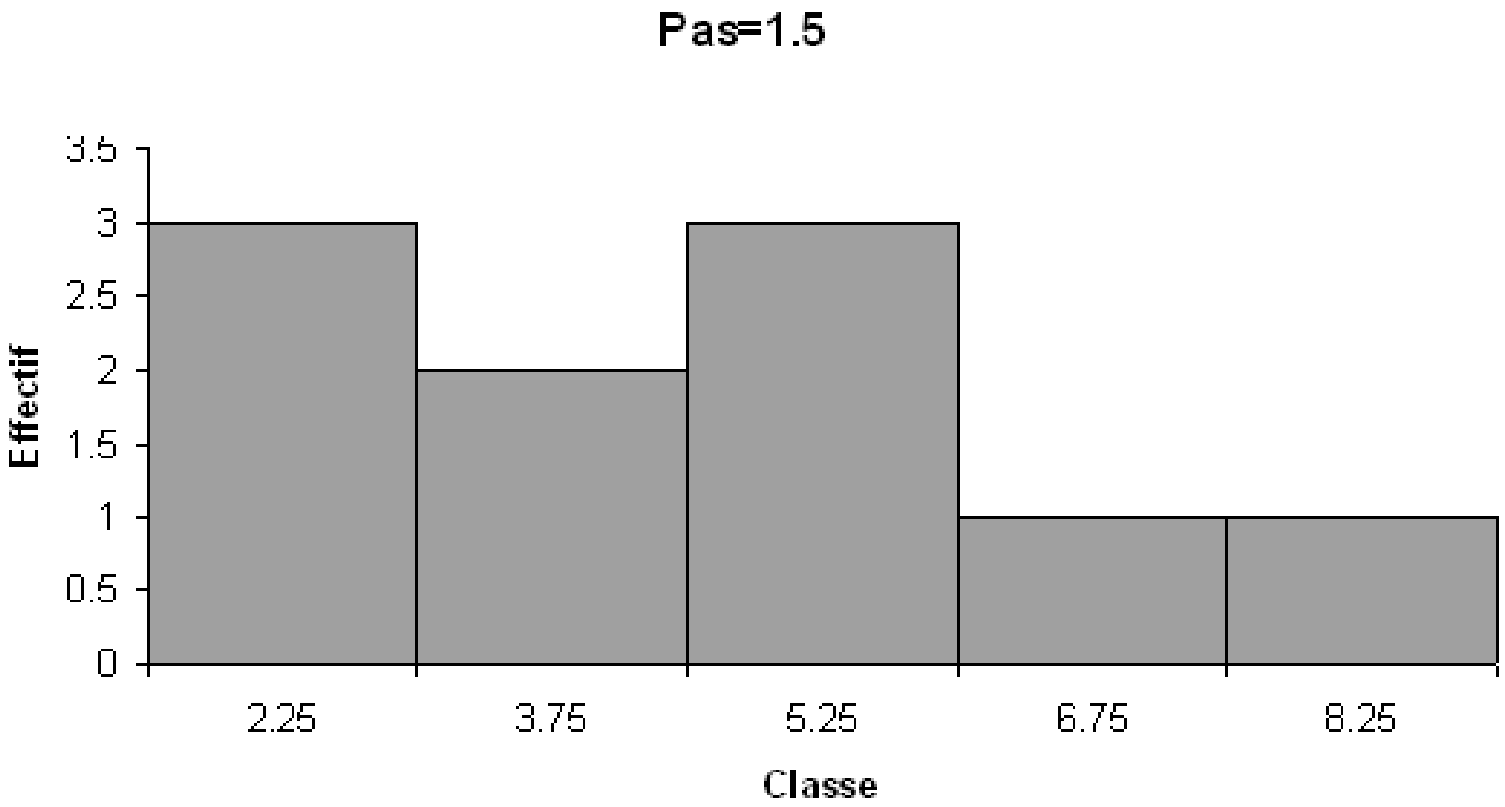
Pas=1.5

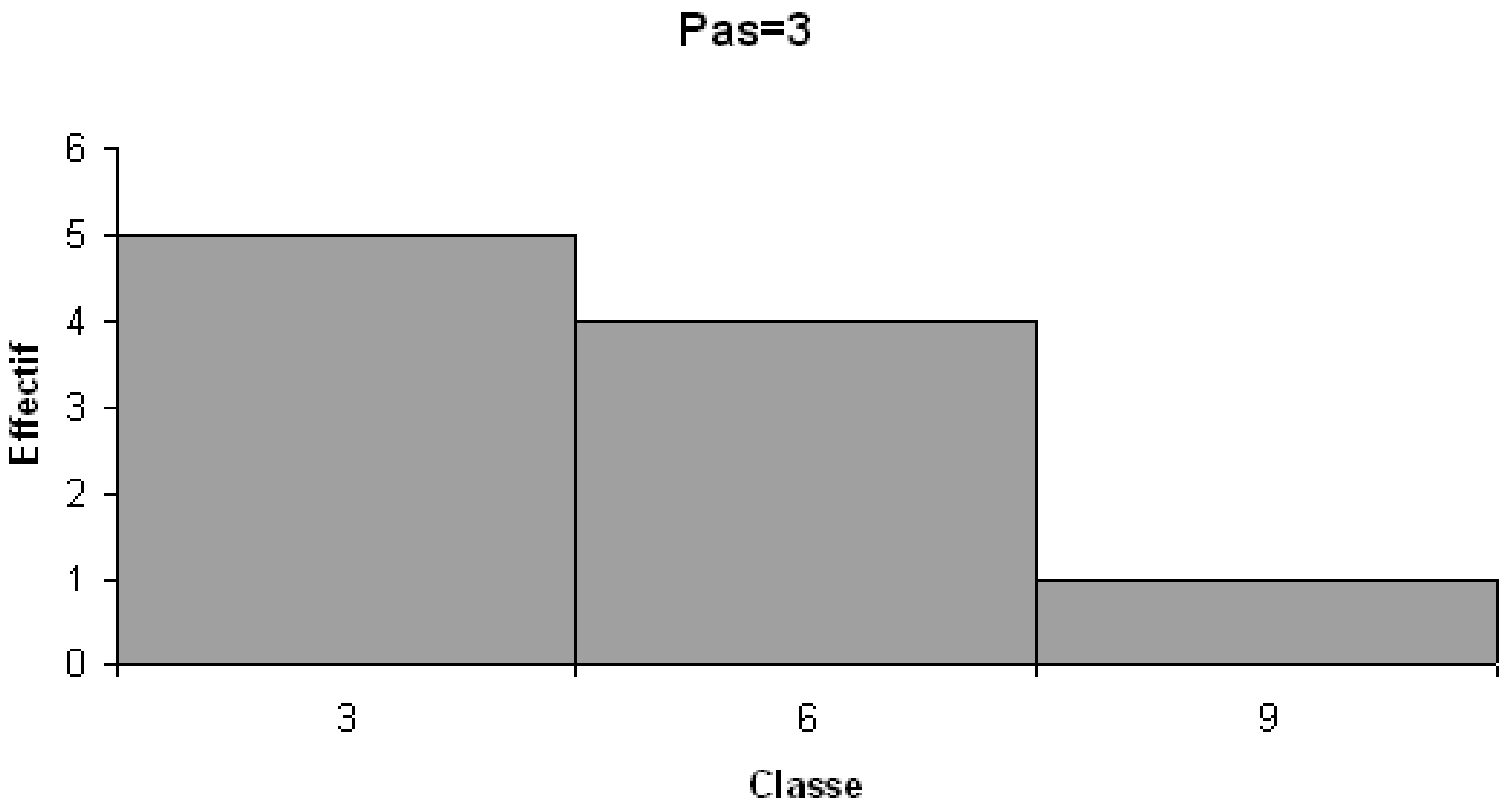
Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,3[	3	0.3
Classe 2	[3,4.5[	2	0.2
Classe 3	[4.5,6[	3	0.3
Classe 4	[6,7.5[	1	0.1
Classe 5	[7.5,9[	1	0.1

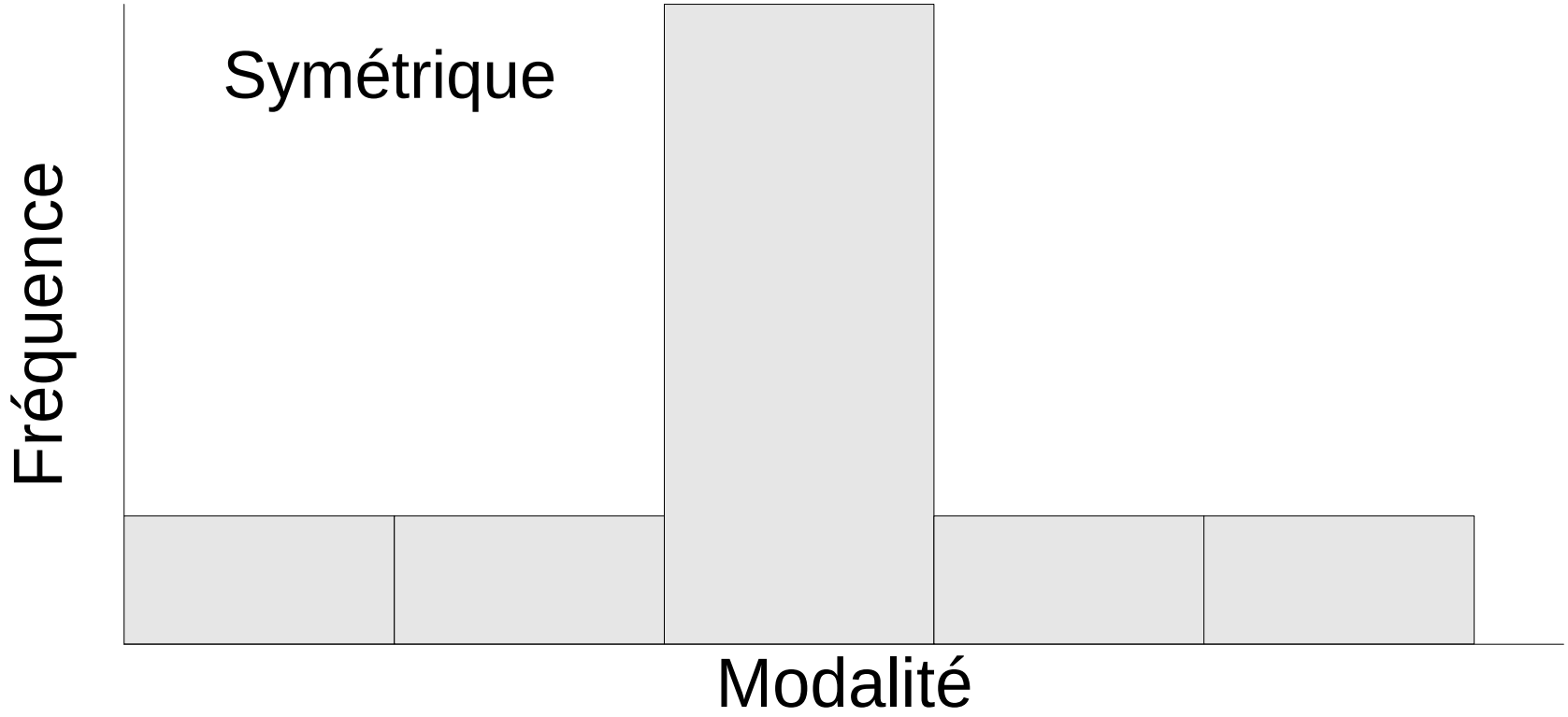
Pas=3.0

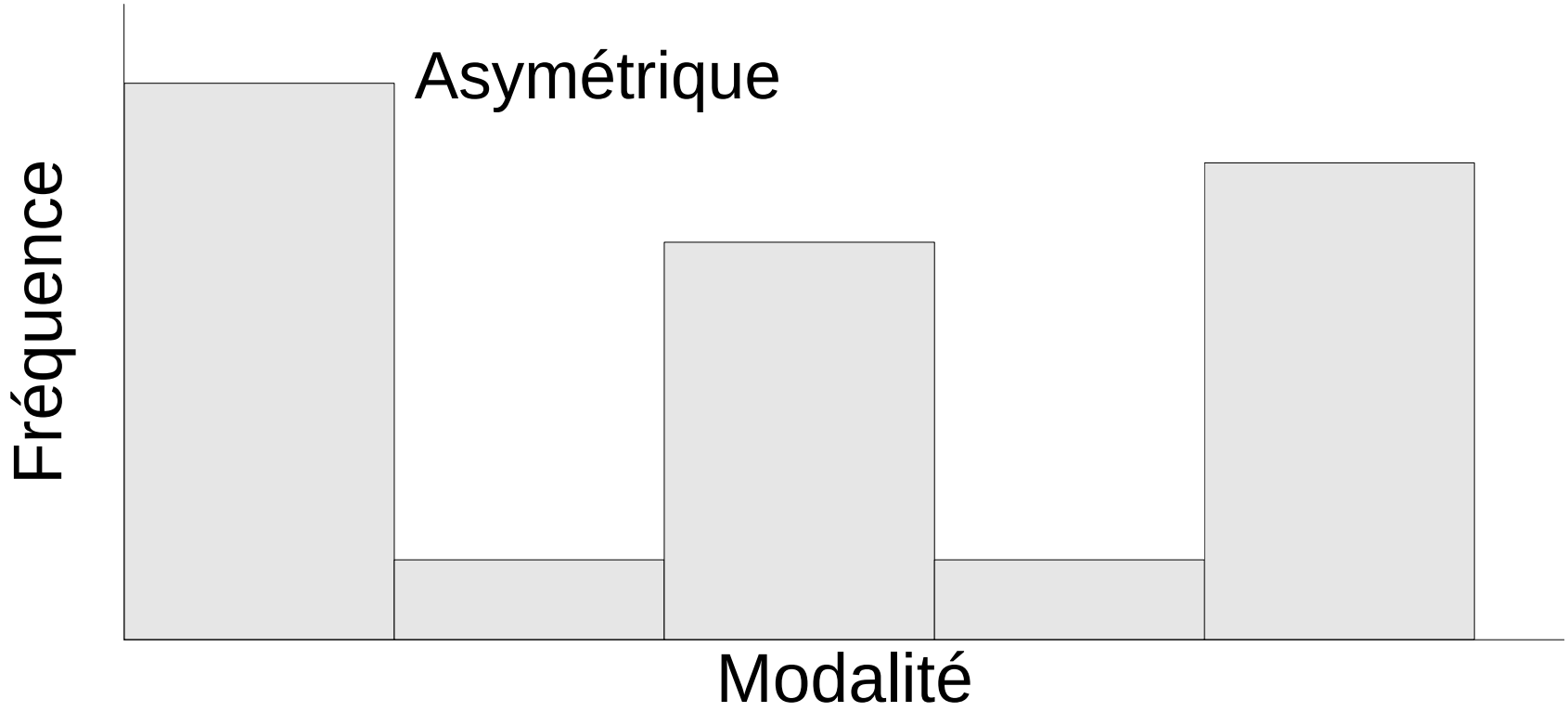
Classe	Intervalle	Nombre d'individu	Fréquence
Classe 1	[1.5,4.5[	5	0.5
Classe 2	[4.5,7.5[	4	0.4
Classe 3	[7.5,9.5[	1	0.1

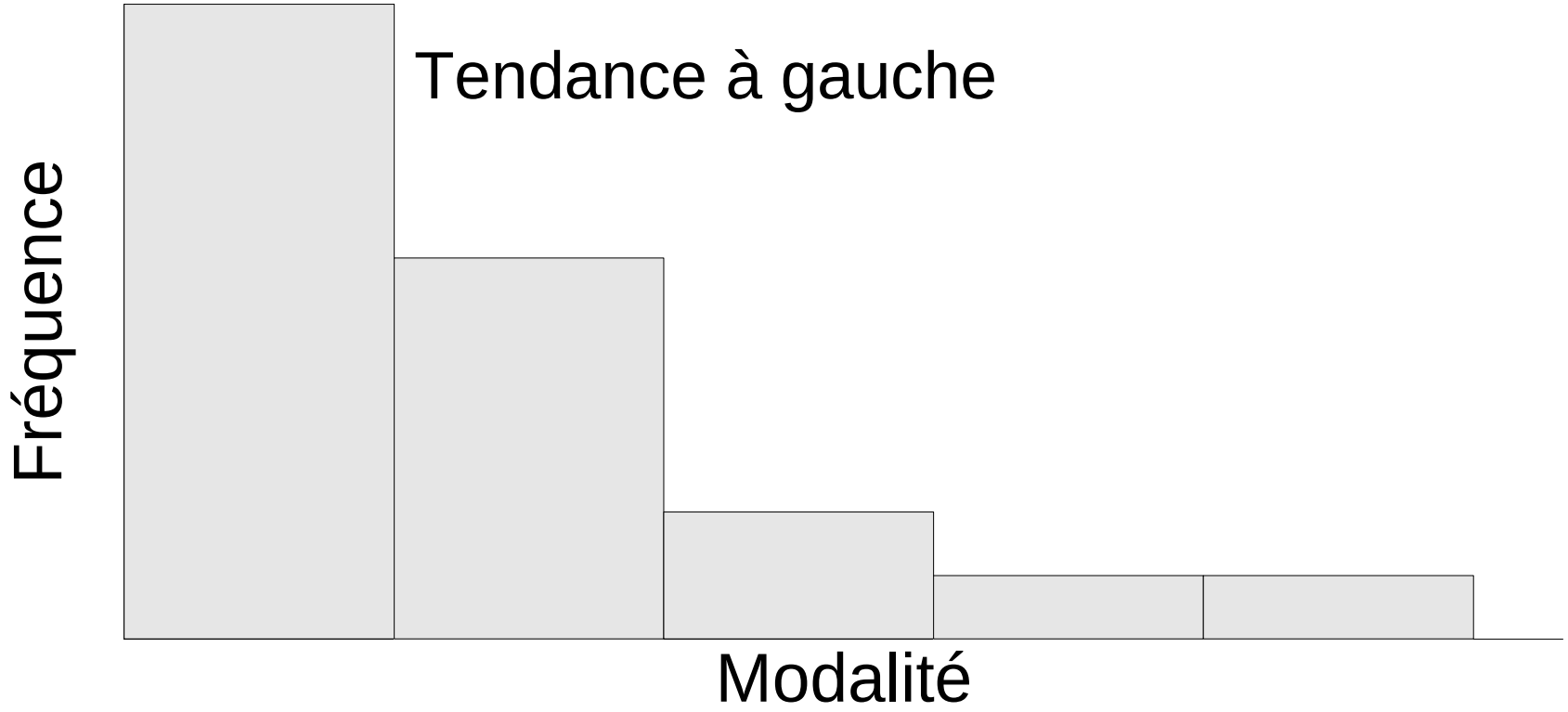

Pas=1
Effectif
3.5
3
2.5
2
1.5
1
0.5
0
2
3
4
5
6
7
8
Classe

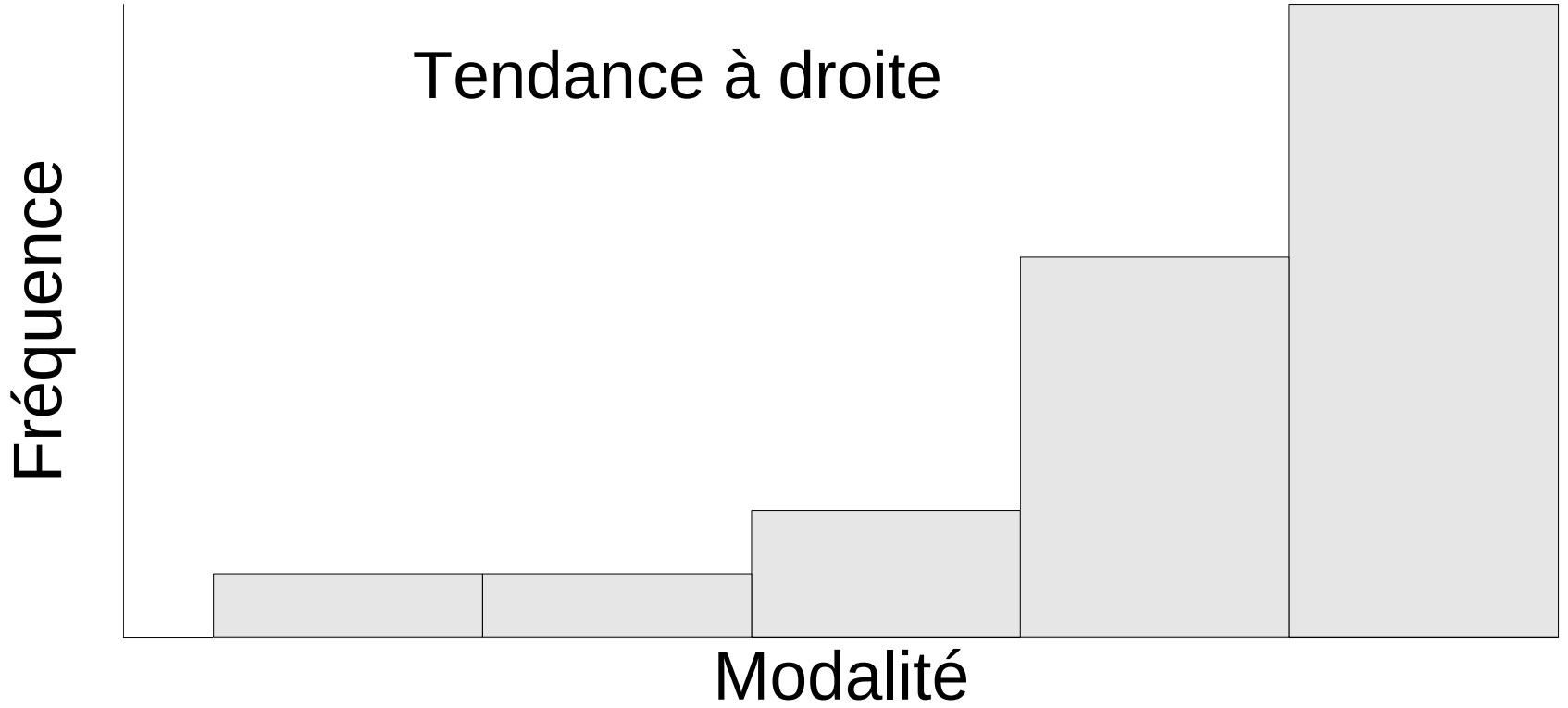

Pas=1.5
Effectif
3.5
3
2.5
2
1.5
1
0.5
0
2.25
3.75
5.25
6.75
8.25
Classe

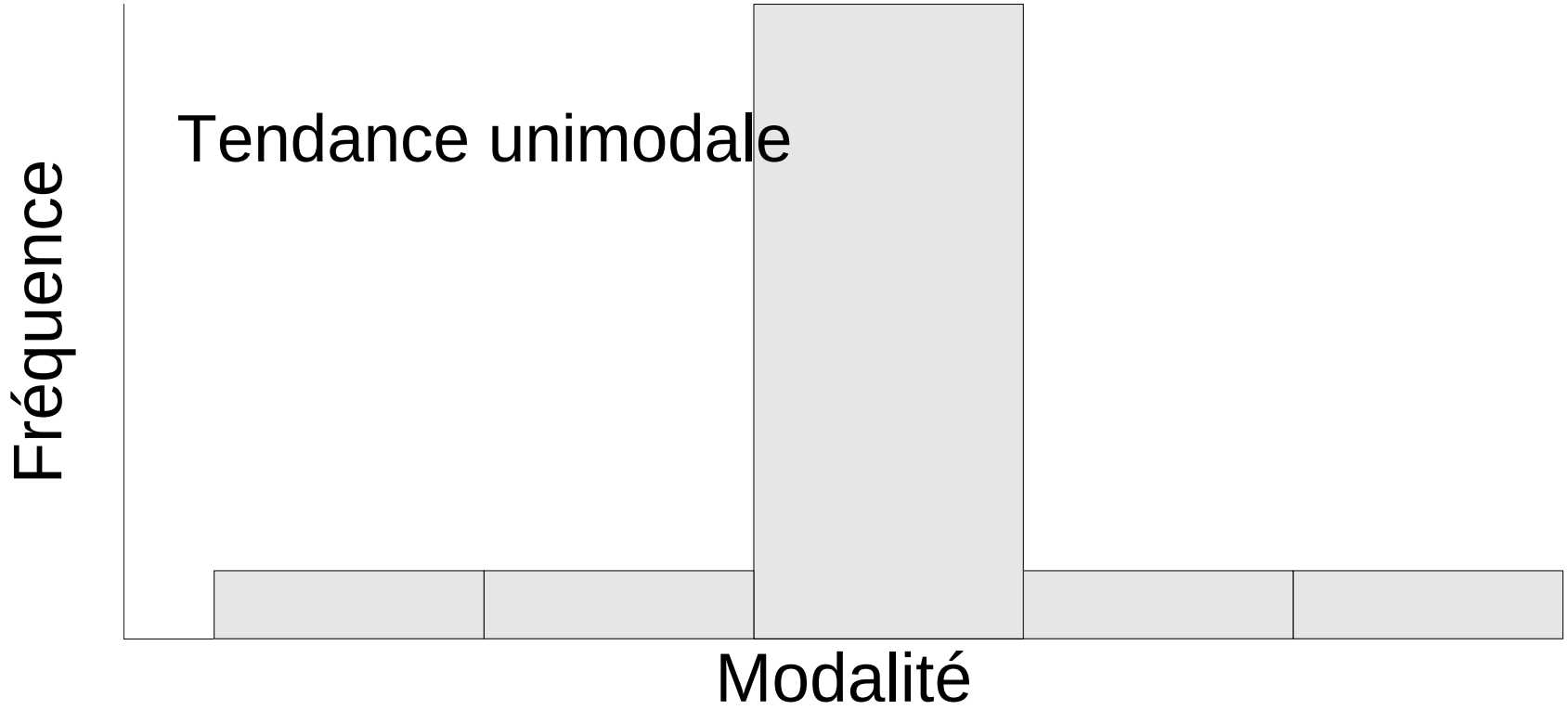

Pas=3
Effectif
0
1
2
3
4
5
6
3
6
9
Classe

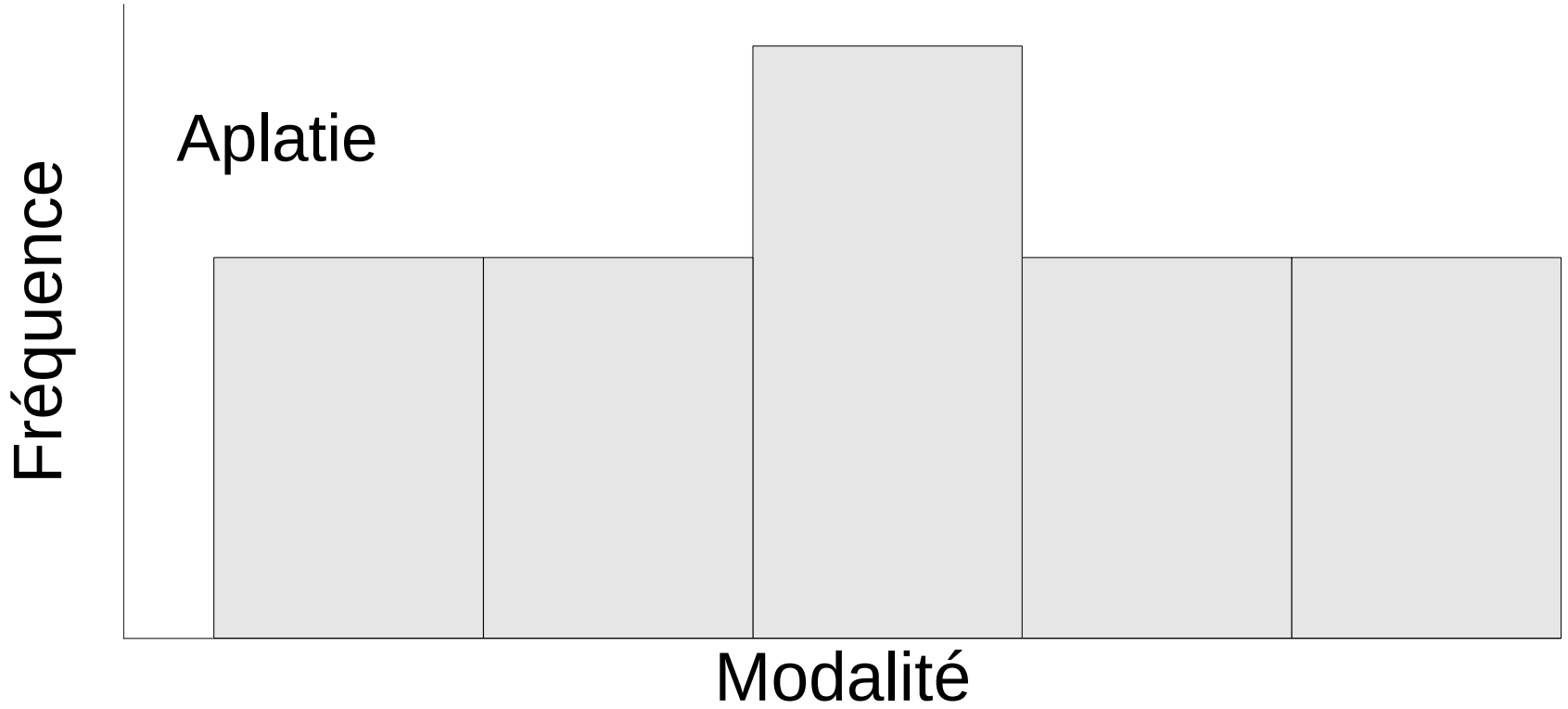

Symétrique
Fréquence
Modalité

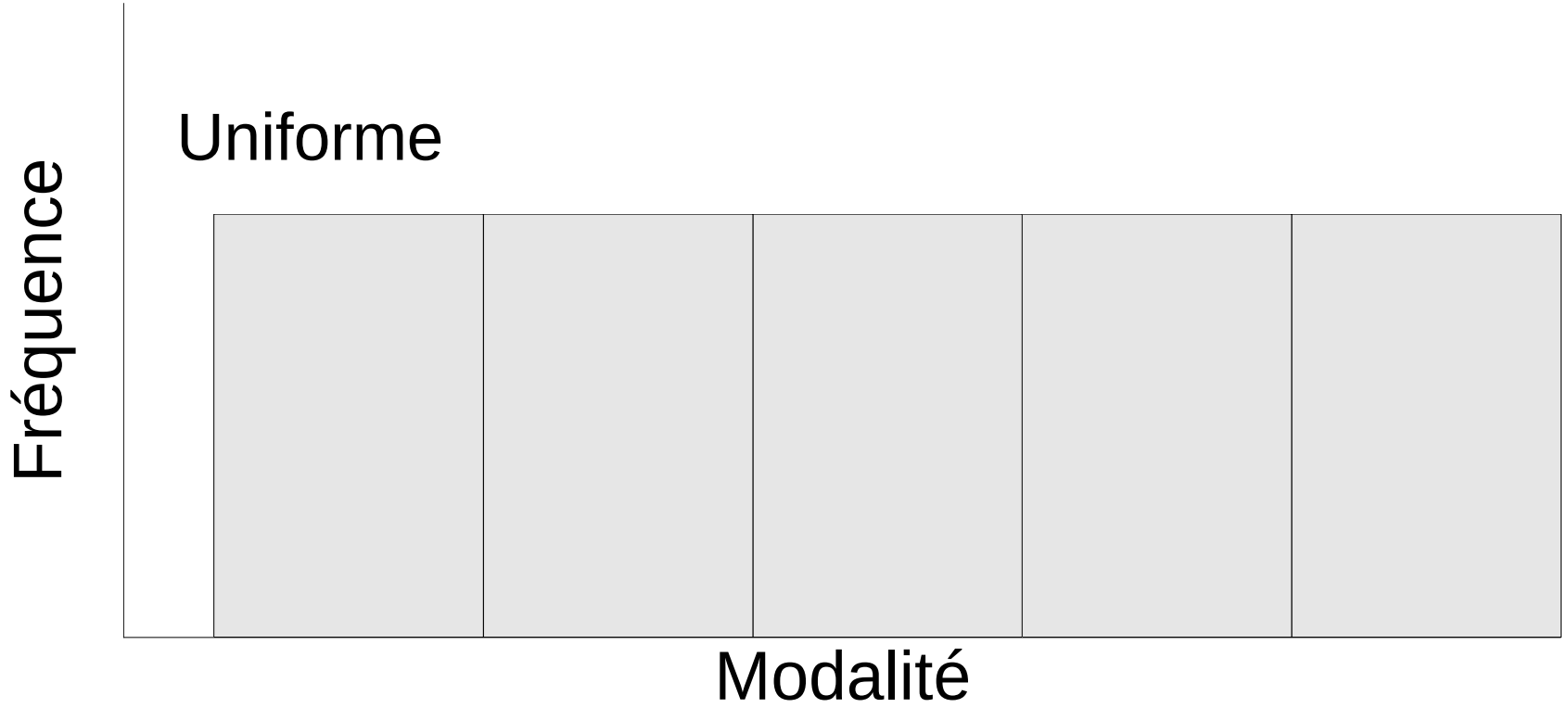

Asymétrique
Fréquence
Modalité


Tendance à gauche
Fréquence
Modalité

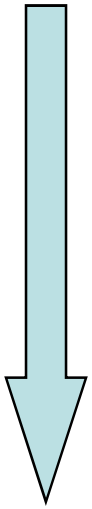

Tendance à droite
Fréquence
Modalité


Tendance unimodale
Fréquence
Modalité


Aplatie
Fréquence
Modalité


Uniforme
Fréquence
Modalité

Quel est le nombre d'individus ayant une valeur (modalité) inférieure à une certaine valeur (par exemple 6)



Fonction cumulative

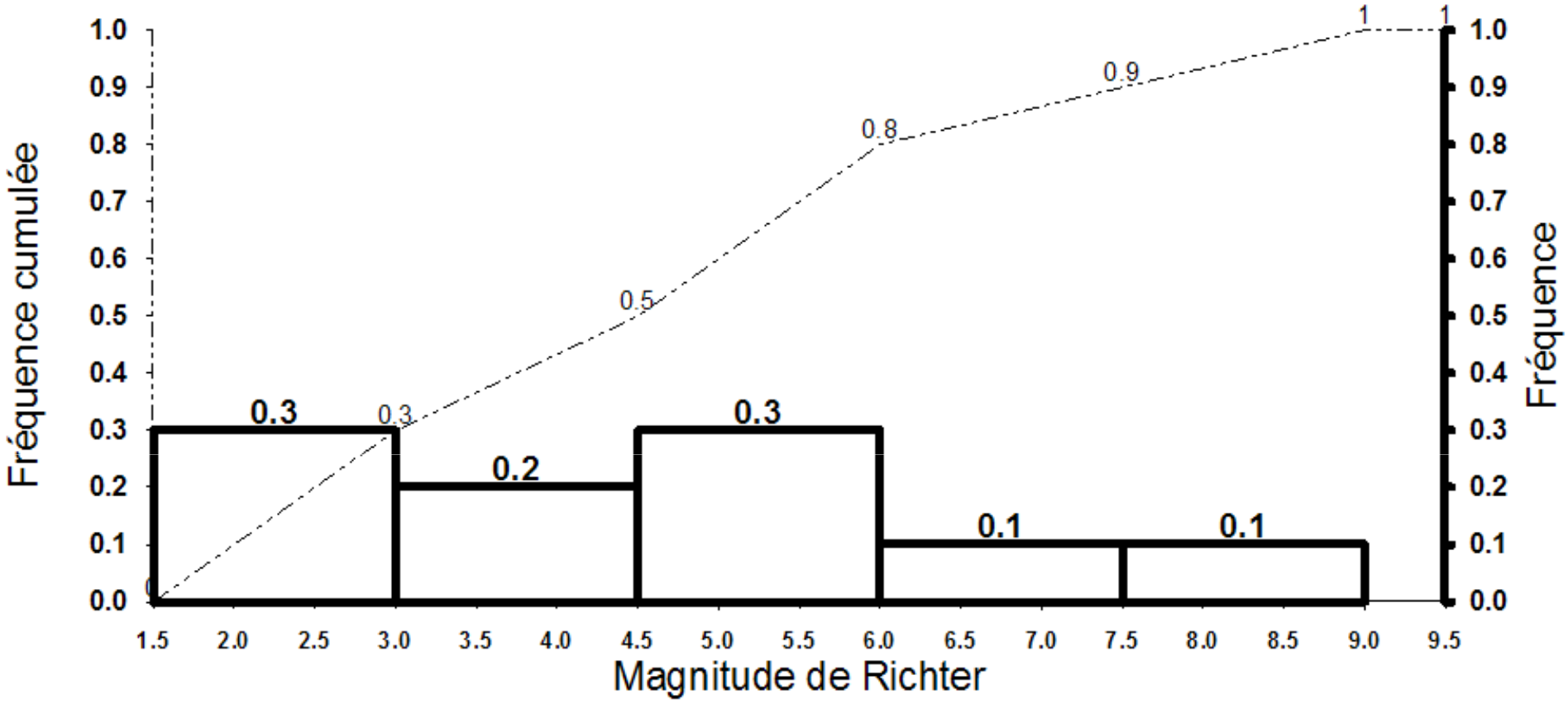
Il suffit pour cela de sommer le nombre de séismes dont la magnitude est inférieure à 6.

Ce nombre divisé par le nombre d'individus de la population s'appelle **Fréquence cumulée**.

Tableau 3.11 Pas=1.5

Classe	Intervalle	Nombre d'individu	Fréquence	Centre de classe	Fréquence cumulée
Classe 1	[1.5,3[	3	0.3	2.25	0.3
Classe 2	[3,4.5[	2	0.2	3.75	0.5
Classe 3	[4.5,6[	3	0.3	5.25	0.8
Classe 4	[6,7.5[	1	0.1	6.75	0.9
Classe 5	[7.5,9[	1	0.1	8.25	1.0

Pour chaque valeur, il existe une **fréquence cumulée**.



Le tracé des couples (valeur,**fréquence cumulée)** donne ce qu'on appelle la ***Fonction cumulative***.

Déterminer le nombre de séismes dont la magnitude est inférieure à 6.0.

Séisme n°	Magnitude (Modalité)
1	2.2
2	3.1
3	5.5
4	7.5
5	2.2
6	1.5
7	4.3
8	6.8
9	4.9
10	4.7

8 sur 10
soit 0.80 ou
80%

Question :

Déterminer le nombre de séismes dont la magnitude est inférieure à 4.5.

Séisme n°	Magnitude (Modalité)
1	2.2
2	3.1
3	5.5
4	7.5
5	2.2
6	1.5
7	4.3
8	6.8
9	4.9
10	4.7

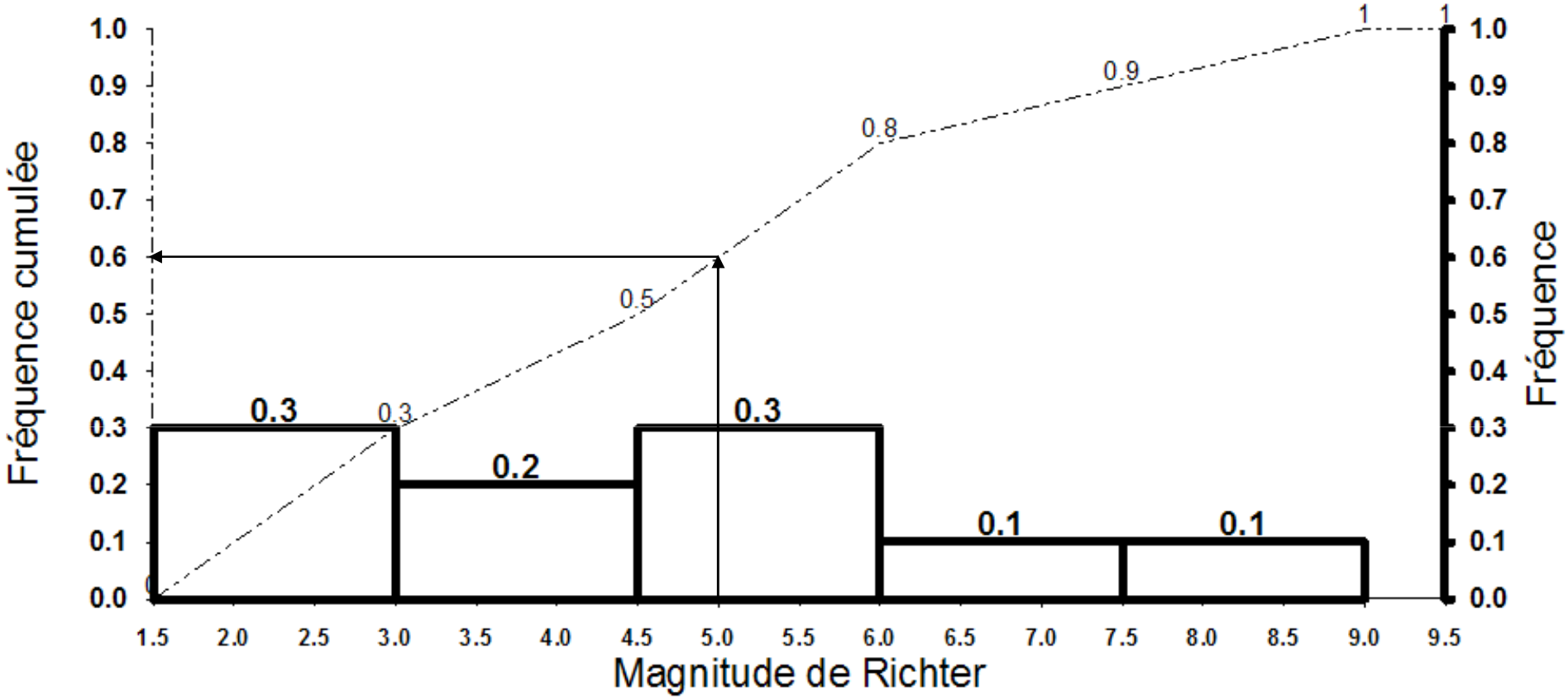
5 sur 10 soit 0.50 ou 50%

Question :

Déterminer le nombre de séismes dont la magnitude est inférieure à 5.0.

Séisme n°	Magnitude (Modalité)
1	2.2
2	3.1
3	5.5
4	7.5
5	2.2
6	1.5
7	4.3
8	6.8
9	4.9
10	4.7

7 sur 10
soit 0.70 ou
70%



Soit 0.60 ou **60%**

7 sur 10 soit
0.70 ou **70%**

3.1 Description graphique

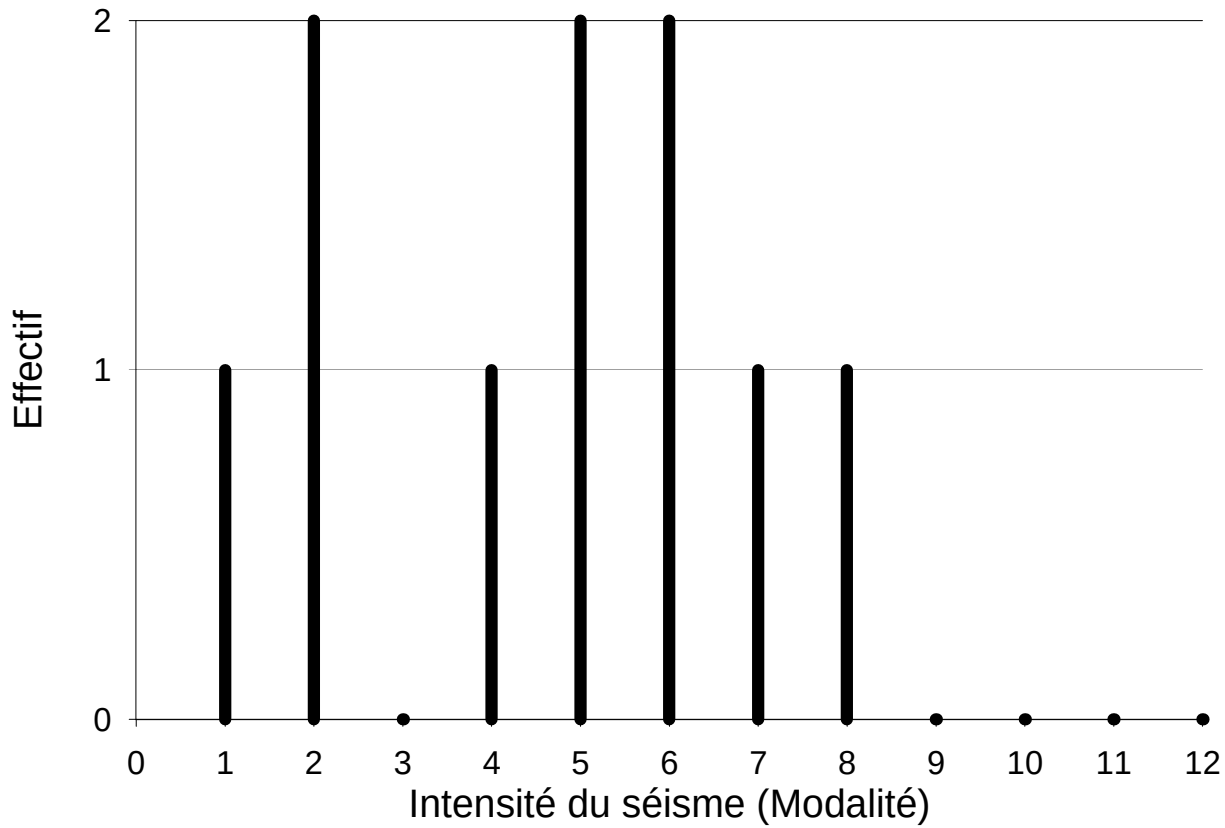
3.1.2 Variable quantitative discrète

La série statistique

Séisme n°	Intensité (Modalité)
1	2
2	4
3	6
4	8
5	2
6	1
7	6
8	7
9	5
10	5

La distribution statistique

Intensité (Modalité)	Effectif	Fréquence de la modalité
1	1	0.1
2	2	0.2
3	0	0
4	1	0.1
5	2	0.2
6	2	0.2
7	1	0.1
8	1	0.1
9	0	0
10	0	0
11	0	0
12	0	0

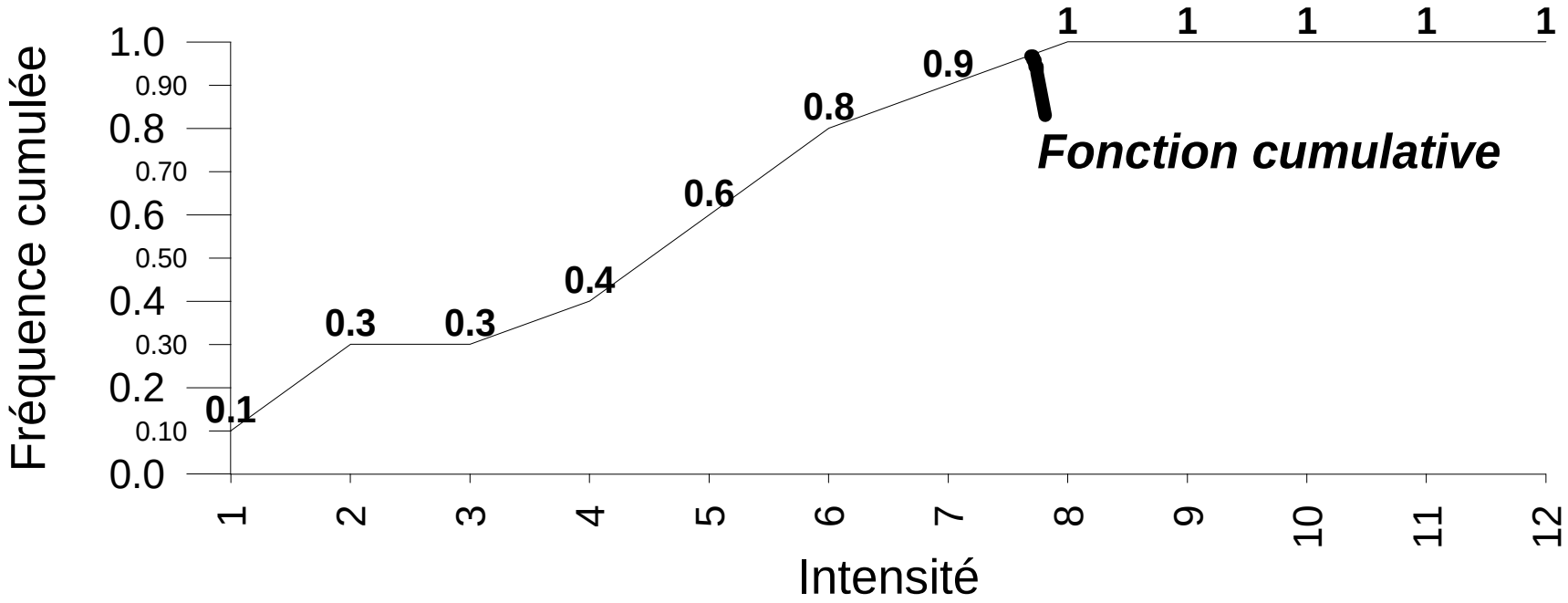


Représentation par bâtonnet

Intensité (Modalité)	Effectif	Fréquence de la modalité	Fréquence cumulée
1	1	0.1	0.1
2	2	0.2	0.3
3	0	0	0.3
4	1	0.1	0.4
5	2	0.2	0.6
6	2	0.2	0.8
7	1	0.1	0.9
8	1	0.1	1
9	0	0	1
10	0	0	1
11	0	0	1
12	0	0	1
	10	1	

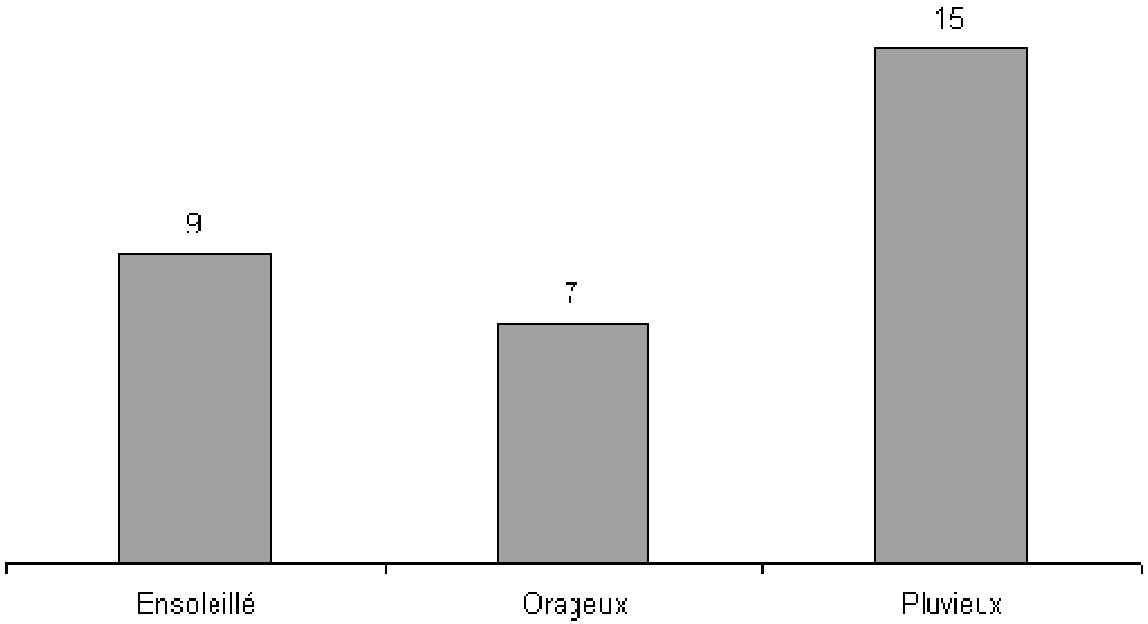
Le nombre de séismes dont l'intensité est inférieure à 5.

Ce nombre divisé par le nombre d'individus de la population s'appelle Fréquence cumulée.

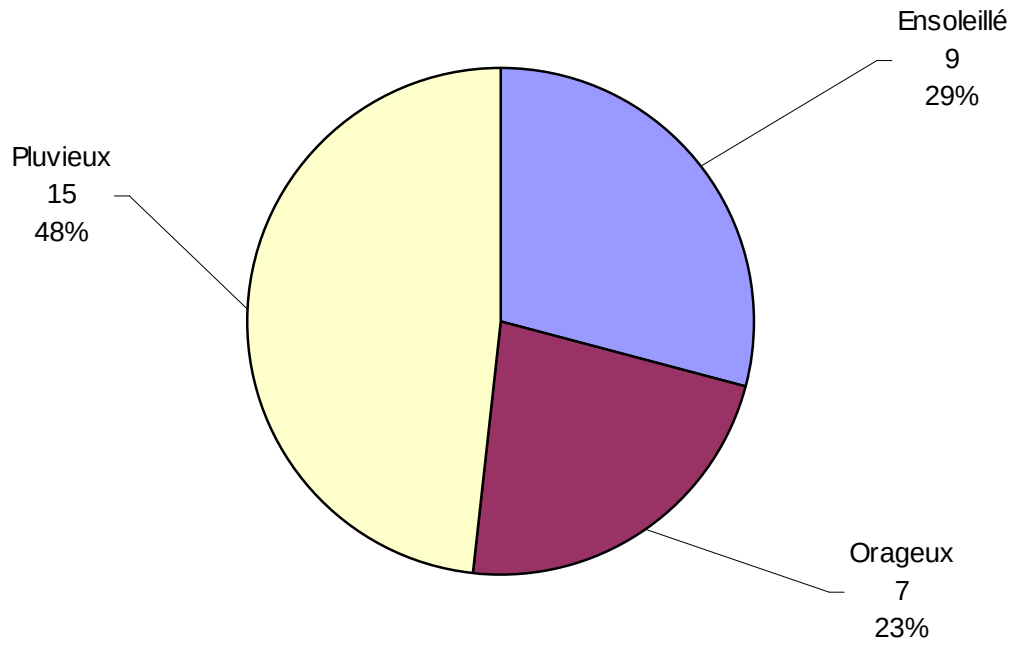

Fréquence cumulée
1.0
0.90
0.8
0.70
0.6
0.50
0.4
0.30
0.2
0.10
0.0
0.1
0.3
0.3
0.4
0.6
0.8
0.9
1
1
1
1
1
Fonction cumulative
1
2
3
4
5
6
7
8
9
10
11
12
Intensité

3.1 Description graphique

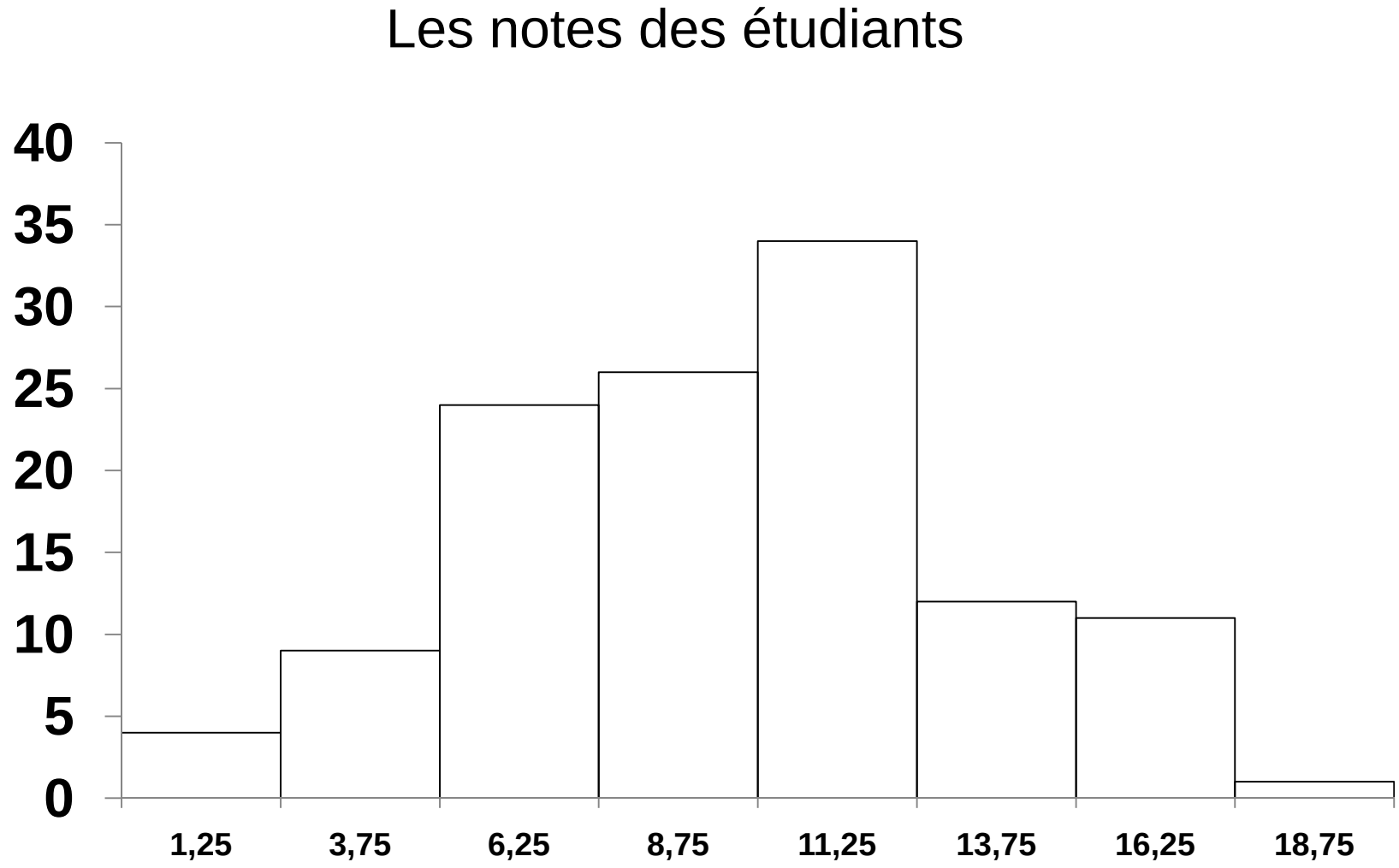
3.1.3 Variable qualitative



Représentation par diagramme



Représentation Diagrammes circulaires


Les notes des étudiants
40
35
30
25
20
15
10
5
0
1,25
3,75
6,25
8,75
11,25
13,75
16,25
18,75

Est-ce la description graphique est suffisante ?

Les modèles mathématiques nécessitent l'introduction d'une valeur et non l'appréciation sur la variation des modalités

3.2 Description numérique

3.2.1 Caractéristique de tendance centrale

Exemple: X=F/K

F: Force (mesurée en Newton) et K: Raideur: Connue=10 N/cm

F: Force (mesurée en Newton)

10	10.3	22.1	23	26	15.2	12.3	18	18.3	14.5

$$n = 10$$

$$k = 1 + 3.3\log_{10}(10) = 4.33$$

Individus	Force (Modalité) N
1	10
2	10.3
3	22.1
4	23
5	26
6	15.2
7	12.3
8	18
9	18.3
10	14.5

$$k = 5$$

$$V_{\min} = 10 \text{ N}$$

$$V_{\max} = 26 \text{ N}$$

$$Pas = \frac{26 - 10}{5} = 3.2 \text{ N}$$

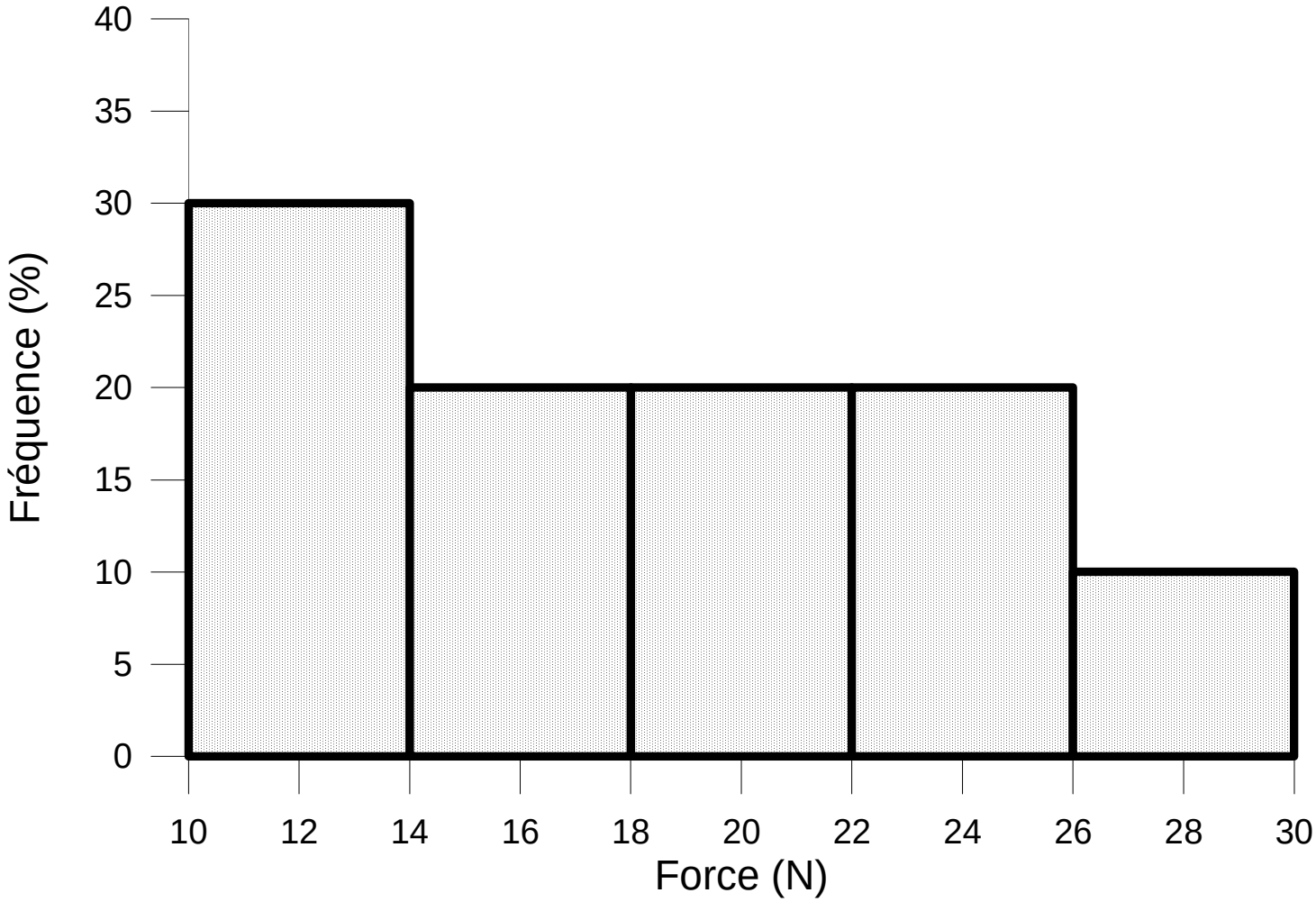
$$Pas = 4 \text{ N}$$

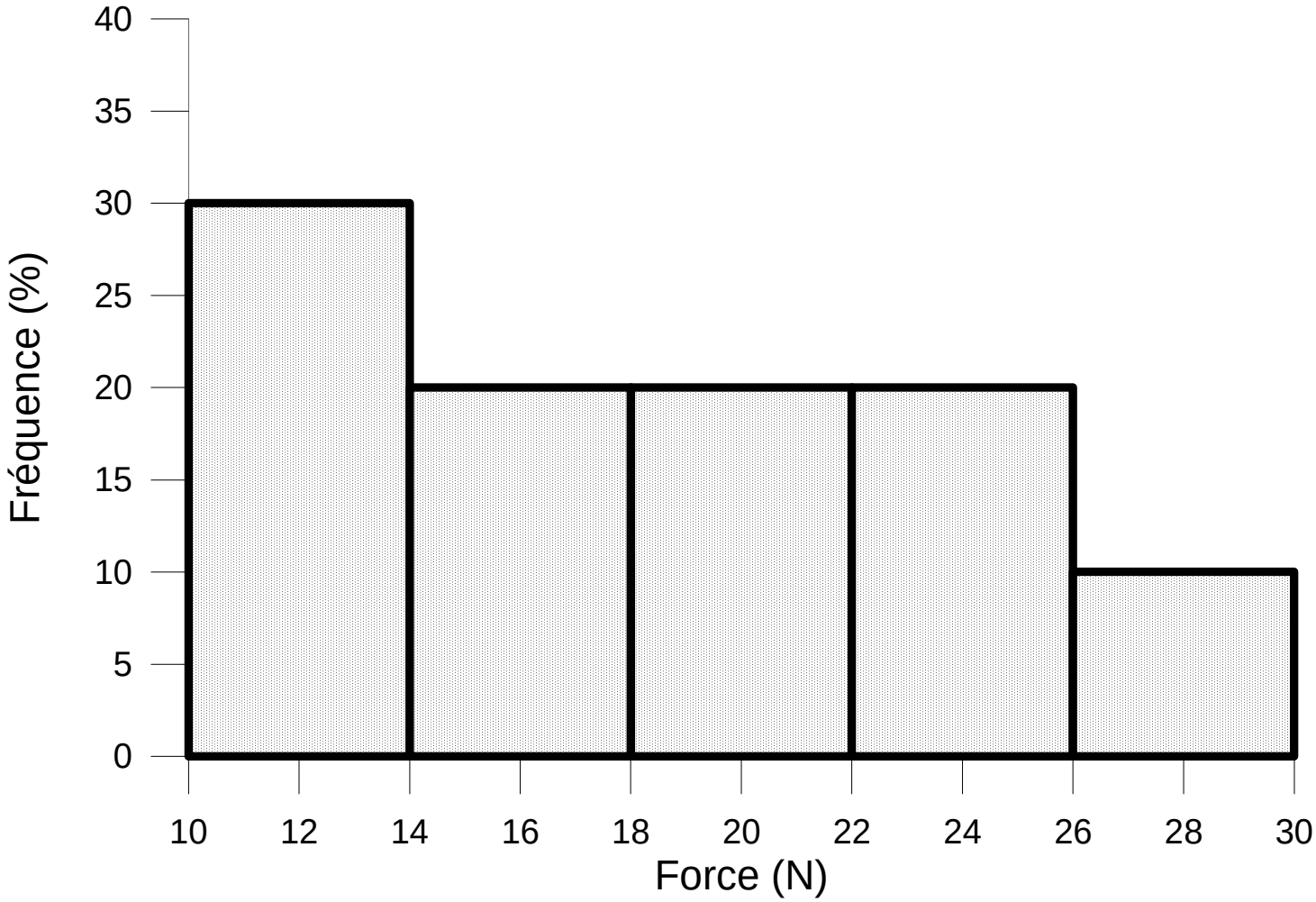
Individus	Force (Modalité) en Newton (N)
1	10
2	10.3
3	22.1
4	23
5	26
6	15.2
7	12.3
8	18
9	18.3
10	14.5

Limites des classes (N)		Centre de classe (N)	Effectif	Fréquence (%)
Limite inf	Limite sup			
10	14	12	3	30
14	18	16	2	20
18	22	20	2	20
22	26	24	2	20
26	30	28	1	10

Individus	Force (Modalité) en Newton (N)
1	10
2	10.3
3	22.1
4	23
5	26
6	15.2
7	12.3
8	18
9	18.3
10	14.5

Individus	Force (Modalité) en Newton (N)
1	12
2	12
3	24
4	24
5	28
6	16
7	12
8	20
9	20
10	16


40
35
30
25
20
15
10
5
0
Fréquence (%)
10
12
14
16
18
20
22
24
26
28
30
Force (N)


40
35
30
25
20
15
10
5
0
Fréquence (%)
10
12
14
16
18
20
22
24
26
28
30
Force (N)

10 valeurs de F sont observées.

Question:

Quelle valeur de F doit-on utiliser dans l'équation X=F/K

La plus petite ?
La plus grande?

La plus significative

Que recherche le concepteur

- Réduire un ensemble de données
- Conserver une partie de l'information

Résumer l'ensemble des valeurs par une seule valeur

Parfois c’est difficile à accepter

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

Caractéristique de tendance centrale

Elle décrit l'ordre de grandeur des valeurs et aussi la valeur centrale autour de laquelle se regroupent les observations.

Caractéristique de tendance centrale (paramètres de position)

Choisir

- Série statistique
- Tableau statistique

1. Mode
2. Médiane
3. Moyenne ($\underline{s}$)
4. Quantile

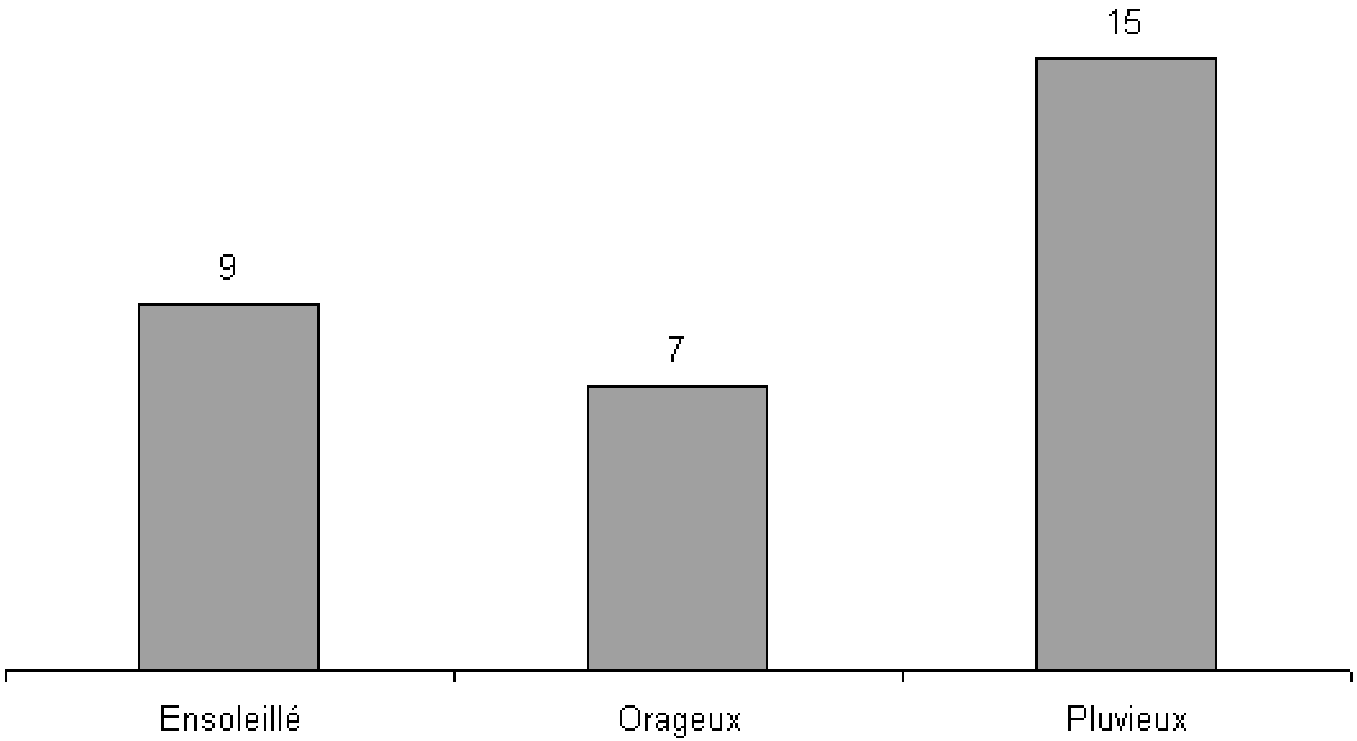
Mode

Le mode, noté *Mo*, est la modalité qui admet **la plus grande fréquence** :

Il est parfaitement défini pour une variable qualitative ou une variable quantitative discrète.

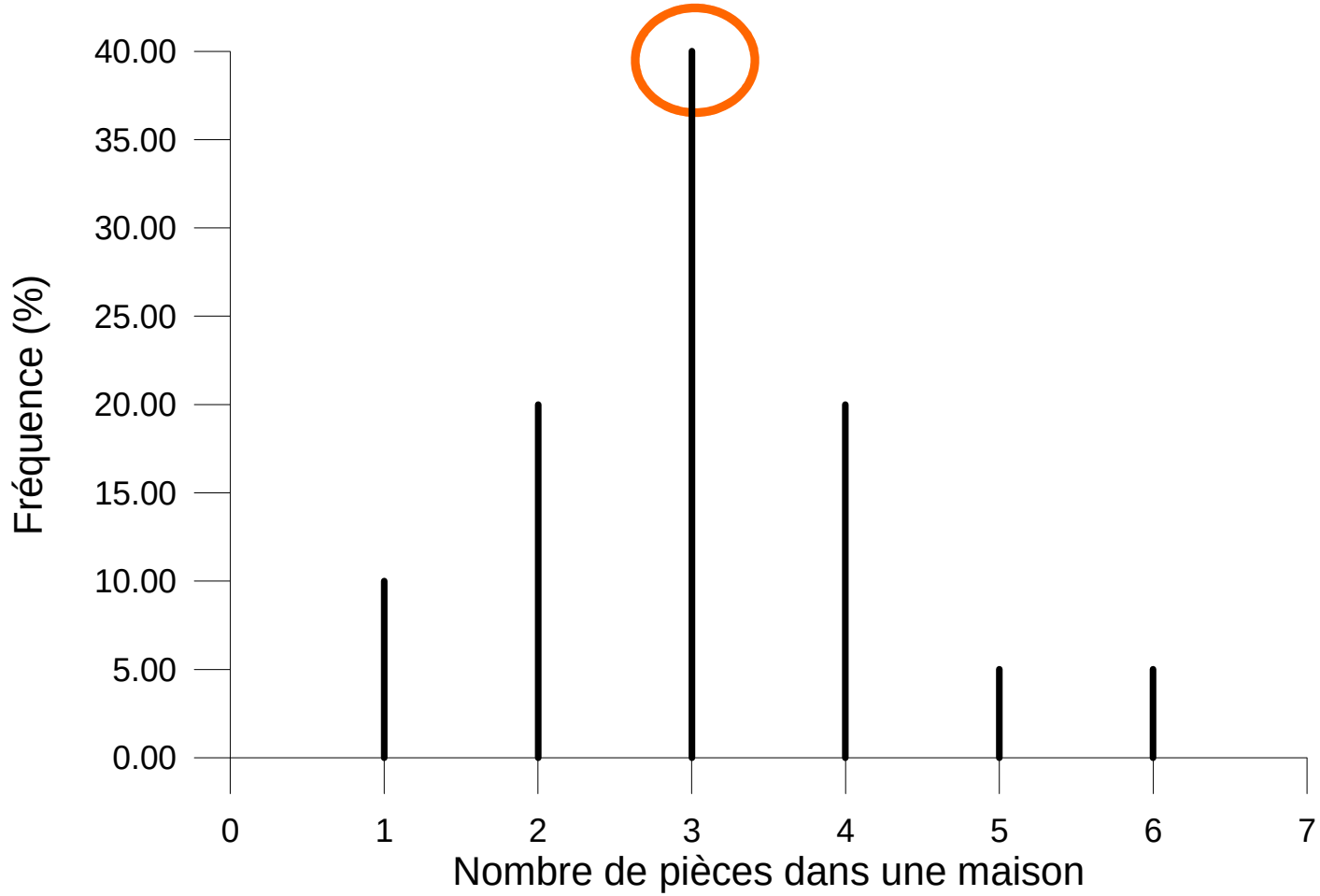
Pour une variable quantitative continue nous parlons de **classe modale** : c'est la classe dont la densité de fréquence est maximum.

Variable qualitative



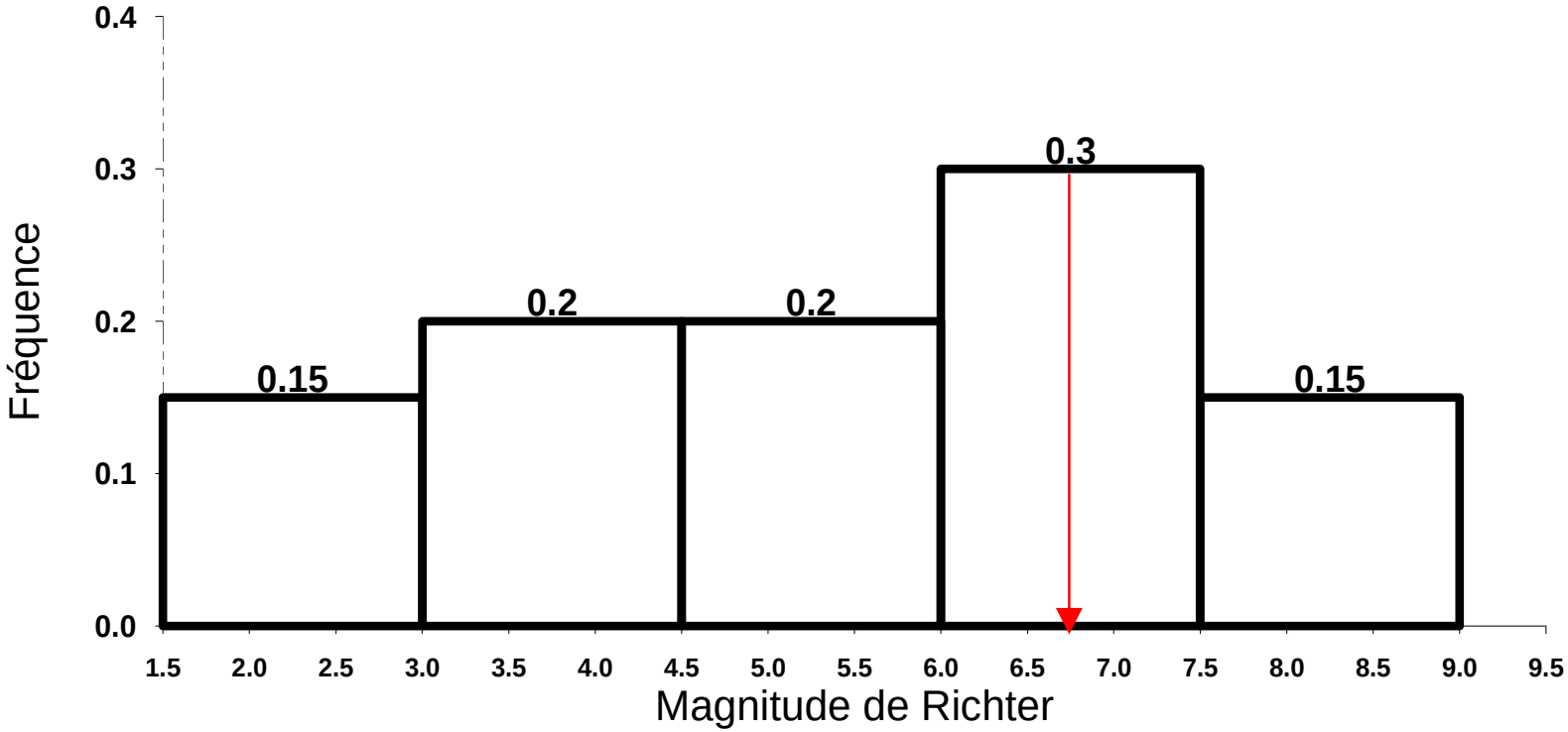
Mo= Pluvieux

Variable quantitative discrète



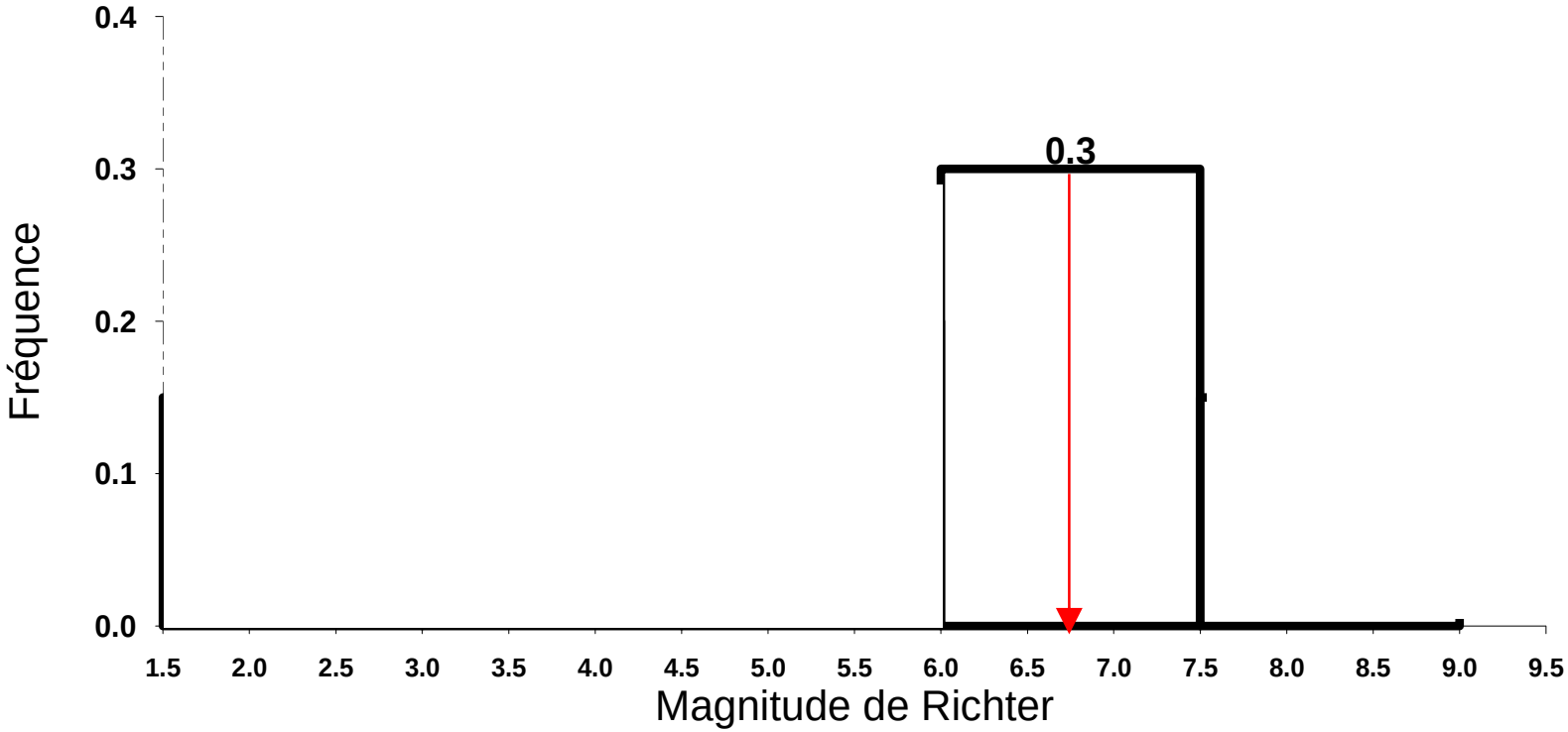
Mo=3 pièces

Variable quantitative continue



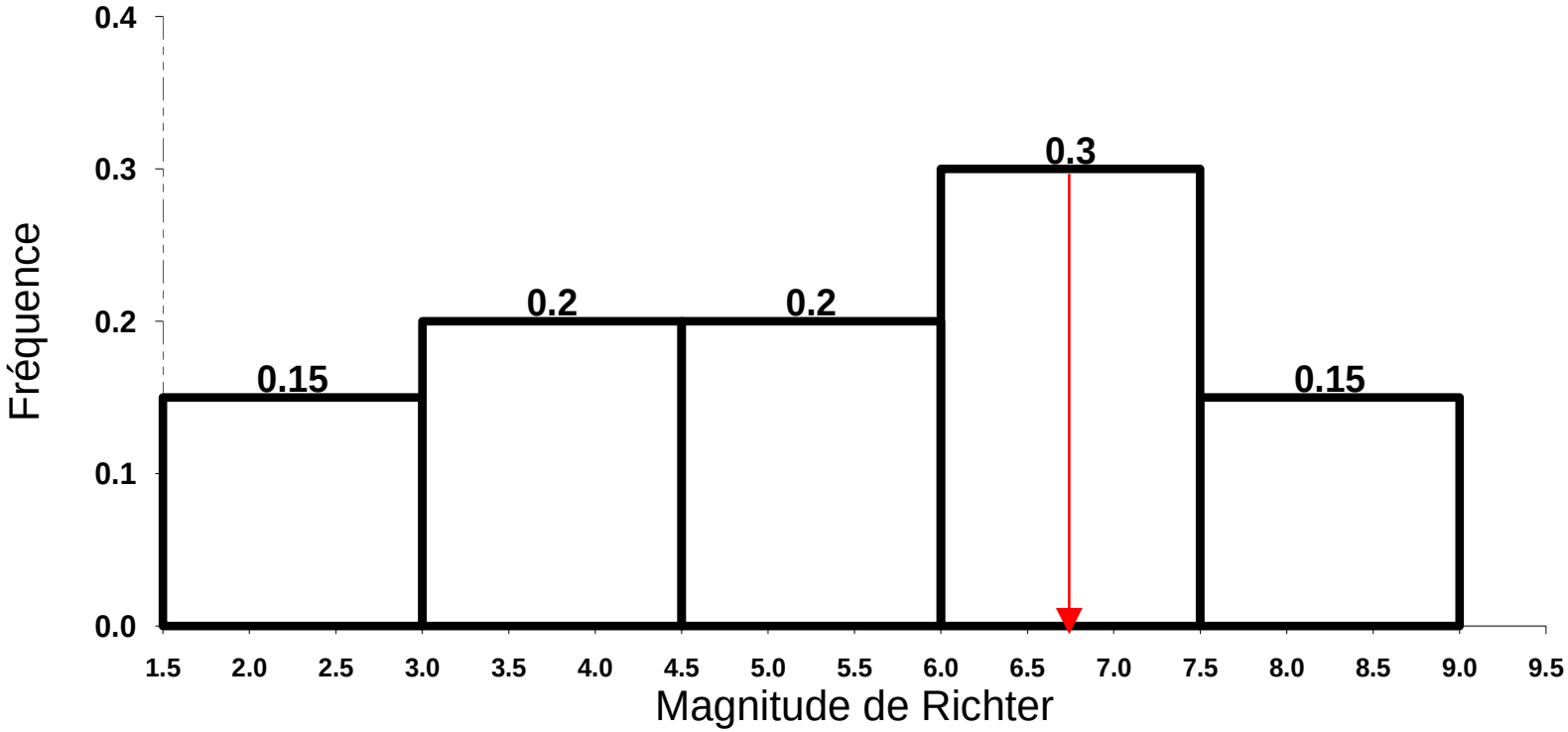
Milieu de la classe dont la fréquence est la plus grande
Mo=6.75

Variable quantitative continue



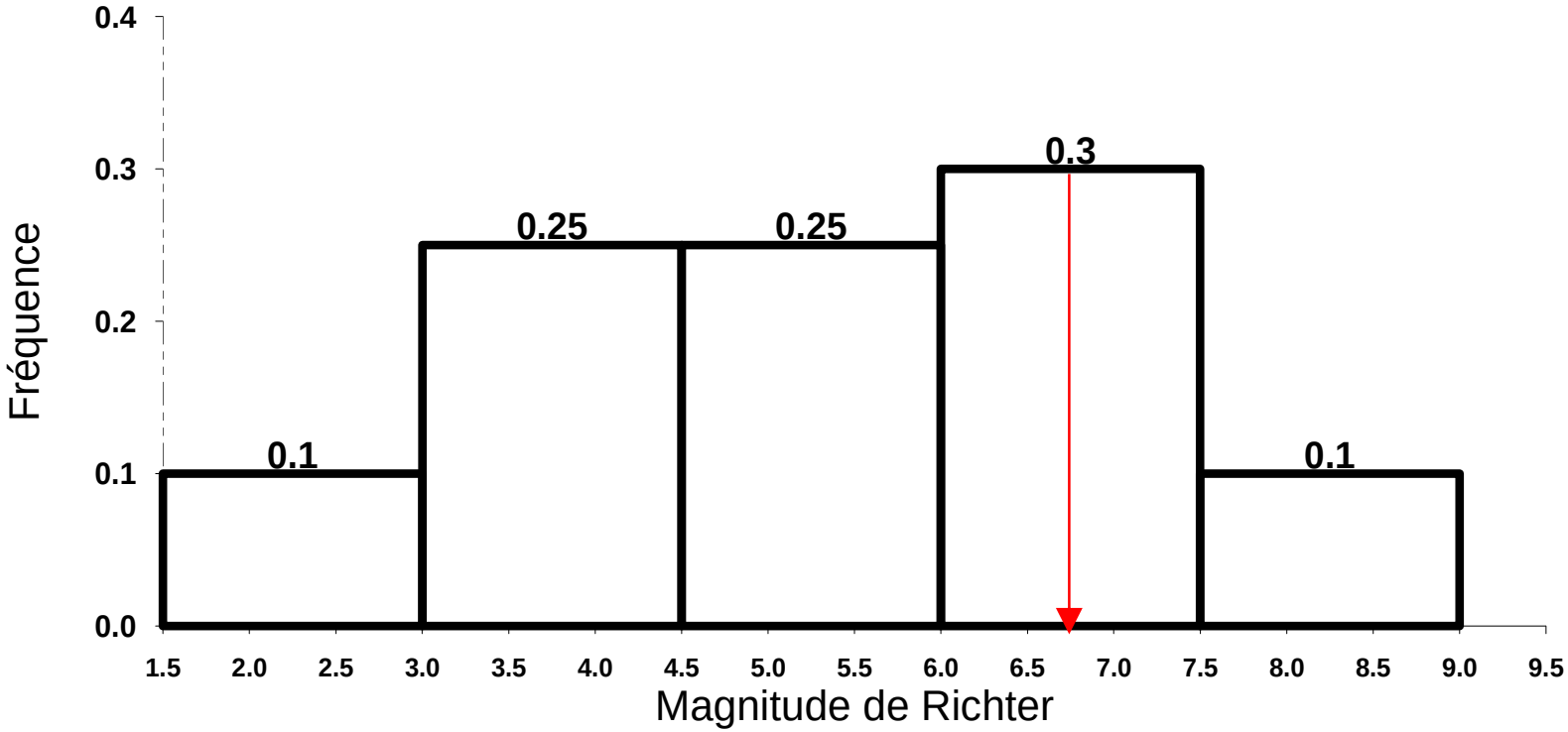
Milieu de la classe dont la fréquence est la plus grande
Mo=6.75

Variable quantitative continue



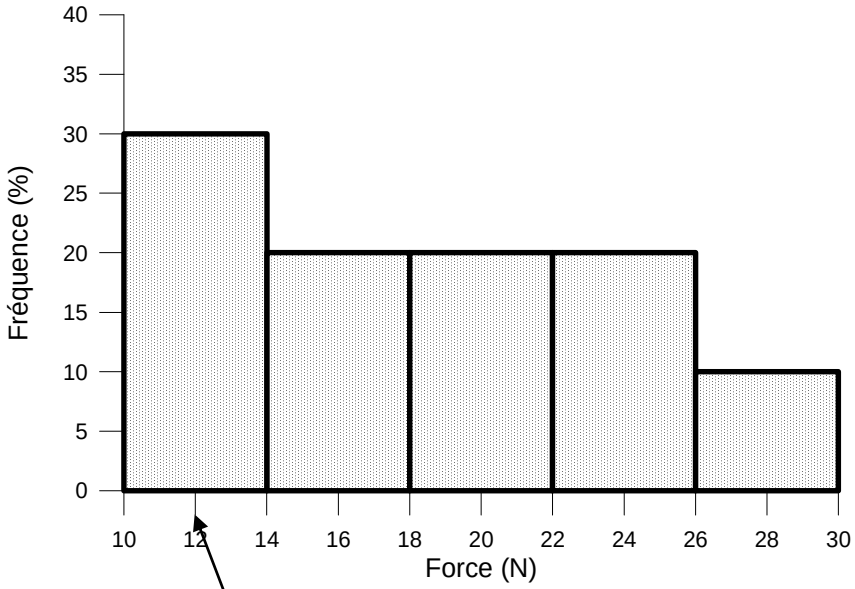
Milieu de la classe dont la fréquence est la plus grande
Mo=6.75

Variable quantitative continue

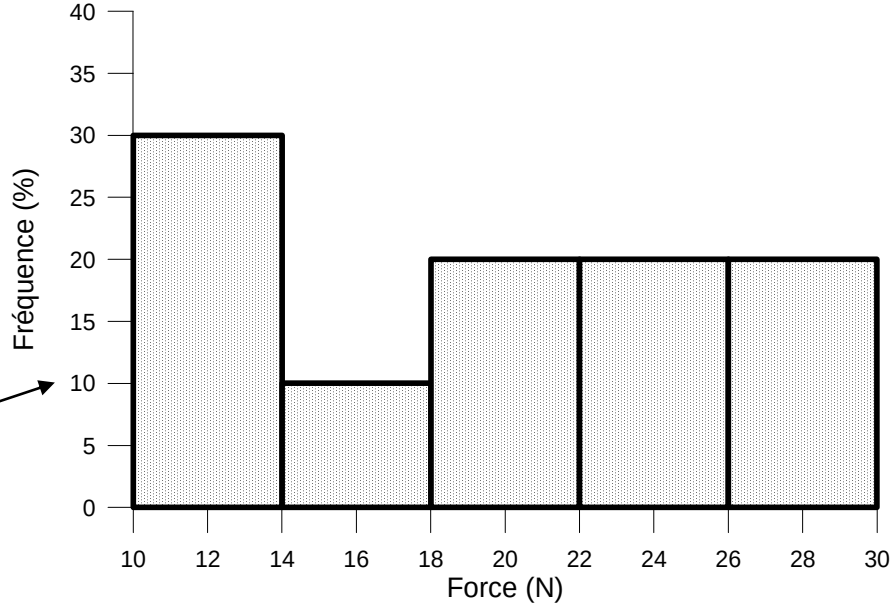


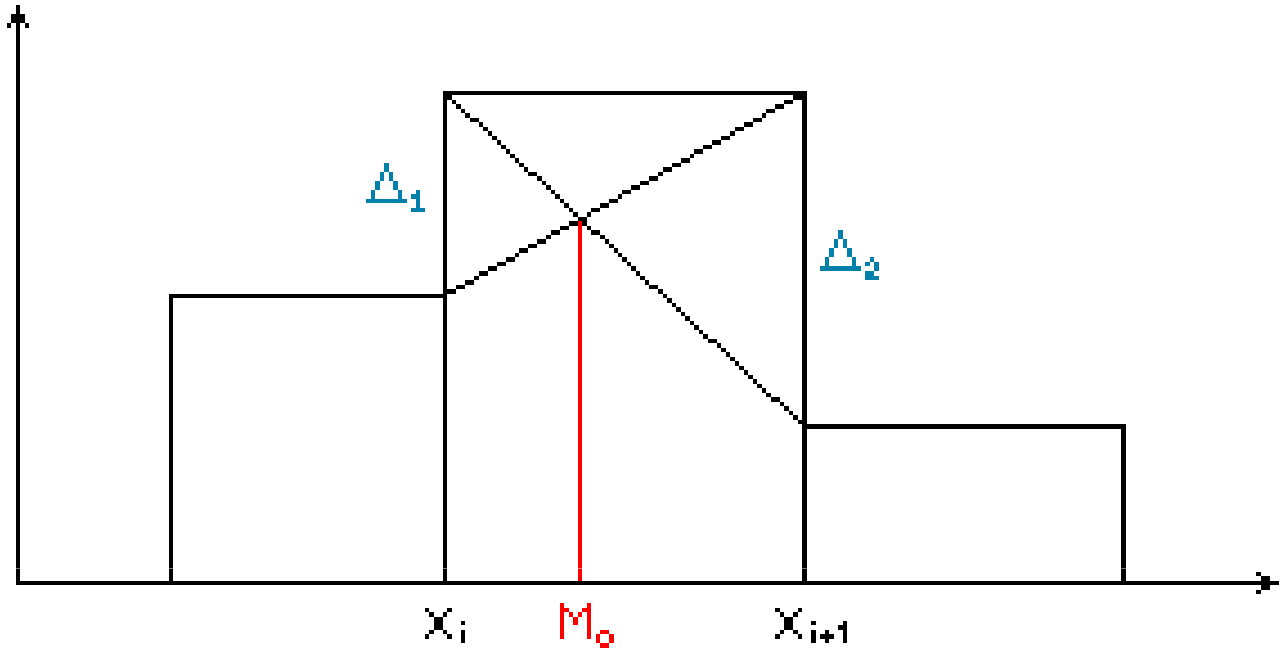
Mo=6.75

Le même résultat a été obtenu
pourtant les deux distributions ne sont pas identiques


40
35
30
25
20
15
10
5
0
Fréquence (%)
10
12
14
16
18
20
22
24
26
28
30
Force (N)

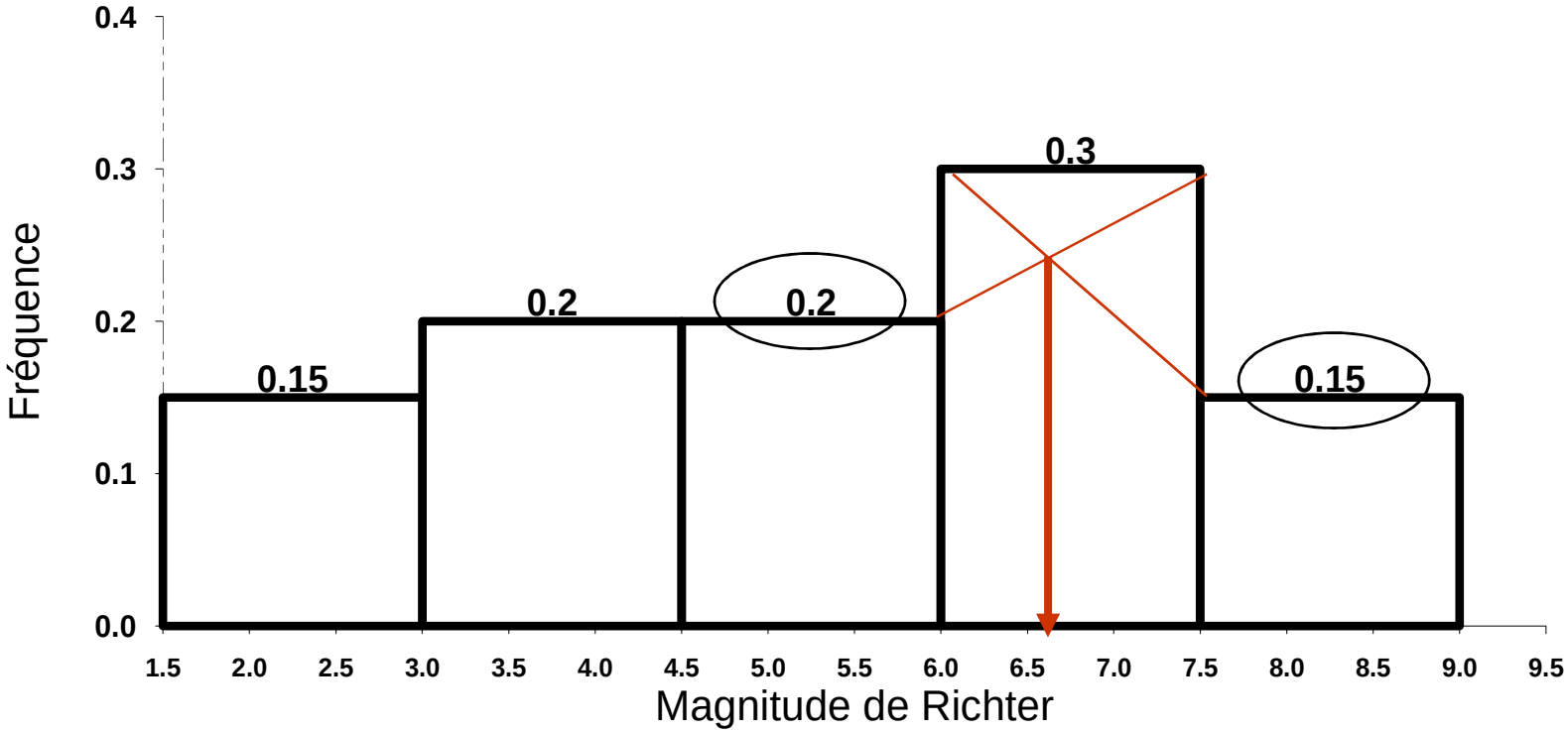
Mode

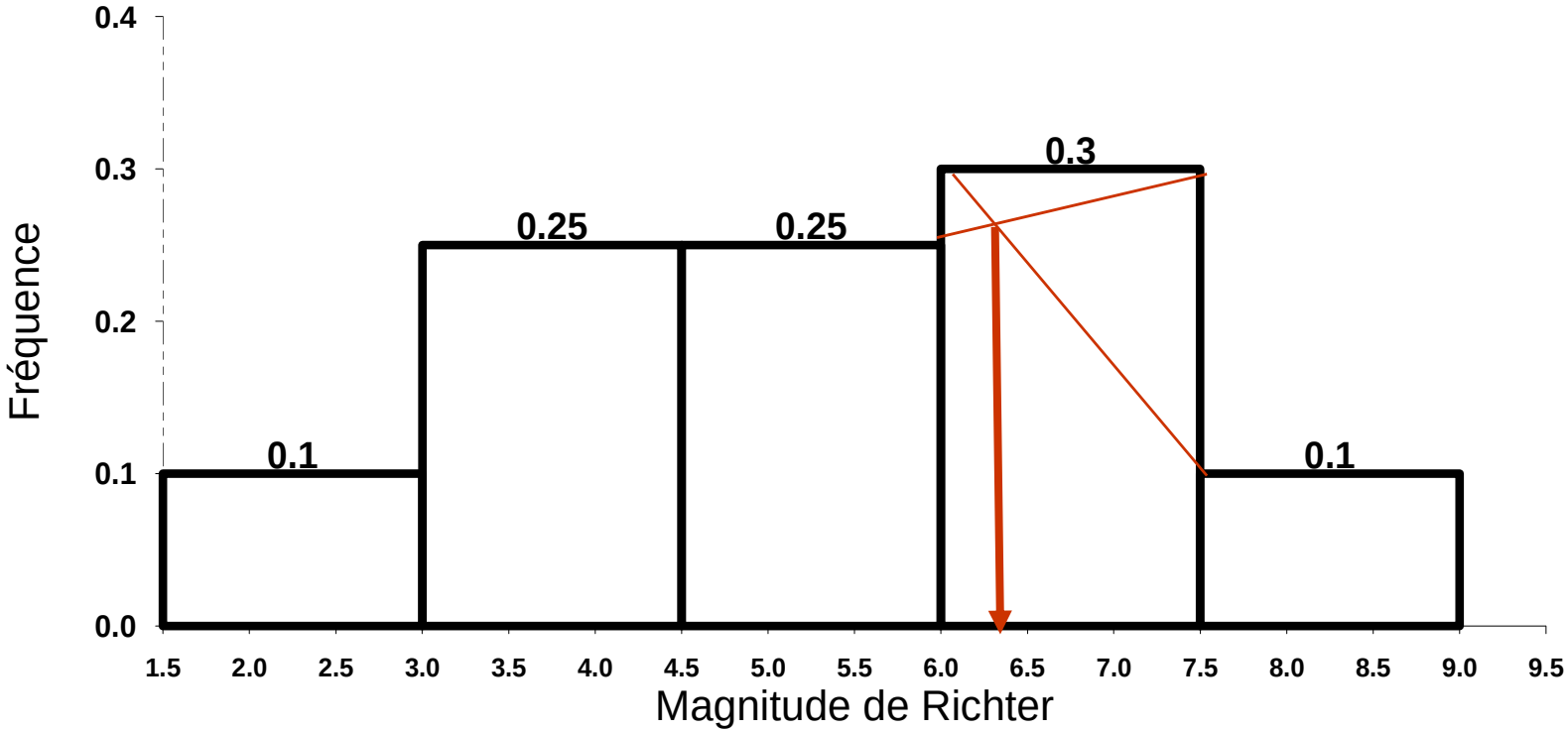

40
35
30
25
20
15
10
5
0
Fréquence (%)
10
12
14
16
18
20
22
24
26
28
30
Force (N)



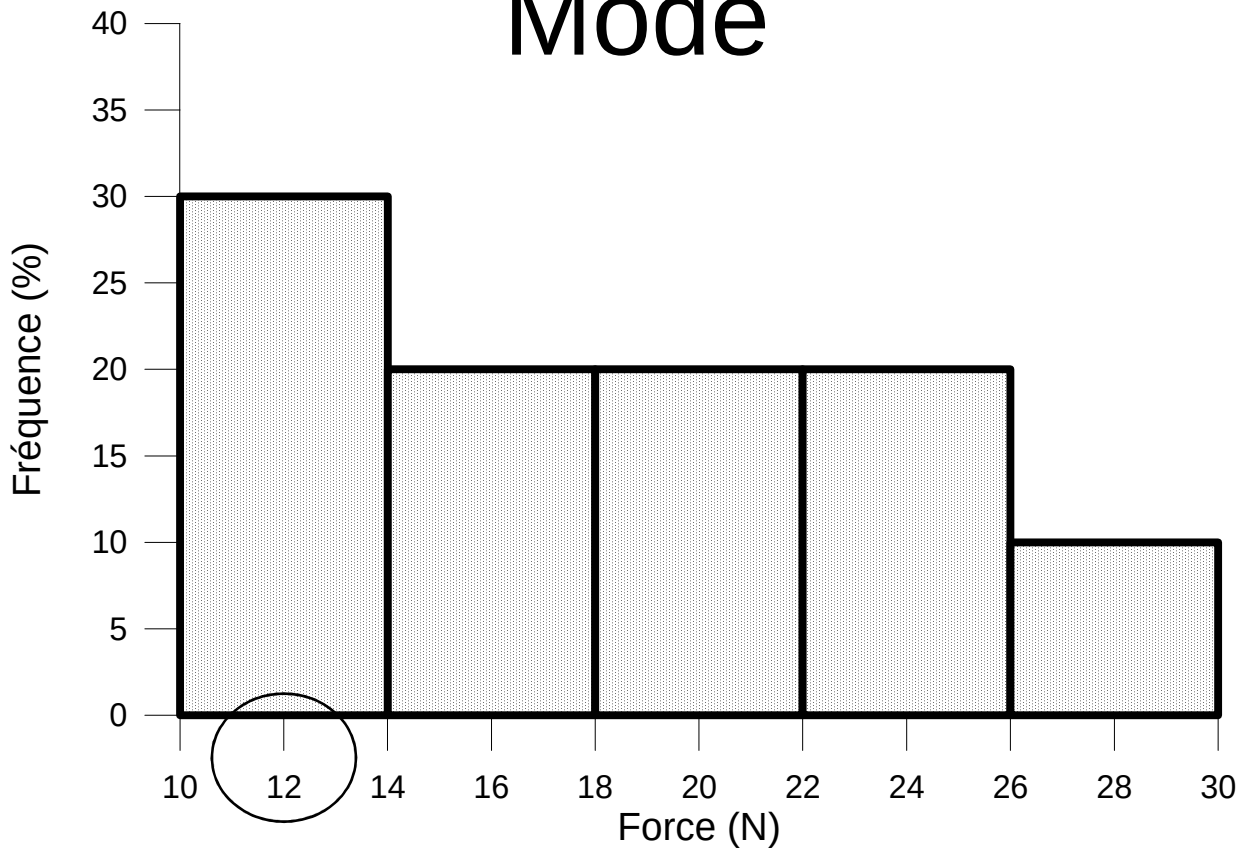
$$\frac{M_o - x_i}{\Delta_1} = \frac{x_{i+1} - M_o}{\Delta_2}$$

$$M_o = x_i + \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i).$$


0.4
0.3
0.2
0.1
0.0
Fréquence
0.15
0.2
0.2
0.3
0.15
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
Magnitude de Richter


0.4
0.3
0.2
0.1
0.0
Fréquence
0.1
0.25
0.25
0.3
0.1
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
Magnitude de Richter

Mode



F=12 N

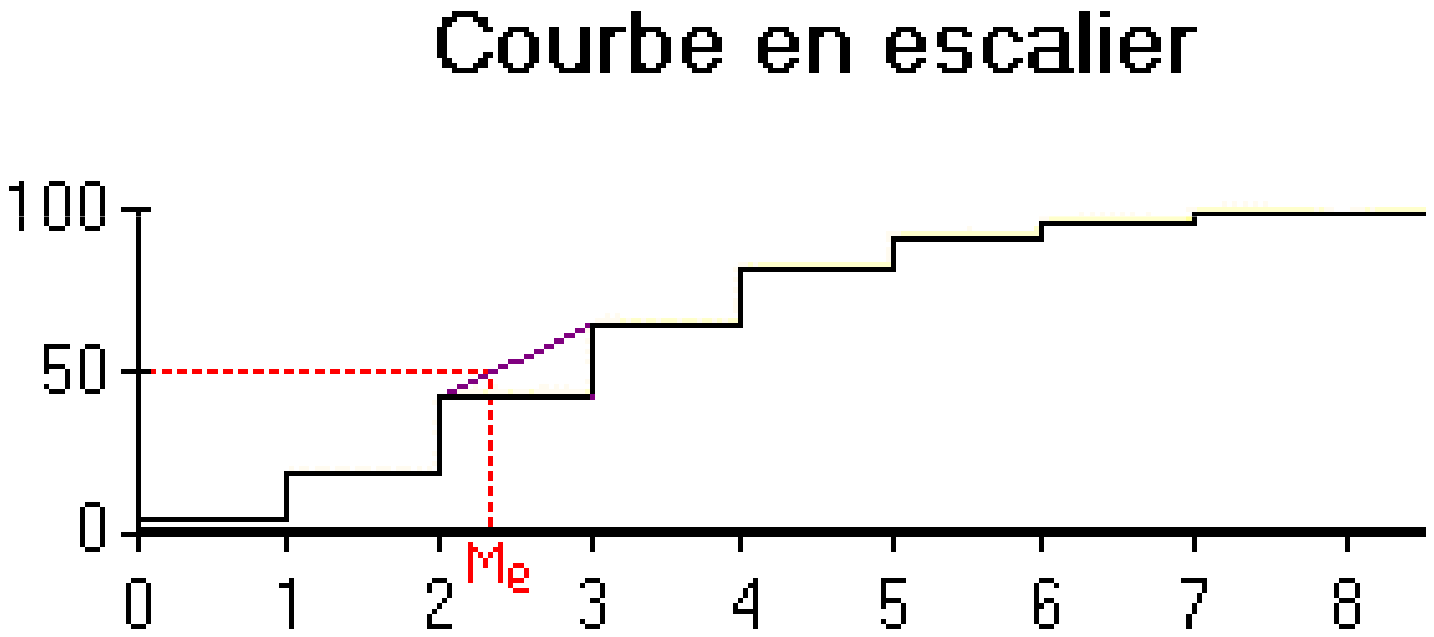
X=12/10=1.2cm

Médiane

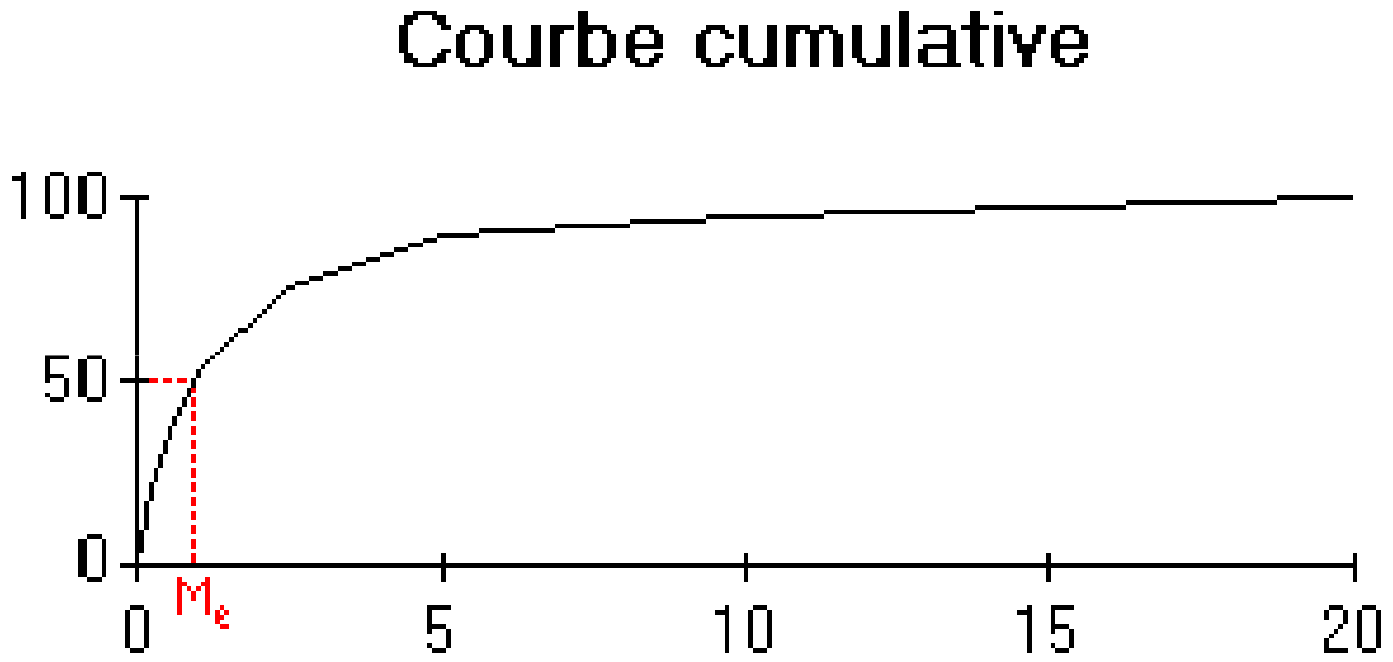
La médiane *Me* est telle que l'effectif des observations dont les modalités sont inférieures à *Me* est égal à l'effectif des observations dont les modalités sont supérieures à *Me*.

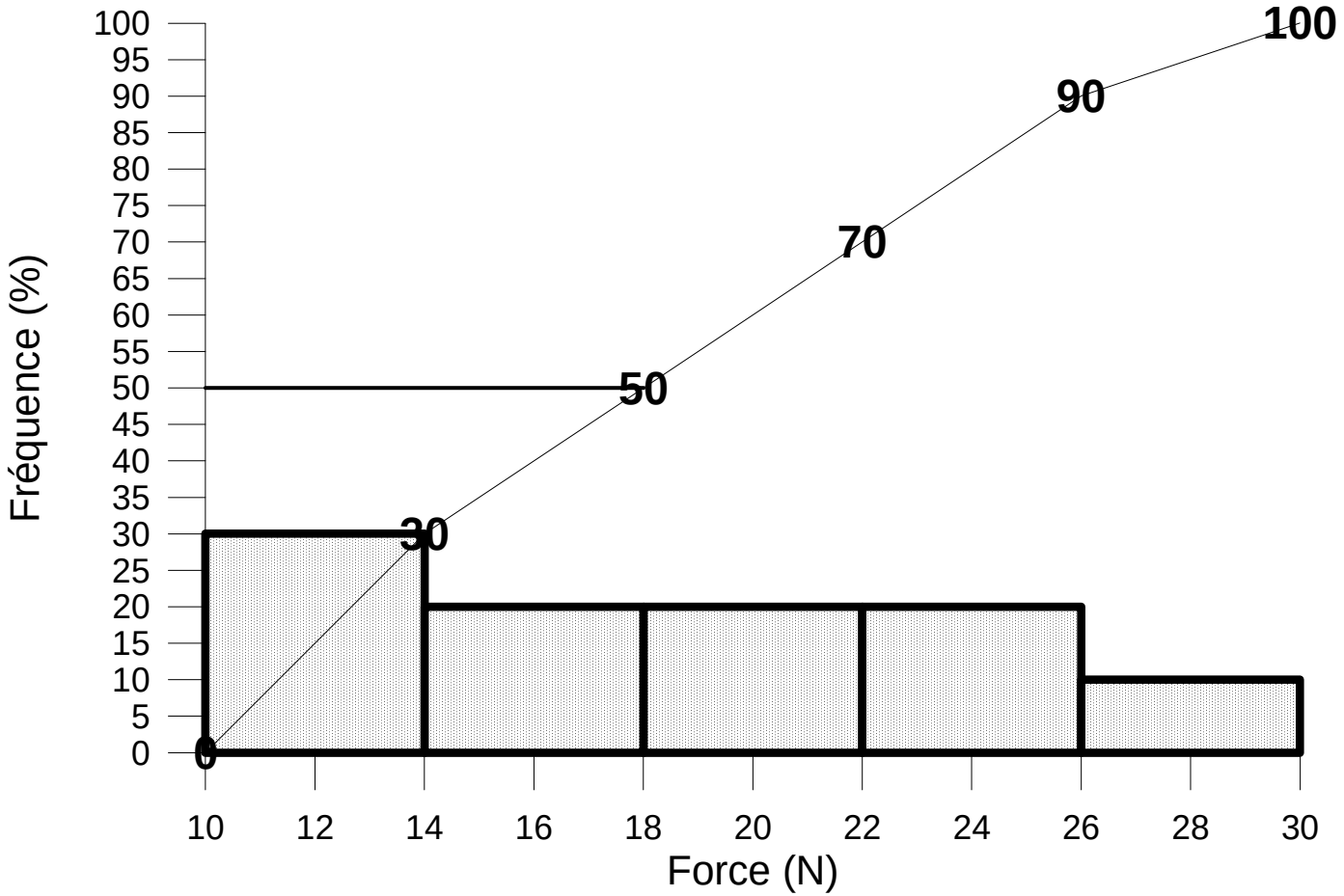
Utilise la notion de fonction cumulative

Quantitative discret



Quantitative continue





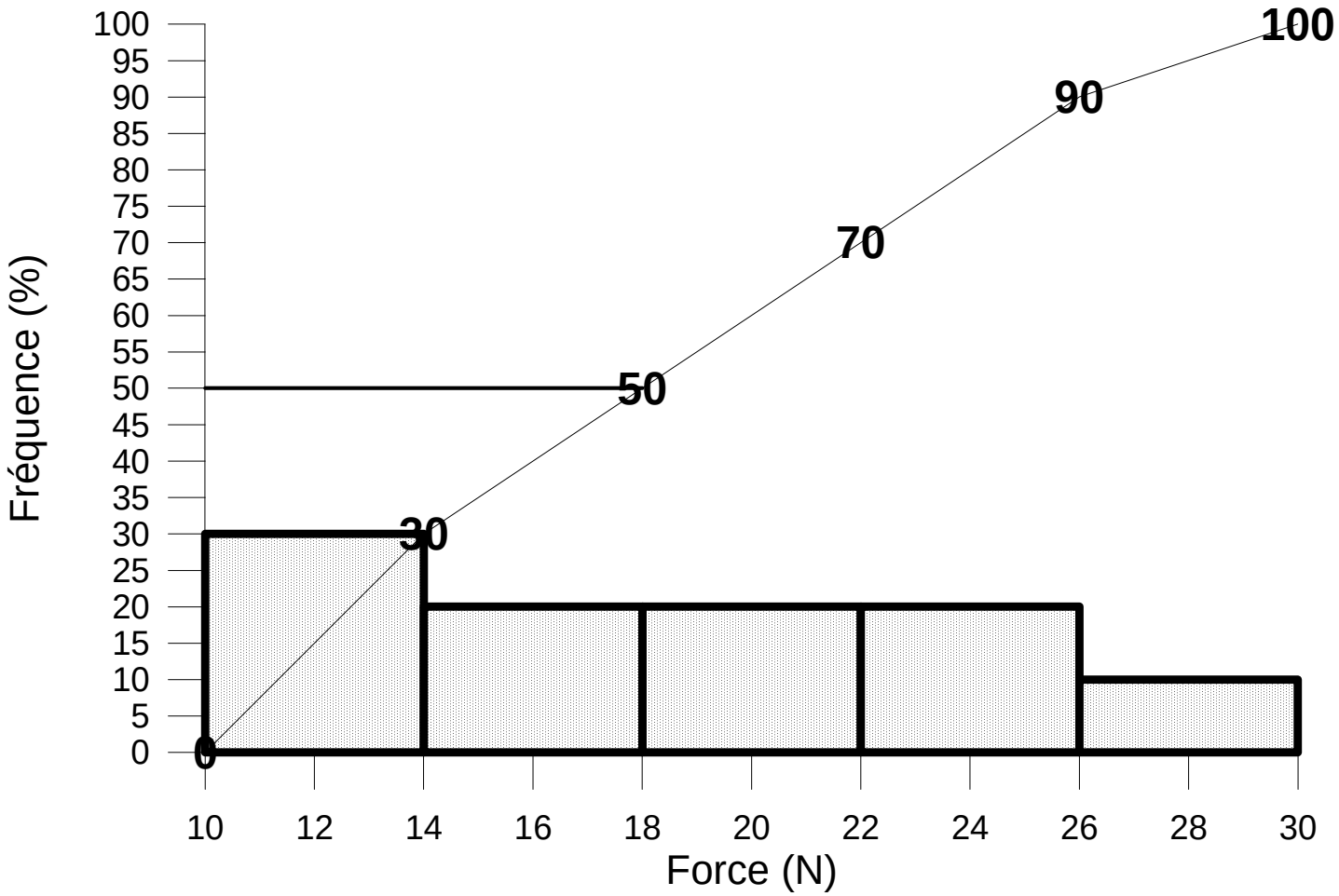
F=18 N

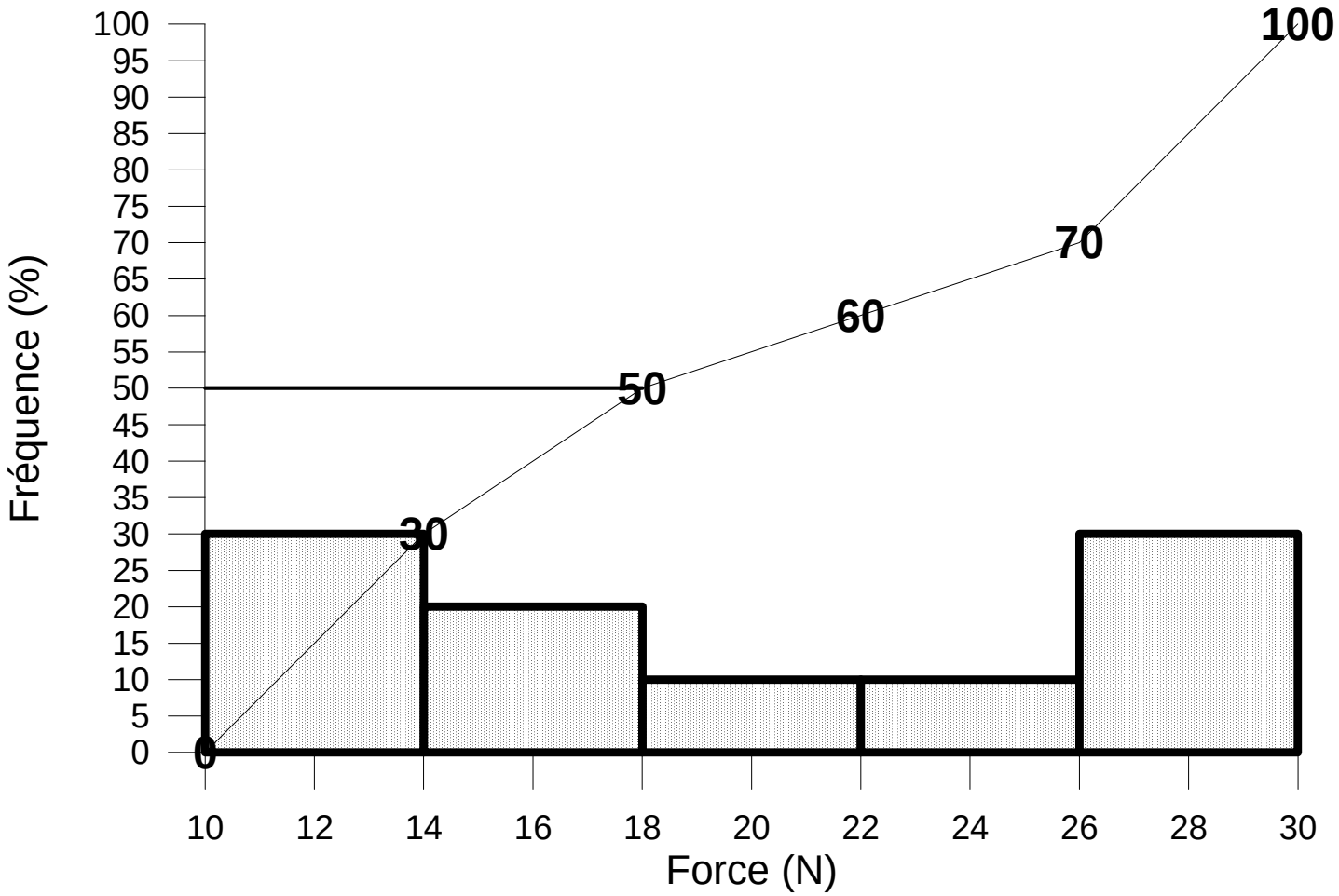
X=18/10=1.8cm

Médiane

Avantage:
•simple à calculer
•Prend en compte une partie des valeurs

Inconvénient:
ne tient pas compte de la distribution des modalités supérieures à la médiane


Fréquence (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
30
50
70
90
100
10
12
14
16
18
20
22
24
26
28
30
Force (N)

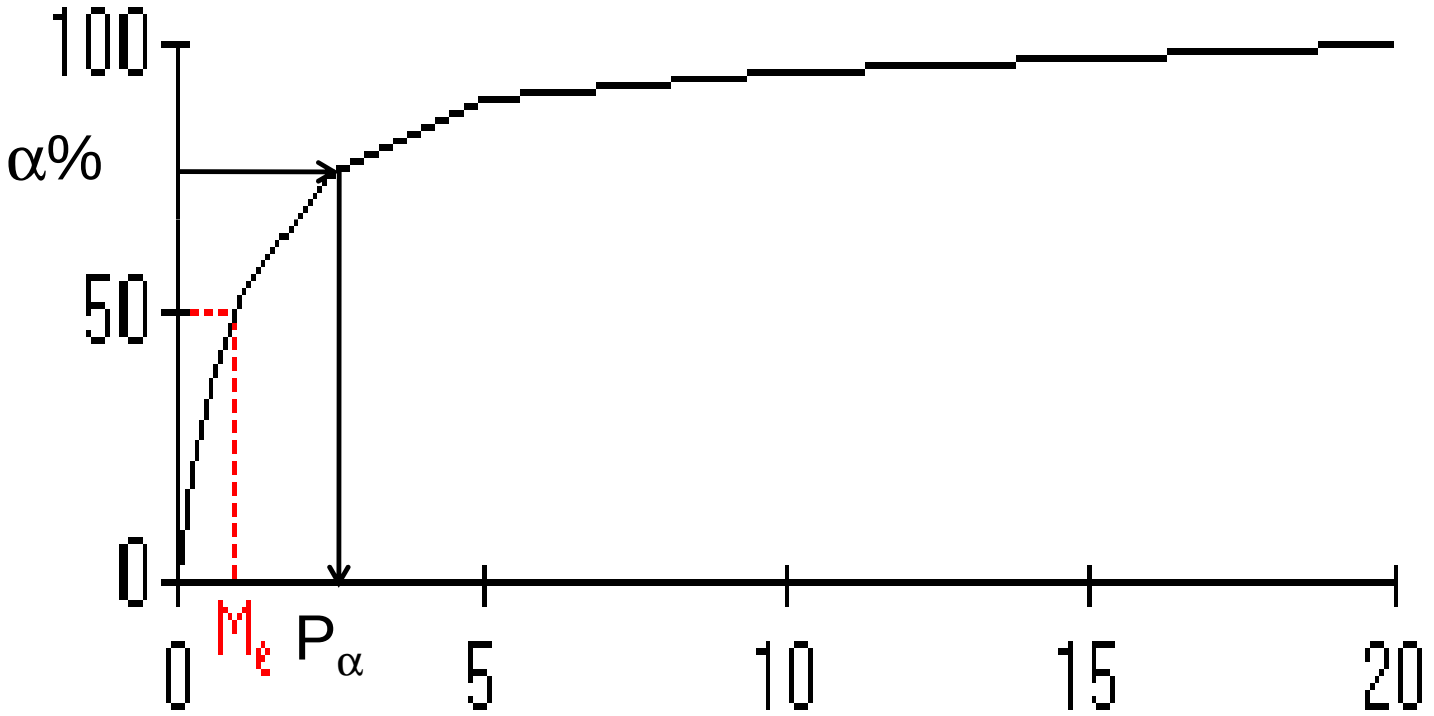

Fréquence (%)
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
10
12
14
16
18
20
22
24
26
28
30
Force (N)
0
30
50
60
70
100

Valeur de tendance centrale

Médiane=18 N

Mode=12 Nm

Quantile



Quantile

Le quantile α est la valeur P_α qui laisse α% des observations en-dessous et $(1-\alpha)$% des observations au-dessus d'elle.

Les deux "quartiles" les plus importants sont P_{25} (qui laisse 25 % des observations en-dessous) et P_{75}.

Moyenne (s)

La moyenne ne se définit que
pour une variable statistique quantitative.

Moyenne arithmétique dite par abréviation MOYENNE

Individus	Force (Modalité) N
1	10
2	10.3
3	22.1
4	23
5	26
6	15.2
7	12.3
8	18
9	18.3
10	14.5

$$\overline{F} = \frac{\sum_{i=1}^{10} F_i}{10} = \sum_{i=1}^{10} \left(\frac{1}{10} F_i \right)$$

$$= 16.97 N$$

	Limites des classes (N)					
	Limite inf	Limite sup	**Centre de classe (N)** F_j	**Effectif**	**Fréquence (%)** f_j	$f_j F_j$
$j=1$	10	14	12	3	30	3.6
$j=2$	14	18	16	2	20	3.2
$j=3$	18	22	20	2	20	4
$j=4$	22	26	24	2	20	4.8
$j=5$	26	30	28	1	10	2.8

$$\overline{F} = \sum_{j=1}^{5} f_j F_j$$

$$= (3.6 + 3.2 + 4 + 4.8 + 2.8)$$

$$= 18.4N$$

Série statistique: Moyenne arithmétique=16.97 N

Tableau statistique: Moyenne arithmétique=18.40 N

DIFFERENCE ??????????????

Individus	Force (Modalité) en Newton (N)
1	10
2	10.3
3	22.1
4	23
5	26
6	15.2
7	12.3
8	18
9	18.3
10	14.5

$$\overline{F} = 16.97N$$

Individus	Force (Modalité) en Newton (N)
1	12
2	12
3	24
4	24
5	28
6	16
7	12
8	20
9	20
10	16

$$\overline{F} = 18.4N$$

Valeur de tendance centrale

⇩

Existe-t-il une seule moyenne

$$\overline{F}_R = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^R \right] \right)^{\frac{1}{R}}$$

R=1----→ Moyenne arithmétique

$$\overline{F_1} = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^1 \right] \right)^{\frac{1}{1}}$$

$$\overline{F_1} = \left(\sum_{j=1}^{j=k} \left[f_j F_j \right] \right)$$

R=0----→ Moyenne géométrique

$$\overline{F_0} = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^0 \right] \right)^{\frac{1}{0}}$$

$$\log\left(\overline{F_0}\right) = \sum_{j=1}^{j=k} f_j \log\left(F_j\right)$$

R=-1----→ Moyenne harmonique

$$\overline{F}_{-1} = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^{-1} \right] \right)^{\frac{1}{-1}}$$

$$\overline{F}_{-1} = \frac{1}{\sum_{j=1}^{j=k} \left[f_j \frac{1}{F_j} \right]}$$

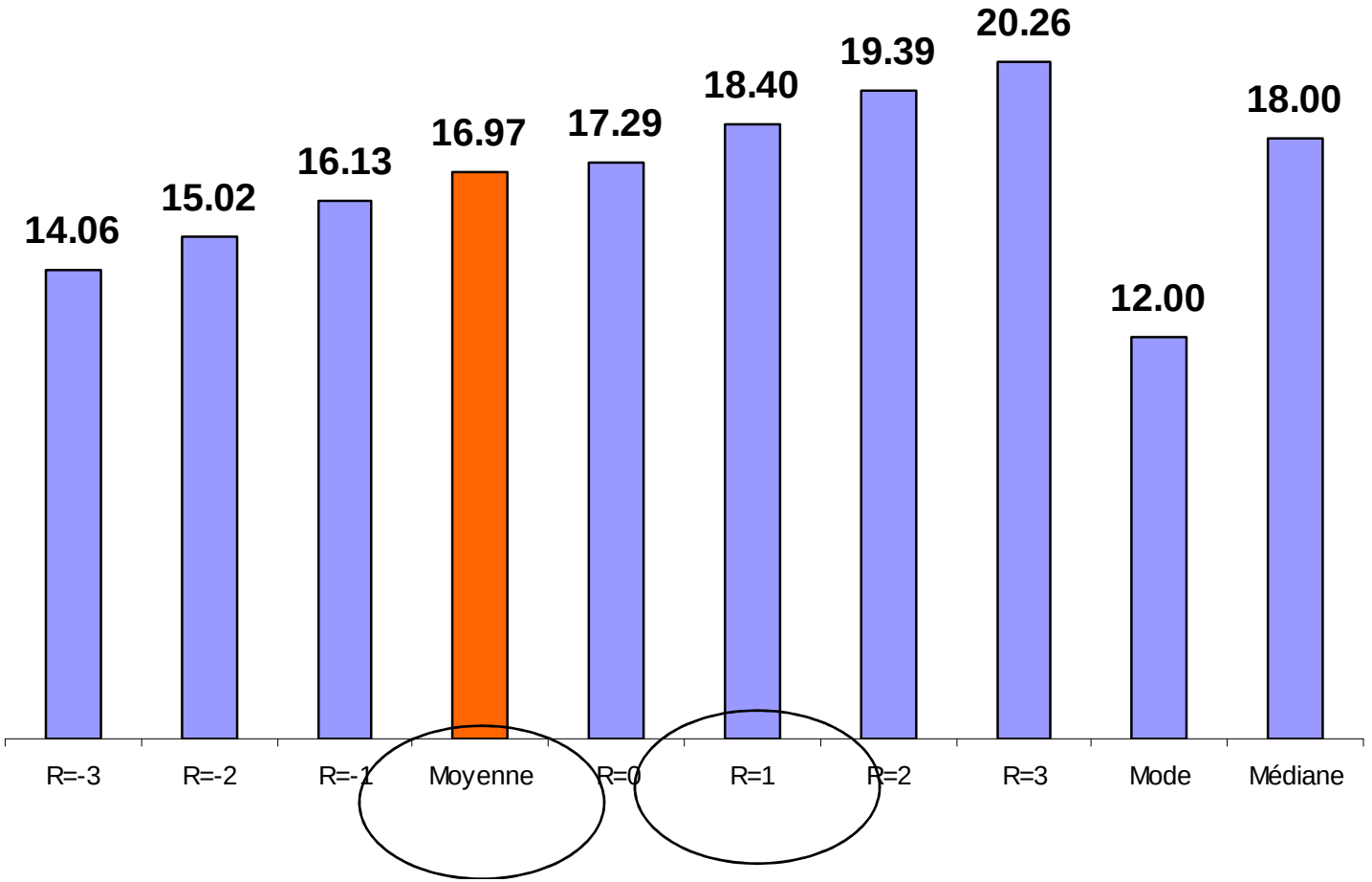
R=2----→ Moyenne quadratique

$$\overline{F}_2 = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^2 \right] \right)^{\frac{1}{2}}$$

$$\overline{F}_2 = \sqrt{\sum_{j=1}^{j=k} \left[f_j \left(F_j \right)^2 \right]}$$

Quelle moyenne choisir????

Etude comparative entres les moyennes et la moyenne réelle obtenue directement du tableau brut



Démontrer que

$$\sum_{j=1}^{j=k}\left[f_j\left(F_j-\overline{F}\right)\right]=0$$

$$\phi(a)=\sum_{j=1}^{j=k}\left[f_j\left(F_j-\overline{F}\right)^2\right]+\left(\overline{F}-a\right)^2$$

$$=\sum_{j=1}^{j=k}\left[f_j\left(F_j-a\right)^2\right]$$

Est-ce la réduction d’un ensemble de valeurs à un seule valeur est une étape suffisante

- Réduire un ensemble de données

- Conserver une partie de l'information

On souhaite donner un prix:
Meilleur classe

Critère de classement: **Moyenne**

Pour faire simple, on suppose que deux classes sont candidates pour ce prix et que chacune des deux classes est composée de 10 élèves

12	13	08	07	11	09	10,5	9,5	10	10

CLASSE –A-

17	13	03	07	19	18	01	02	16	04

CLASSE –B-

MOYENNE=10

Qui mérite de recevoir le prix? ???

Question
Est-ce que les caractéristiques de tendance centrale sont suffisantes pour identifier une valeur de F à utiliser

Tableau n°1

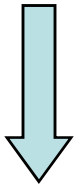
15.25	16.25	18.36	19.26	19.23	15.2	11.2	20.35	18.3	16.3

Tableau n°2

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

Les deux tableaux présentent la même moyenne arithmétique=16.97 N

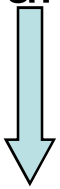
Question: Est-ce que les deux tableaux sont identiques



Trouver une valeur qui reflète

au mieux

La dispersion des valeurs de notre échantillon



Valeur de dispersion

3.2 Description numérique

3.2.2 Caractéristique de dispersion

Caractéristique de dispersion

Choisir

- Série statistique
- Tableau statistique

1. Etendue
2. L’intervalle inter-quartiles
3. Ecart moyen
4. Ecart type

Tableau n°1

15.25	16.25	18.36	19.26	19.23	15.2	11.2	20.35	18.3	16.3

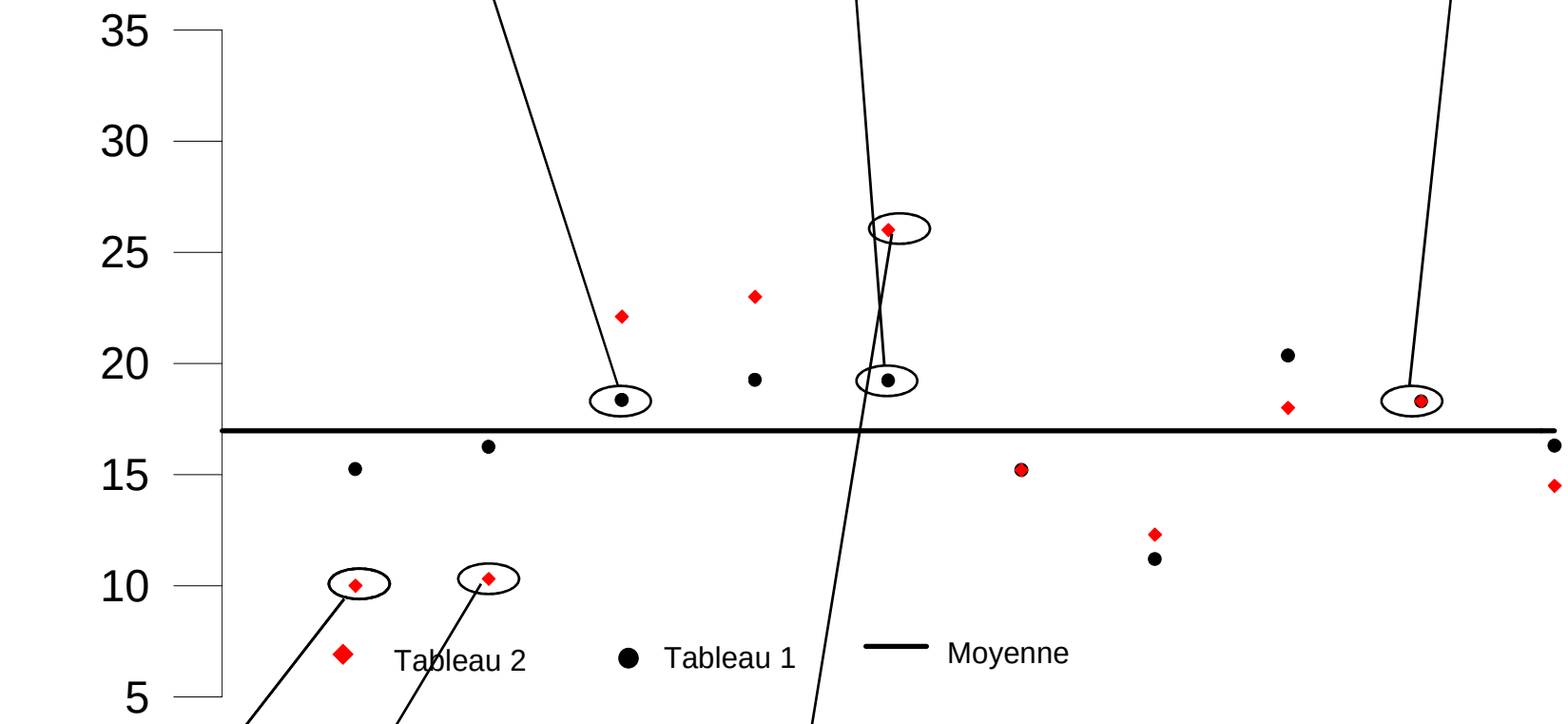


Tableau n°2

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

Caractéristiques de dispersion

Mesurer la différence qui existe entre
la valeur max et min
ETENDUE

Tableau n°1

15.25	16.25	18.36	19.26	19.23	15.2	11.2	20.35	18.3	16.3

ETENDUE=20.35-11.2=**9.15 N**

Tableau n°2

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

ETENDUE=26-10=**16.0 N**

Caractéristiques de dispersion

ETENDUE (Tableau 1)= **9.15 N**

ETENDUE (Tableau 2)= **16.0 N**

ETENDUE (Tableau 2)>ETENDUE (Tableau 1)

C’est logique mais attention cette valeur peut vous fausser l’interprétation car elle se base sur les valeurs extrêmes

Tableau n°1

15.25	16.25	18.36	19.26	19.23	15.2	11.2	20.35	18.3	16.3

Moyenne= 16.97 N et ETENDUE=20.35-11.2=**9.15 N**

Tableau n°2

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

Moyenne= 16.97 N et ETENDUE=26-10=**16.0 N**

Tableau n°3

10.00	17.00	17.40	18.00	26.00	16.00	14.30	17.20	17.30	16.50

Moyenne= 16.97 N et ETENDUE=26-10=**16.0 N**

Tableau n°1

Moyenne= 16.97 N et ETENDUE=20.35-11.2=**9.15 N**

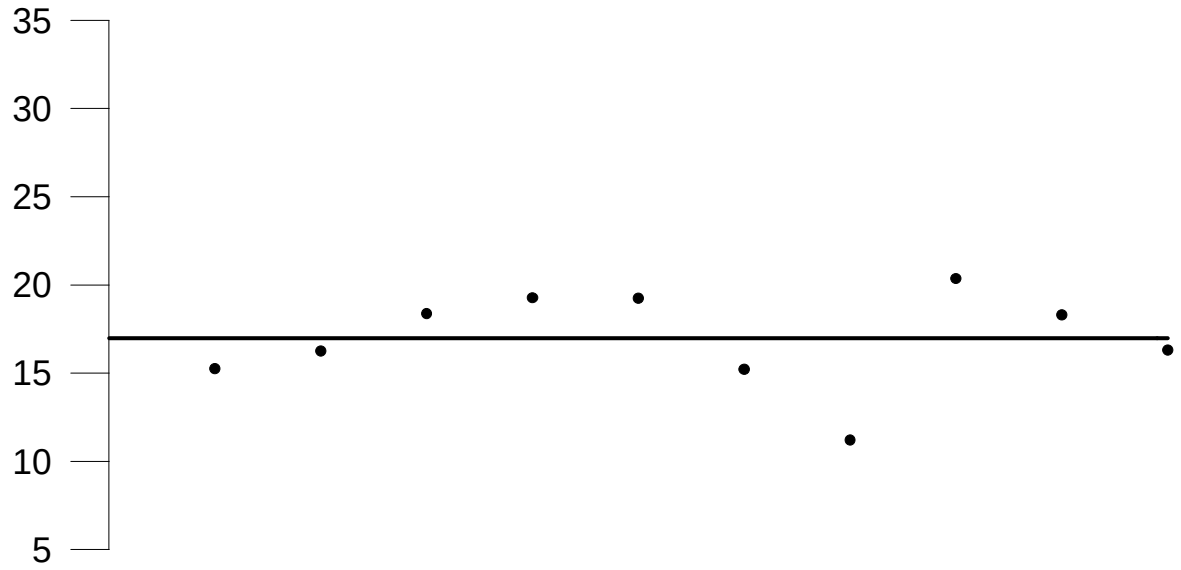


Tableau n°2

Moyenne= 16.97 N et ETENDUE=26-10=**16.0 N**

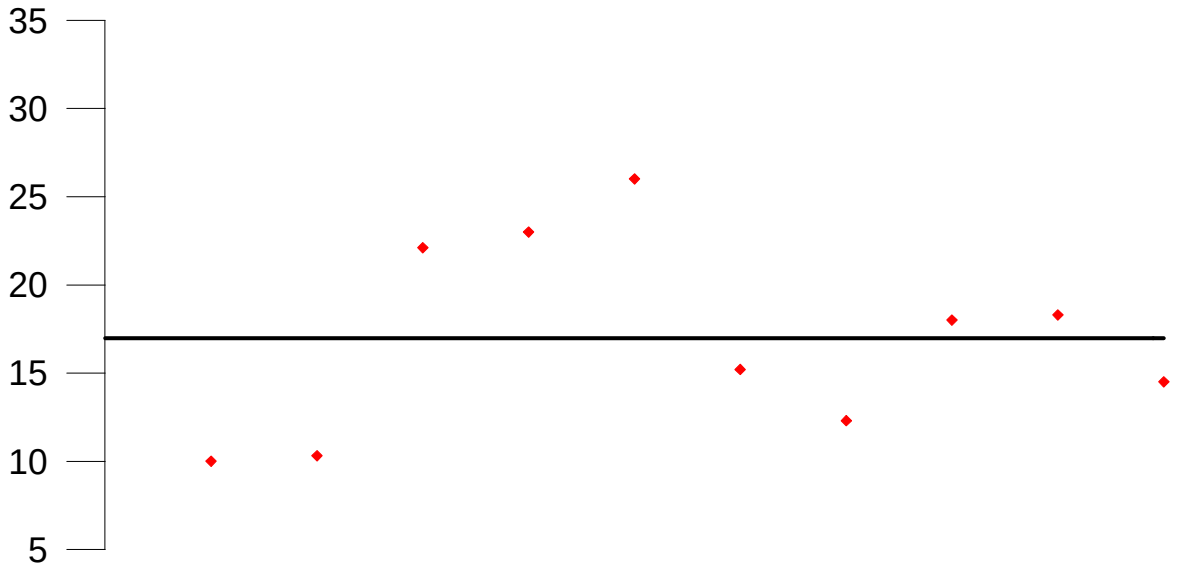
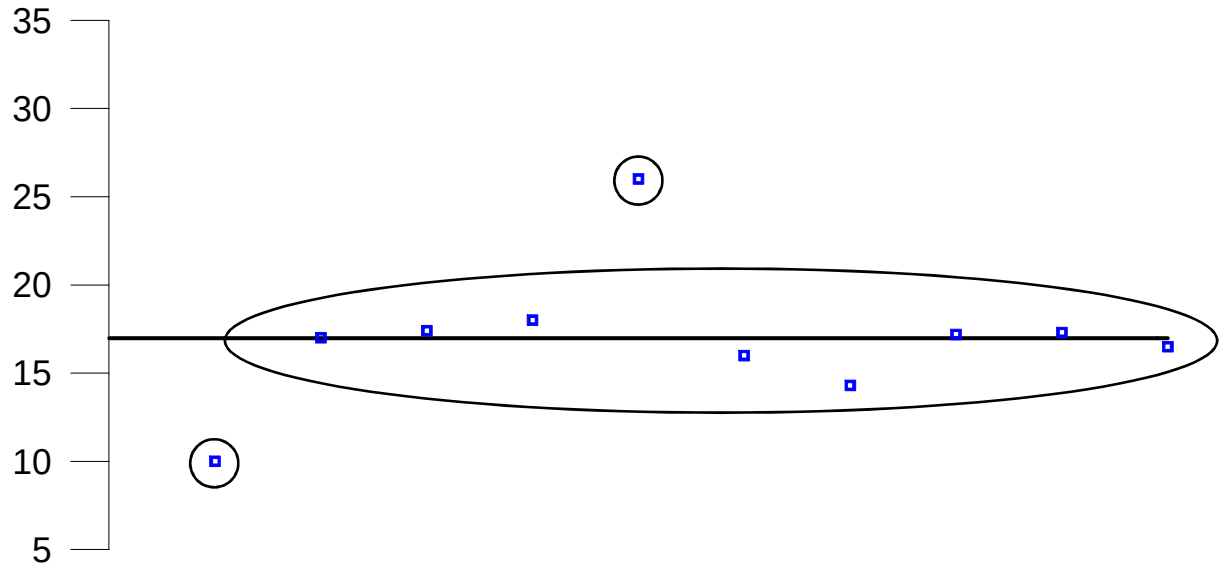
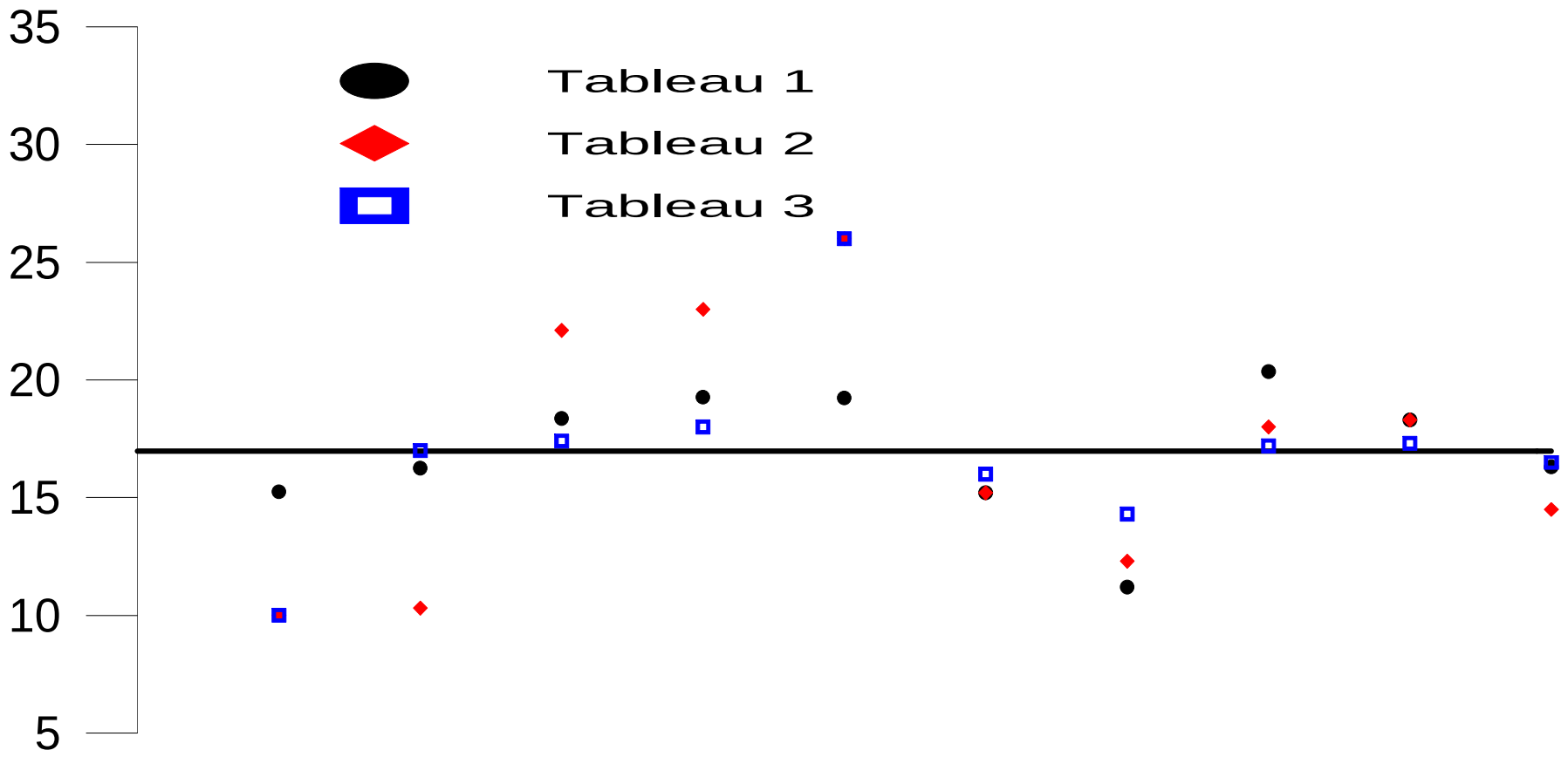


Tableau n°3

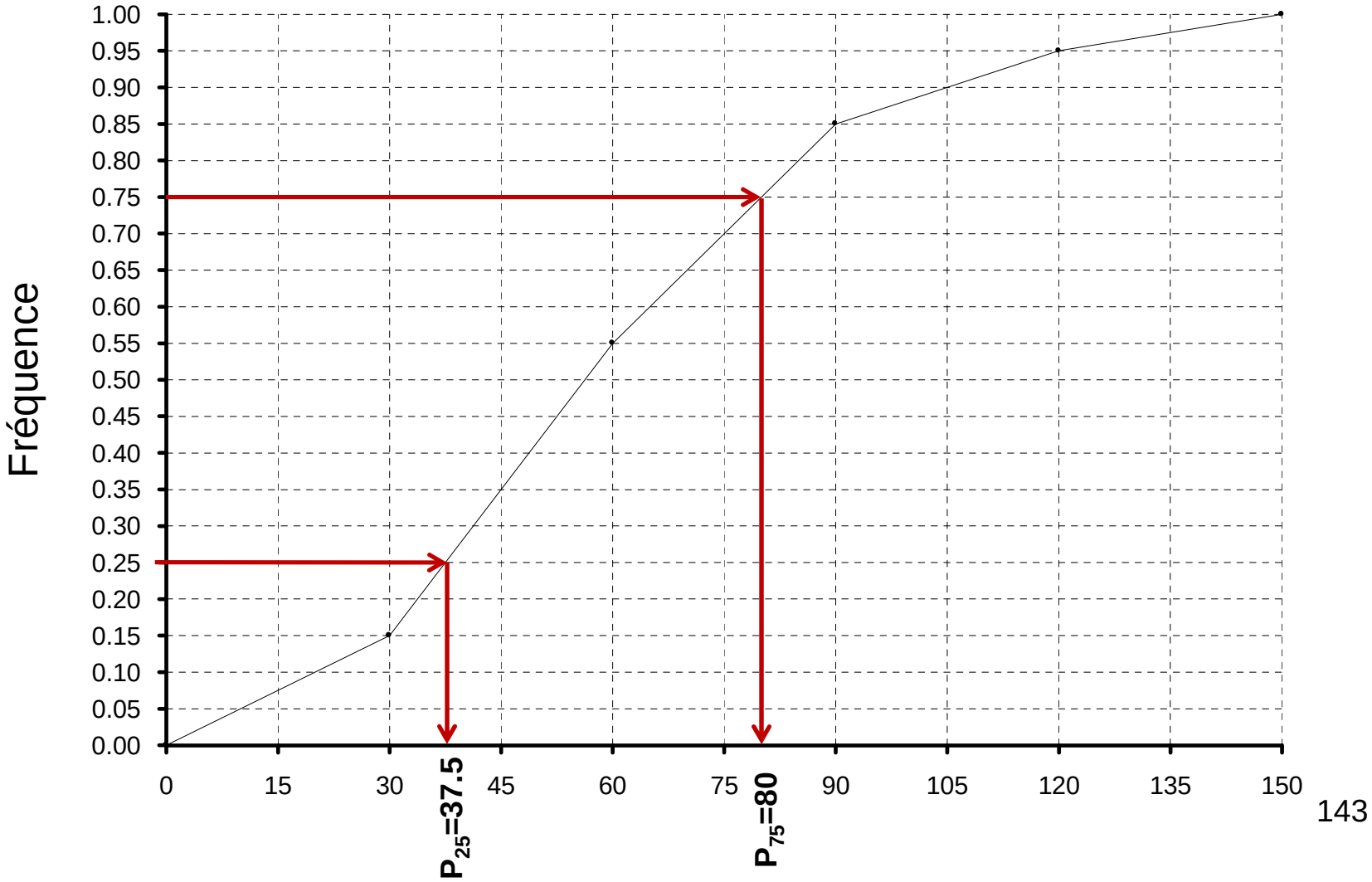
Moyenne= 16.97 N et ETENDUE=26-10=**16.0 N**




35
30
25
20
15
10
5
Tableau 1
Tableau 2
Tableau 3

Caractéristiques de dispersion

L'intervalle inter-quartiles $H = P_{75} - P_{25}$



Caractéristiques de dispersion

Mesurer la dispersion des valeurs par rapport à la moyenne

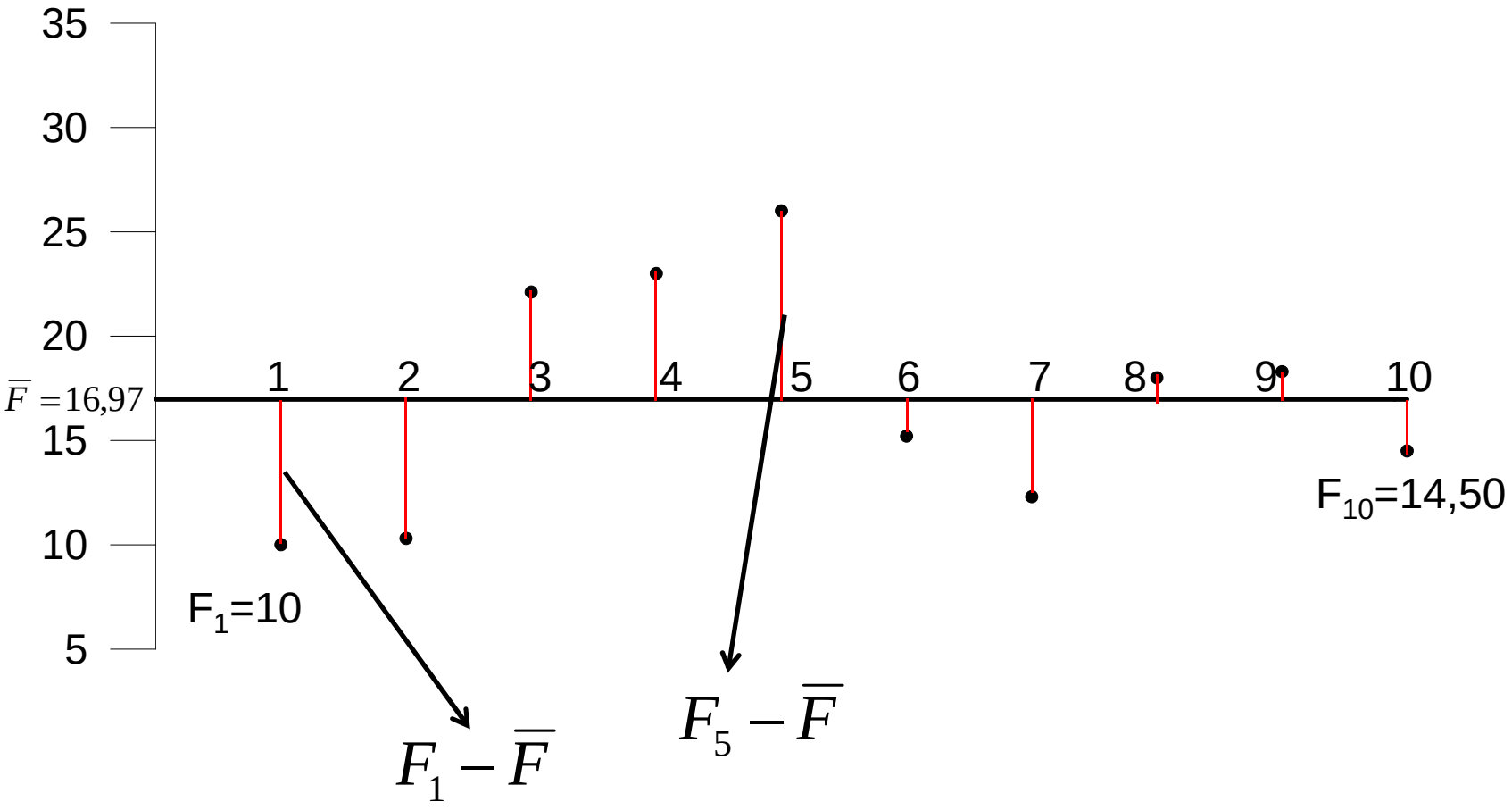
Estimer la différence entre

les valeurs observées

Et

la moyenne

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50



Valeur de dispersion

Cette valeur DOIT être très petite

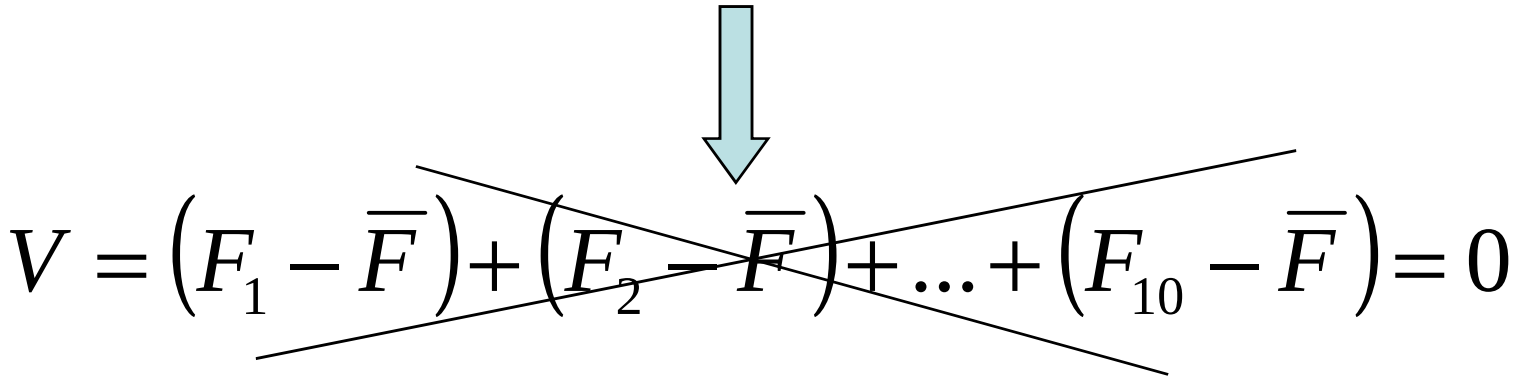
$$V = \left(F_1 - \overline{F}\right) + \left(F_2 - \overline{F}\right) + \ldots + \left(F_{10} - \overline{F}\right)$$

$$V = \left(F_1 + F_2 + \ldots + F_{10} - 10\overline{F}\right)$$

$$V = \left(10\overline{F} - 10\overline{F}\right) = 0$$

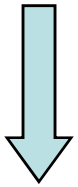
Valeur de dispersion

Cette valeur DOIT être très petite



$$V = \left(F_1 - \overline{F}\right) + \left(F_2 - \overline{F}\right) + \ldots + \left(F_{10} - \overline{F}\right) = 0$$

Or cette valeur est en fait nulle et ne peut donc mesurer la ***dispersion***



Autre méthode

Valeur de dispersion

$$V = \left|F_1 - \overline{F}\right| + \left|F_2 - \overline{F}\right| + \ldots + \left|F_{10} - \overline{F}\right|$$

Cette valeur permet d'estimer la différence

La valeur absolue est une fonction mathématique difficilement dérivable

Valeur de dispersion

$$V = \left(F_1 - \overline{F}\right)^2 + \left(F_2 - \overline{F}\right)^2 + ... + \left(F_{10} - \overline{F}\right)^2$$

Cette valeur est toujours positif et permet d'estimer la différence entre les valeurs observées et leur moyenne

Valeur de dispersion

Mesurer la dispersion des valeurs par rapport à la moyenne

$$V = \sum_{i=1}^{10} \left(F_i - \overline{F} \right)^R \text{ et } R \text{ pair}$$

Série statistique ***Tableau statistique***

Moyenne

$$\overline{F} = \sum_{i=1} \frac{F_i}{n} \qquad \overline{F} = \sum_{j=1}^{j=k} f_j F_j$$

Écart moyen

$$EM_F = \sum_{i=1}^{10} \left[\frac{1}{n} \left| F_i - \overline{F} \right| \right] \qquad EM_F = \sum_{j=1}^{j=k} \left[f_j \left| F_j - \overline{F} \right| \right]$$

Série statistique ***Tableau statistique***

Variance

$$(\sigma_F)^2 = \sum_{i=1}\left[\frac{(F_i - \overline{F})^2}{n}\right] \qquad (\sigma_F)^2 = \sum_{j=1}^{j=k}\left[f_j (F_j - \overline{F})^2\right]$$

Ecart type

$$\sigma_F$$

Série statistique ***Tableau statistique***

Différence d'ordre *R*

$$V_R = \left(\sum_{i=1}^{i=n} \left[\frac{1}{n} \left(F_i - \overline{F} \right)^R \right] \right)^{\frac{1}{R}} \quad V_R = \left(\sum_{j=1}^{j=k} \left[f_j \left(F_j - \overline{F} \right)^R \right] \right)^{\frac{1}{R}}$$

Exercice: Déterminer l'Ecart Moyen et l'Ecart type

10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

Limites des classes (N)		**Centre de classe (N)**	**Effectif**	**Fréquence (%)**
Limite inf	Limite sup			
10	14	12	3	30
14	18	16	2	20
18	22	20	2	20
22	26	24	2	20
26	30	28	1	10

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$f_k\left(F_k - \overline{F}\right)$
10	14	12	3				
14	18	16	2				
18	22	20	2				
22	26	24	2				
26	30	28	1				

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$f_k\left(F_k - \overline{F}\right)$
10	14	12	3	30			
14	18	16	2	20			
18	22	20	2	20			
22	26	24	2	20			
26	30	28	1	10			

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$f_k\left(F_k - \overline{F}\right)$
10	14	12	3	30	3.6		
14	18	16	2	20	3.2		
18	22	20	2	20	4		
22	26	24	2	20	4.8		
26	30	28	1	10	2.8		
					$\overline{F} = 18.4$ N		

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$f_k\left(F_k - \overline{F}\right)$
10	14	12	3	30	3.6	-6.4	
14	18	16	2	20	3.2	-2.4	
18	22	20	2	20	4	1.6	
22	26	24	2	20	4.8	5.6	
26	30	28	1	10	2.8	9.6	
					$\overline{F} = 18.4$ N		

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$f_k\left(F_k - \overline{F}\right)$
10	14	12	3	30	3.6	-6.4	-1.92
14	18	16	2	20	3.2	-2.4	-0.48
18	22	20	2	20	4	1.6	0.32
22	26	24	2	20	4.8	5.6	1.12
26	30	28	1	10	2.8	9.6	0.96
					$\overline{F} = 18.4$ N		0.0

$$\sum_{j=1}^{j=k}\left[f_j\left(F_j - \overline{F}\right)\right] = 0$$

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$\left\|F_k - \overline{F}\right\|$	$f_k \left\|F_k - \overline{F}\right\|$
10	14	12	3	30	3.6	-6.4	6.4	1.92
14	18	16	2	20	3.2	-2.4	2.4	0.48
18	22	20	2	20	4	1.6	1.6	0.32
22	26	24	2	20	4.8	5.6	5.6	1.12
26	30	28	1	10	2.8	9.6	9.6	0.96
					$\overline{F} = 18.4$ N			4.8 N

$$V = \sum_{j=1}^{j=k} f_j \left|F_j - \overline{F}\right| = 4.8N$$

Ecart Moyen= 4.8 N

Lim inf	Lim sup	Centre de classe	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$(F_k - \overline{F})^2$	$f_k (F_k - \overline{F})^2$
10	14	12	3	30	3.6	-6.4	40.96	12.288
14	18	16	2	20	3.2	-2.4	5.76	1.152
18	22	20	2	20	4	1.6	2.56	0.512
22	26	24	2	20	4.8	5.6	31.36	6.272
26	30	28	1	10	2.8	9.6	92.16	9.216
					$\overline{F}$ =18.4 N	8		29.44

Ecart type= $\sqrt{29.44} = 5.42N$

Lim inf	Lim sup	Centre de classe F_k	Effectif	Fréquence f_k (%)	$f_k F_k$	$F_k - \overline{F}$	$(F_k - \overline{F})^2$	$f_k (F_k - \overline{F})^2$
11	13.4	12.2	1	10	1.22	-5.04	25.4016	2.54016
13.4	15.8	14.6	2	20	2.92	-2.64	6.9696	1.39392
15.8	18.2	17	2	20	3.4	-0.24	0.0576	0.01152
18.2	20.6	19.4	5	50	9.7	2.16	4.6656	2.3328
					17.24 N			6.2784

Ecart Type= $\sqrt{6.2784} = 2.50$ N

Tableau n°1

15.25	16.25	18.36	19.26	19.23	15.20	11.20	20.35	18.30	16.25

$$(\sigma)^2 = \sum_{j=1}^{j=k} f_j \left(F_j - \overline{F}\right)^2 = 6.28\, N^2$$

Tableau n°2

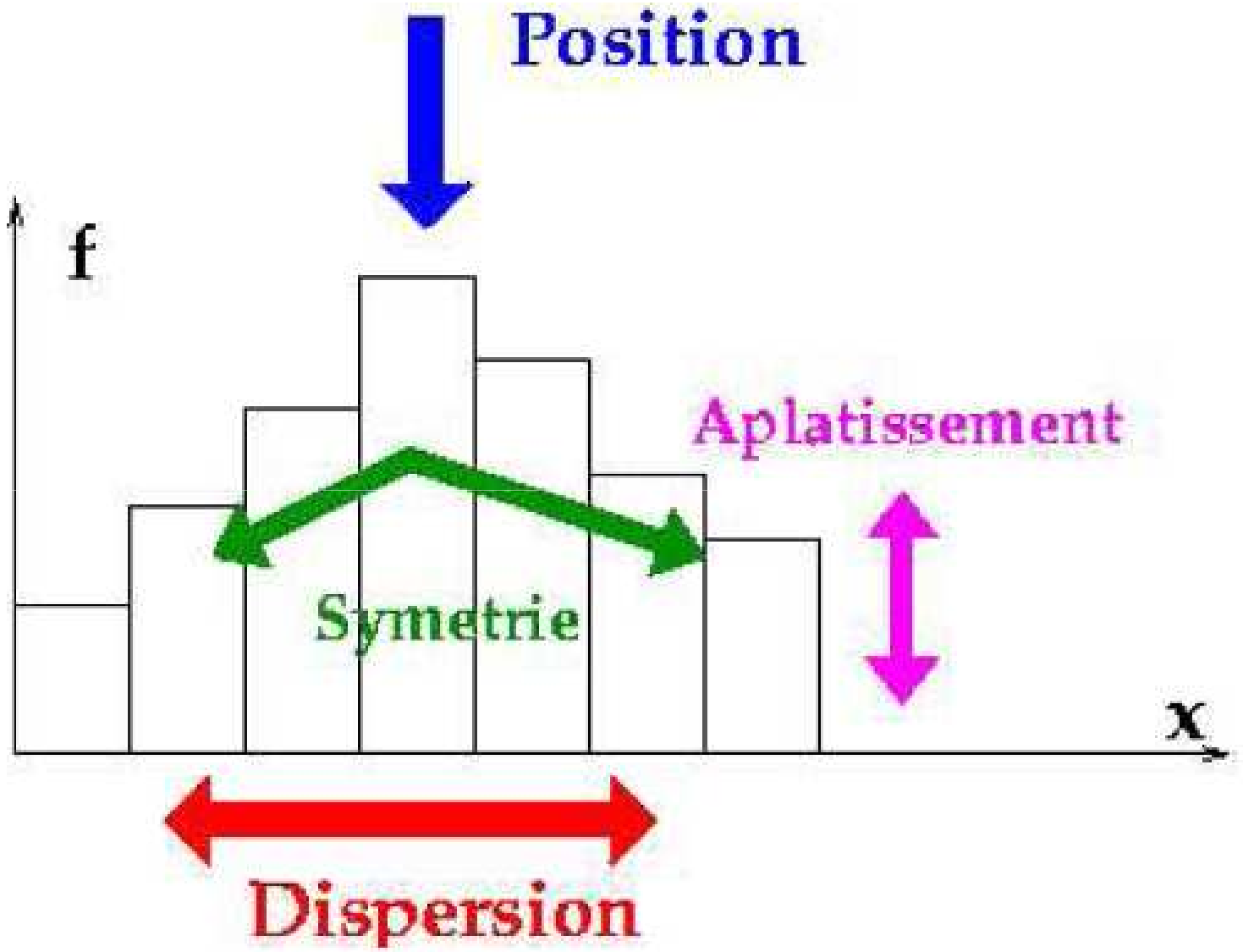
10.00	10.30	22.10	23.00	26.00	15.20	12.30	18.00	18.30	14.50

$$(\sigma)^2 = \sum_{j=1}^{j=k} f_j \left(F_j - \overline{F}\right)^2 = 29.44\, N^2$$

3.3 Caractéristiques de formes

1. Coefficient de variation
2. Coefficient de symétrie
3. Coefficient d'aplatissement

Coefficient de variation (noté **CV**)=Ecart type / Moyenne

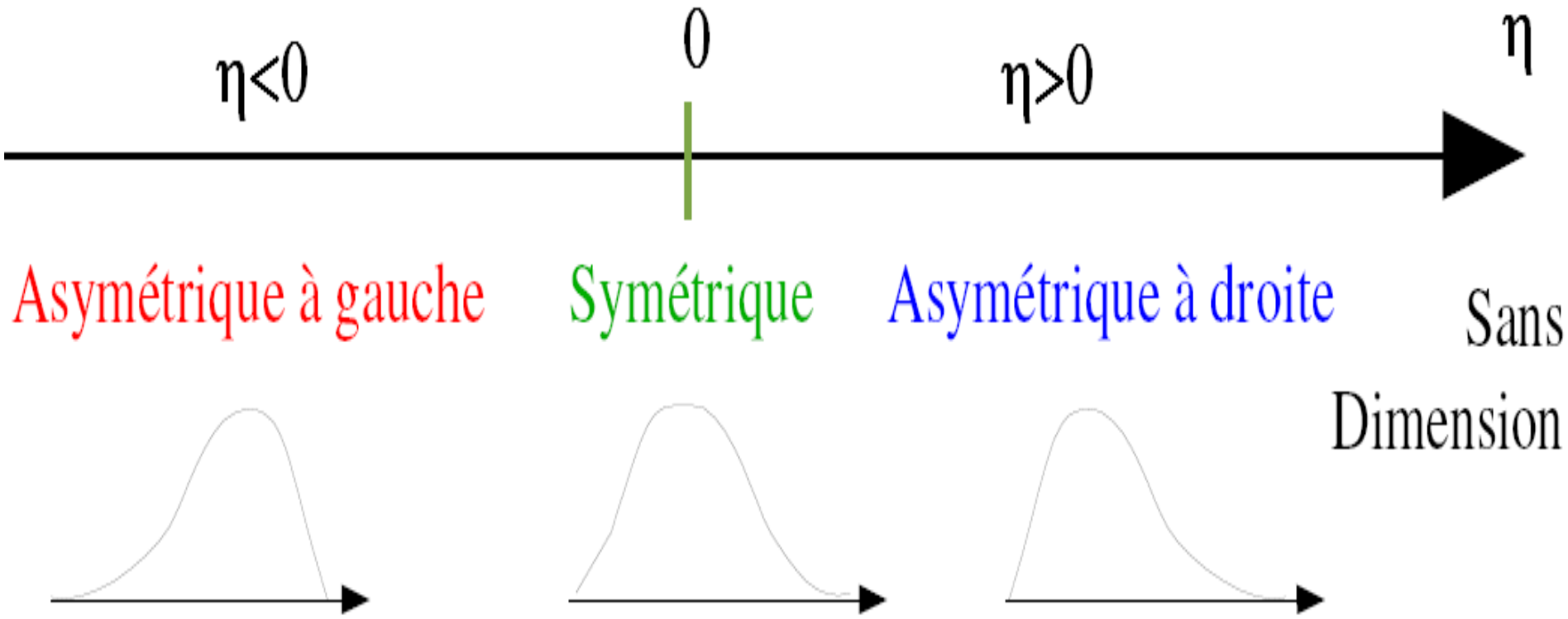

Position
f
Aplatissement
Symetrie
x
Dispersion

Coefficient de symétrie **(Cœfficient of Skewness)**

$$\eta = \sum_{i=1}^{i=n} \frac{1}{n} \frac{\left(F_i - \overline{F}\right)^3}{\left(\sigma_F\right)^3}$$

$$\eta = \sum_{j=1}^{j=k} f_j \frac{\left(F_j - \overline{F}\right)^3}{\left(\sigma_F\right)^3}$$

Coefficient de symétrie **(Cœfficient of Skewness)**



Coefficient d'aplatissement **(Cœfficient of Kurtosis)**
généralement comparé à la valeur de 3 qui est celui de la distribution

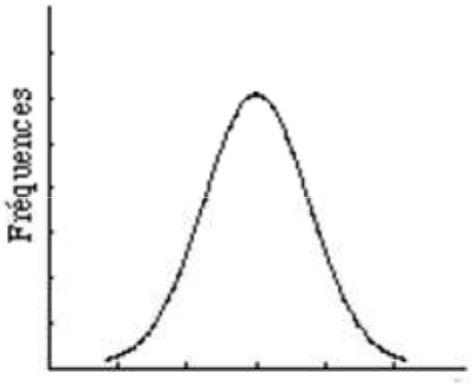
$$\kappa = \sum_{i=1}^{i=n} \frac{1}{n} \frac{\left(F_i - \overline{F}\right)^4}{\left(\sigma_F\right)^4}$$

$$\kappa = \sum_{j=1}^{j=k} f_j \frac{\left(F_j - \overline{F}\right)^4}{\left(\sigma_F\right)^4}$$

Coefficient d'aplatissement **(Cœfficient of Kurtosis)**
généralement comparé à la valeur de 3 qui celle d'une distribution normale

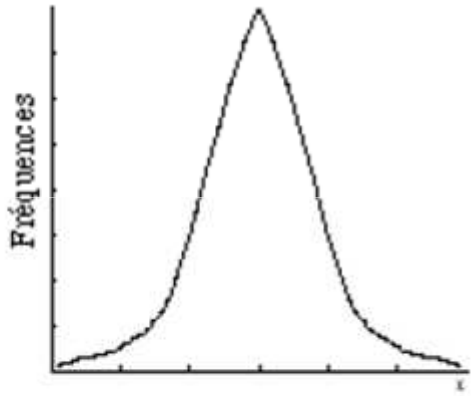
platikurtique

$\kappa - 3 < 0$



normale

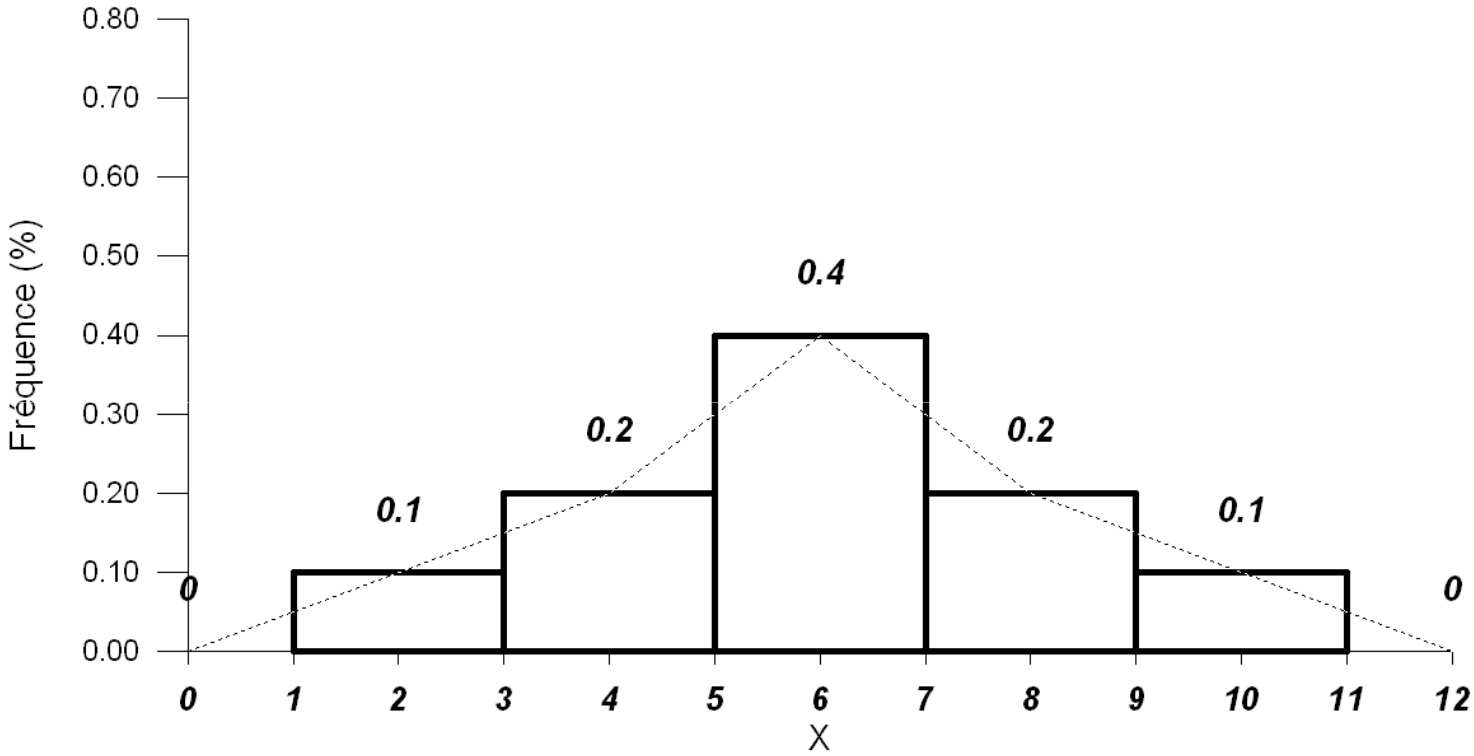
$\kappa - 3 = 0$



leptokurtique

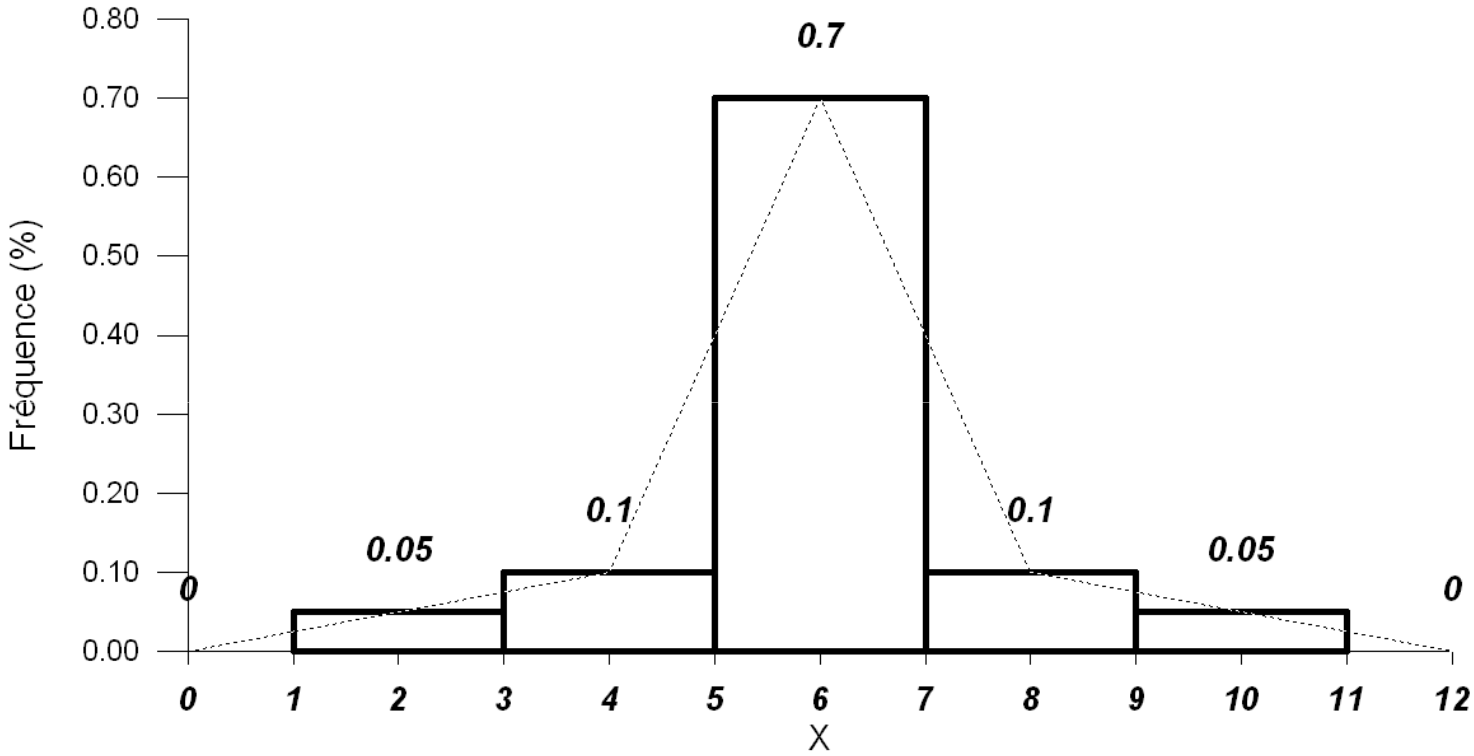
$\kappa - 3 > 0$

Exemple:

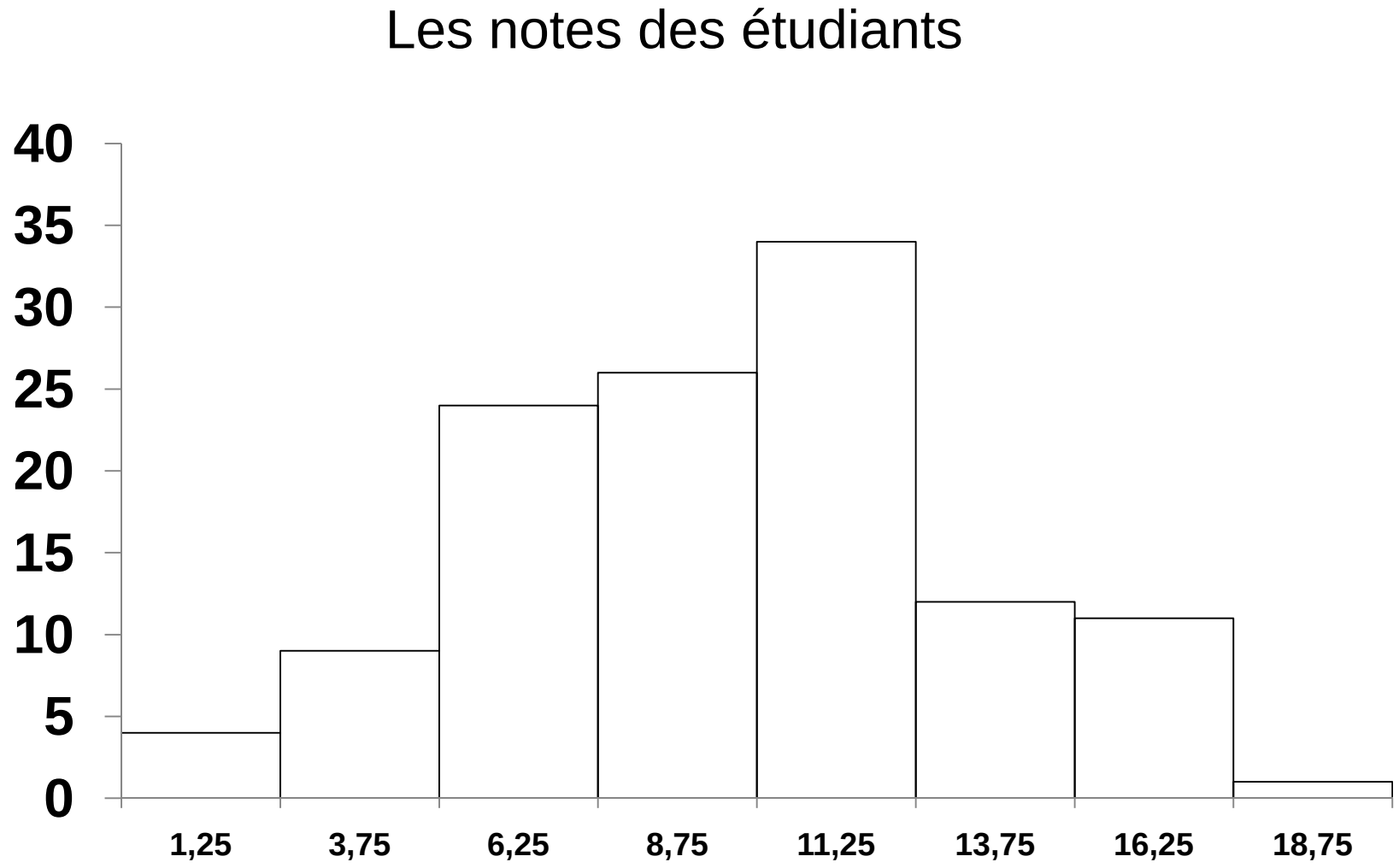


$$\kappa - 3 = -0{,}5 < 0$$

Exemple: La distribution normale est symétrique



$$\kappa - 3 = 2 > 0$$


Les notes des étudiants
40
35
30
25
20
15
10
5
0
1,25
3,75
6,25
8,75
11,25
13,75
16,25
18,75

Chapitre 4

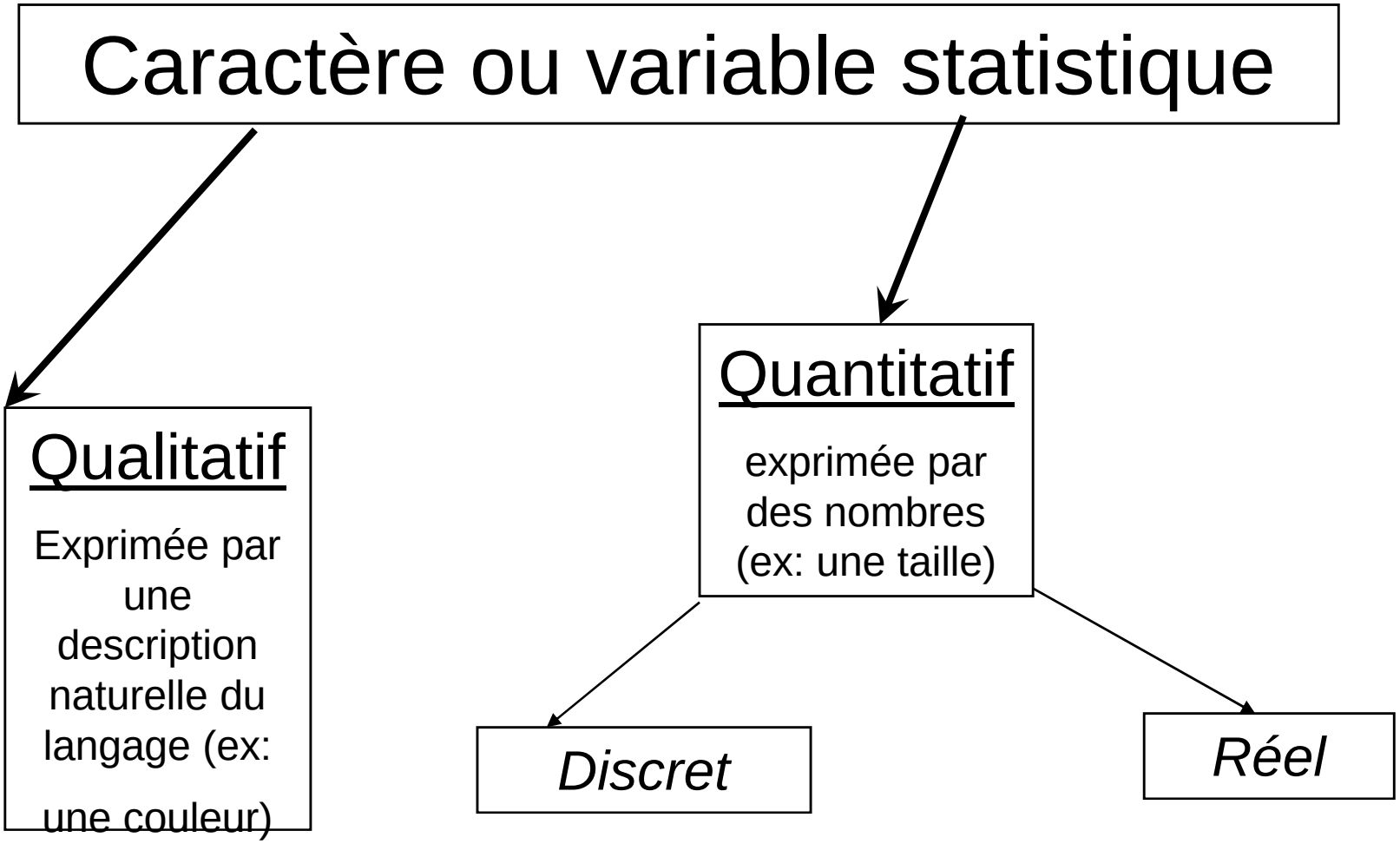
Distribution statistique à deux caractères

4.1 Introduction

4.2 Description numérique

4.3 Principe de la méthode des moindres carrées "Least Square method »

4.4 Covariance



k modalités $c_1, c_2, \ldots, c_k$

4.1 Introduction

Un même individu peut il avoir plusieurs caractères????

Oui

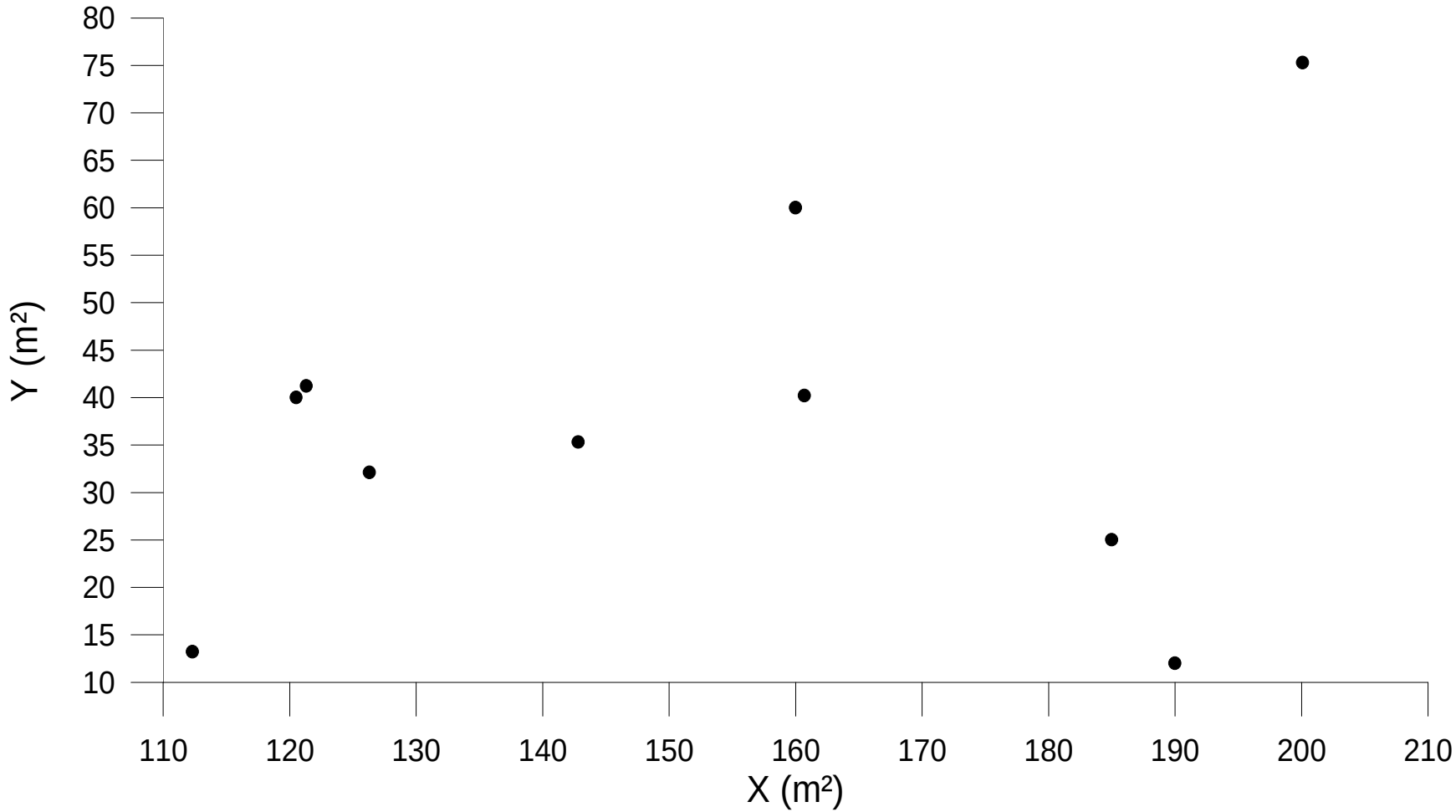
Individu	Maison
Caractère 1	Surface habitable (notée X)
Caractère 2	Surface non habitable (notée Y)

Surface non habitable: Le jardin et autres…

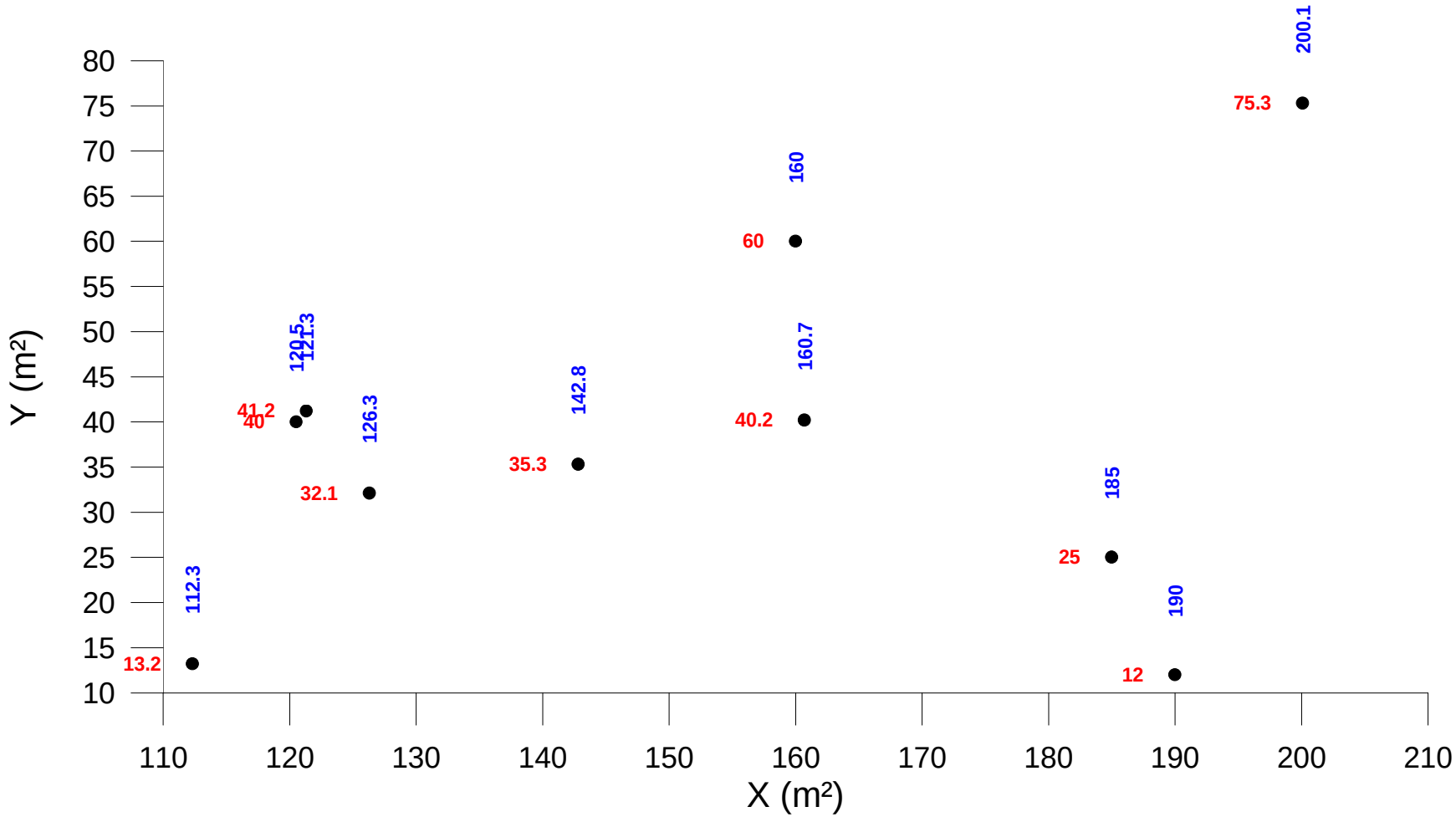
Ce qui est recherché c'est

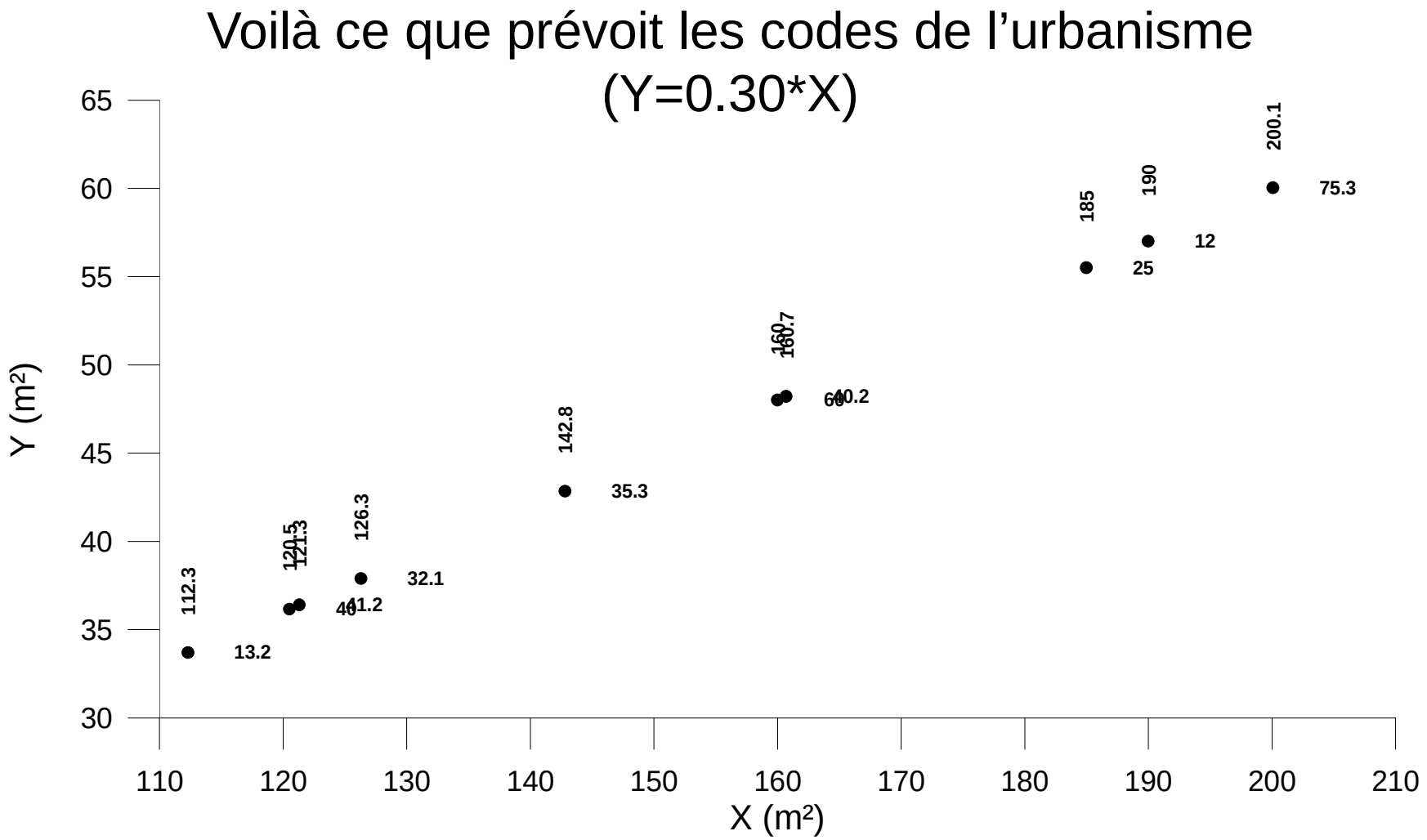
De possibles relations entre le caractère 1 et le caractère 2.

Maison	X (m²)	Y (m²)
1	200,1	75,3
2	121,3	41,2
3	160,7	40,2
4	112,3	13,2
5	190	12
6	120,5	40
7	160	60
8	185	25
9	126,3	32,1
10	142,8	35,3


80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
Y (m²)
110
120
130
140
150
160
170
180
190
200
210
X (m²)

La distribution des valeurs (X et Y)




Voilà ce que prévoit les codes de l'urbanisme
(Y=0.30*X)
Y (m²)
X (m²)
65
60
55
50
45
40
35
30
110
120
130
140
150
160
170
180
190
200
210
112.3
13.2
120.5
121.3
126.3
32.1
41.2
142.8
35.3
160
160.7
40.2
185
190
12
25
200.1
75.3

4.2 Description numérique

Individu	X	Y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
e	x_e	y_e
⋮	⋮	⋮
n	x_n	y_n

Il est possible de passer de la série statistique vers le tableau statistique

Comment?

Tableau statistique

Classe Y / Classe X		***(1)*** $[CY_1, CY_2[$	...	***(j)*** $[CY_j, CY_{j+1}[$	…	***(k)*** $[CY_k, CY_{k+1}[$
		Y_1		Y_j		Y_K
(1) $[CX_1, CX_2[$	X_1	n_{11}		n_{1j}		n_{1k}
⋮						
(i) $[CX_i, CX_{i+1}[$	X_i	n_{i1}		n_{ij}		n_{ik}
⋮						
(m) $[CX_m, CX_{m+1}[$	X_m	n_{m1}		n_{mj}		n_{mk}

n_{11}: l'ensemble des individus ayant la modalité X_1 de X et Y_1 de Y

$$n = \sum_{i=1}^{i=m} \sum_{j=1}^{j=k} n_{ij}$$

Tableau statistique

Classe Y / Classe X		***(1)*** $[CY_1, CY_2[$	...	***(j)*** $[CY_j, CY_{j+1}[$	…	***(k)*** $[CY_k, CY_{k+1}[$
		Y_1		Y_j		Y_K
(1) $[CX_1, CX_2[$	X_1	f_{11}		f_{1j}		f_{1k}
⋮						
(i) $[CX_i, CX_{i+1}[$	X_i	f_{i1}		f_{ij}		f_{ik}
⋮						
(m) $[CX_m, CX_{m+1}[$	X_m	f_{m1}		f_{mj}		f_{mk}

f_{11}: Fréquence des individus ayant la modalité x_1 de X et y_1 de Y

$$f_{ij} = \frac{n_{ij}}{n}$$

Individu	X	Y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
e	x_e	y_e
⋮	⋮	⋮
n	x_n	y_n

$$\overline{\overline{X}} = \sum_{e=1}^{e=n} \frac{x_n}{n} \qquad \overline{\overline{Y}} = \sum_{e=1}^{e=n} \frac{y_n}{n}$$

$$\sigma_X = \sqrt{\sum_{e=1}^{e=n} \left[\frac{1}{n} \left(x_i - \overline{\overline{X}} \right)^2 \right]} \qquad \sigma_Y = \sqrt{\sum_{e=1}^{e=n} \left[\frac{1}{n} \left(y_i - \overline{\overline{Y}} \right)^2 \right]}$$

Fréquence marginale

	Y_1	...	Y_j		Y_K	
x_1	f_{11}		f_{1j}		f_{1k}	$f_{1.}$
⋮						
x_i	f_{i1}		f_{ij}		f_{ik}	$f_{j.}$
⋮						
x_m	f_{m1}		f_{mj}		f_{mk}	$f_{m.}$
	$f_{.1}$		$f_{.j}$		$f_{.k}$	

$f_{.1}$: Fréquence des individus ayant la modalité y1 de Y

$$f_{.1} = \sum_{i=1}^{i=m} f_{i1}$$

Fréquence conditionnelle

Répond à la question suivante:

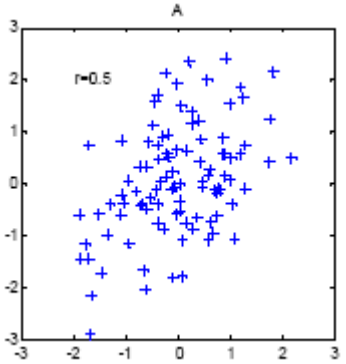
Quelle la fréquence des individus ayant la modalité x_1 de X sachant que $Y=y_1$

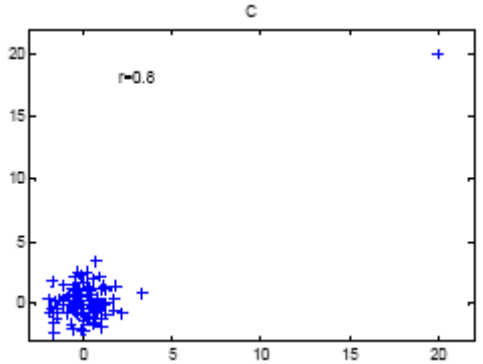
	Y_1	...	Y_j		Y_K	
x_1	f_{11}		f_{1j}		f_{1k}	$f_{1.}$
⋮						
x_i	f_{i1}		f_{ij}		f_{ik}	$f_{j.}$
⋮						
x_m	f_{m1}		f_{mj}		f_{mk}	$f_{m.}$
	$f_{.1}$		$f_{.j}$		$f_{.k}$	

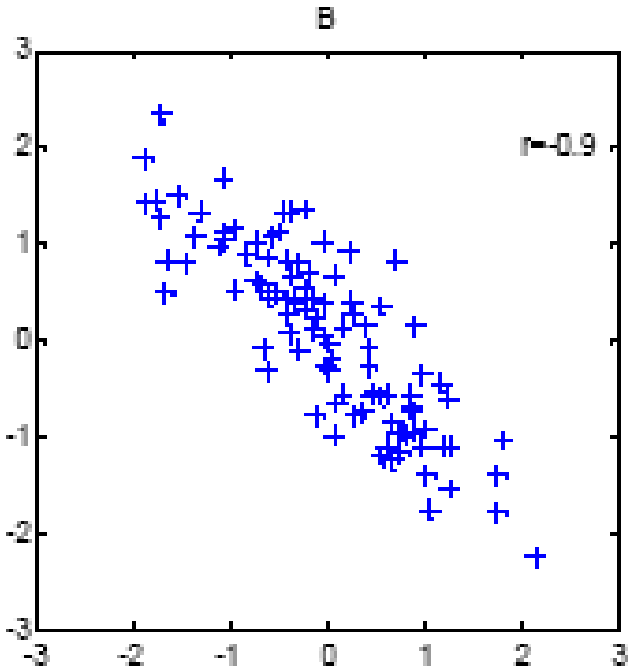
	Y_1	...				
X_1	f_{11}					
⋮						
X_i	f_{i1}					
⋮						
X_m	f_{m1}					
	$f_{.1}$					

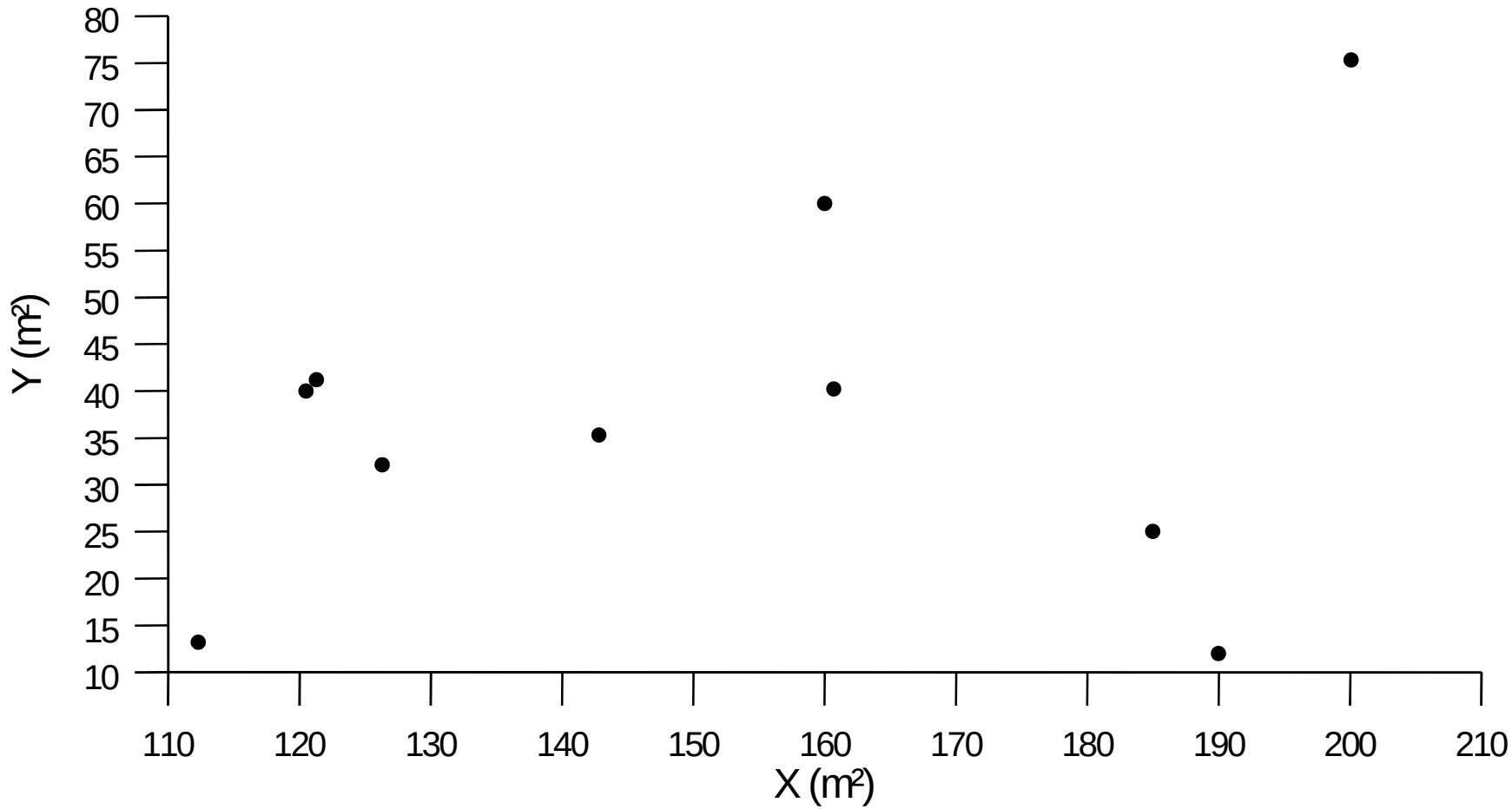
$$\frac{f_{11}}{f_{.1}} = \frac{\frac{n_{11}}{n}}{\frac{\sum_{i=m}^{i=1} n_{i1}}{n}} = \frac{n_{11}}{\sum_{i=m}^{i=1} n_{i1}}$$

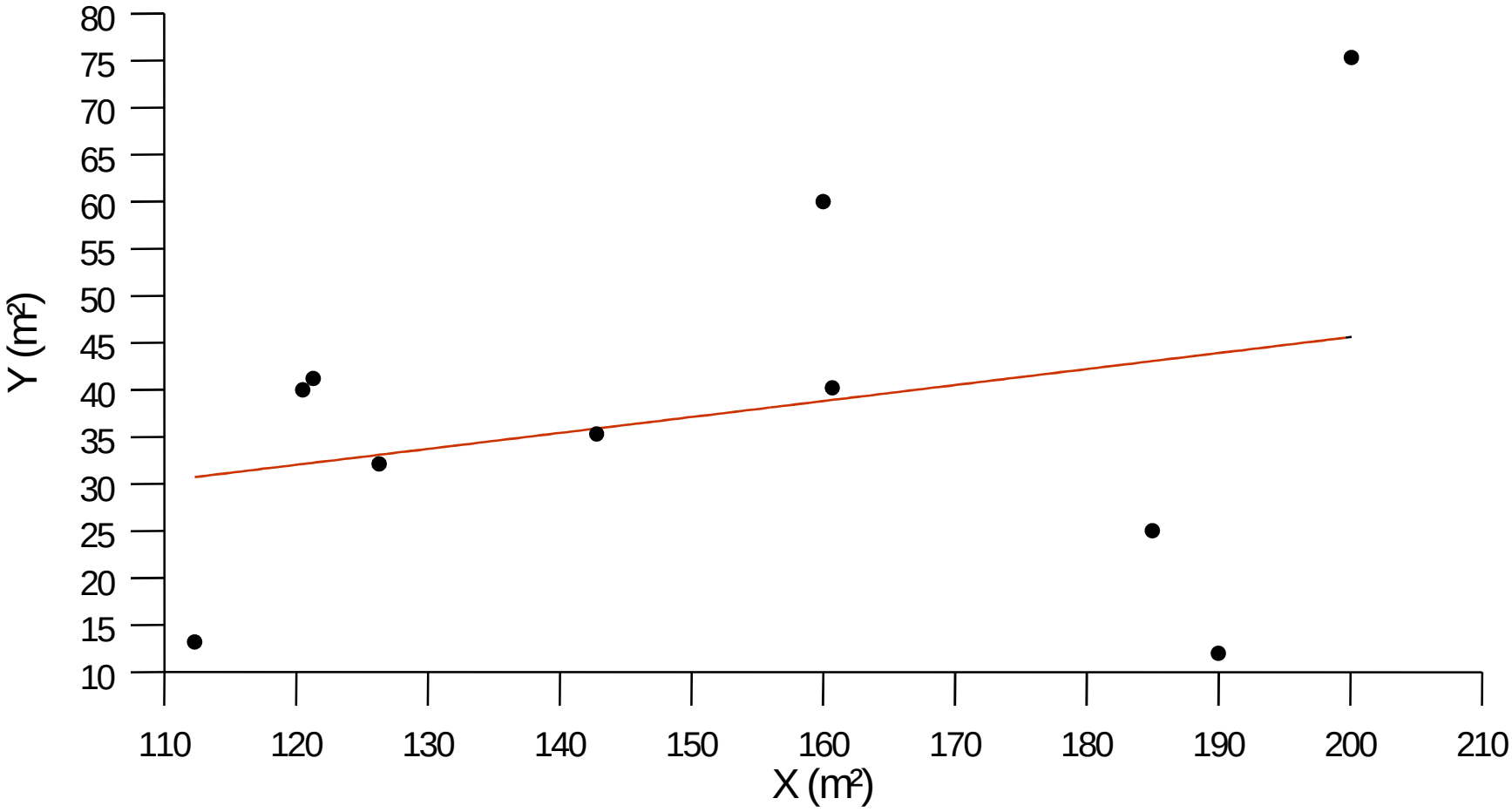
•4.3 Principe de la méthode des moindres carrées "Least Square method »

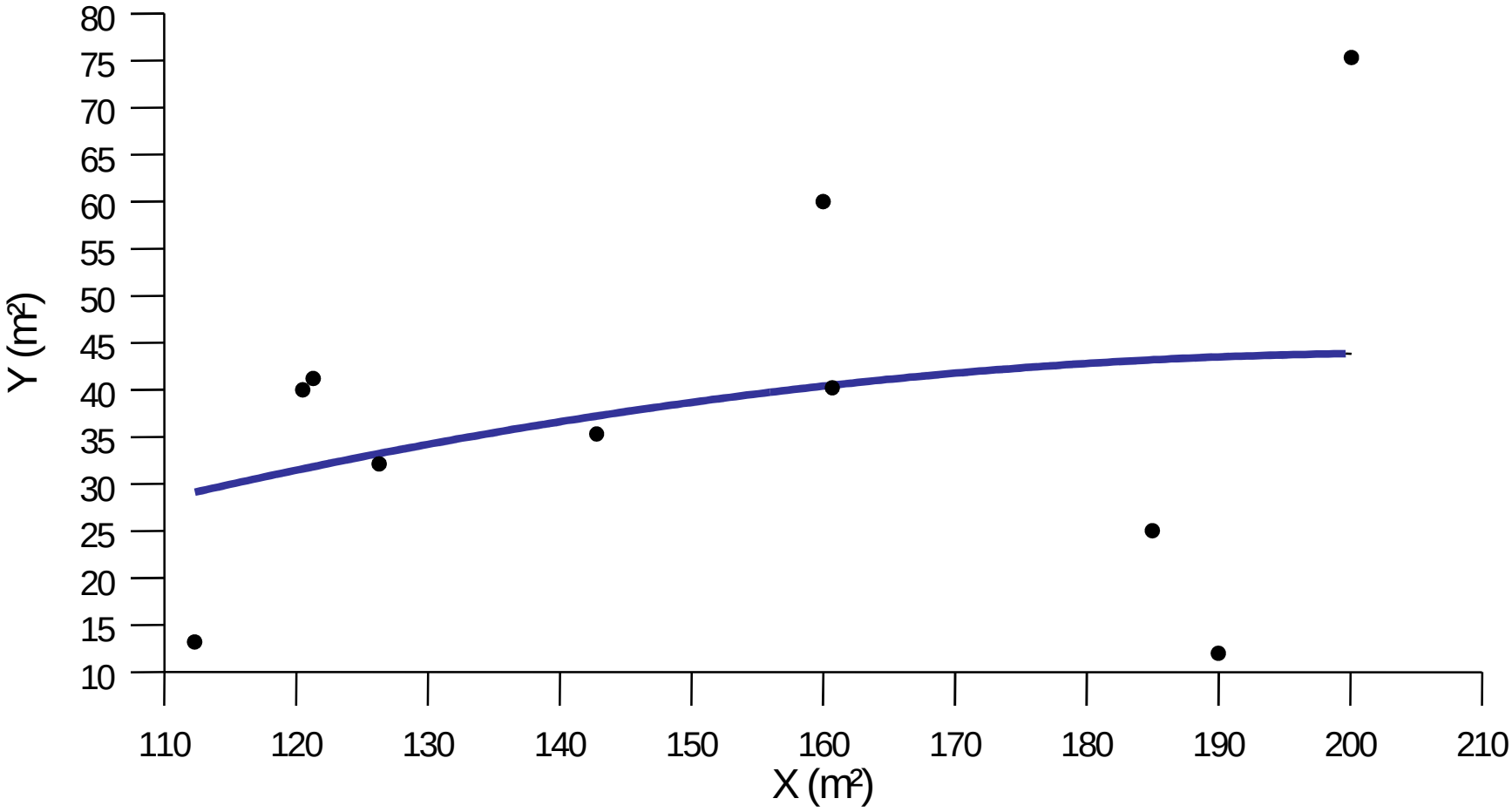







80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
Y (m²)
110
120
130
140
150
160
170
180
190
200
210
X (m²)


80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
Y (m²)
110
120
130
140
150
160
170
180
190
200
210
X (m²)


80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
Y (m²)
110
120
130
140
150
160
170
180
190
200
210
X (m²)

Problème étudié:

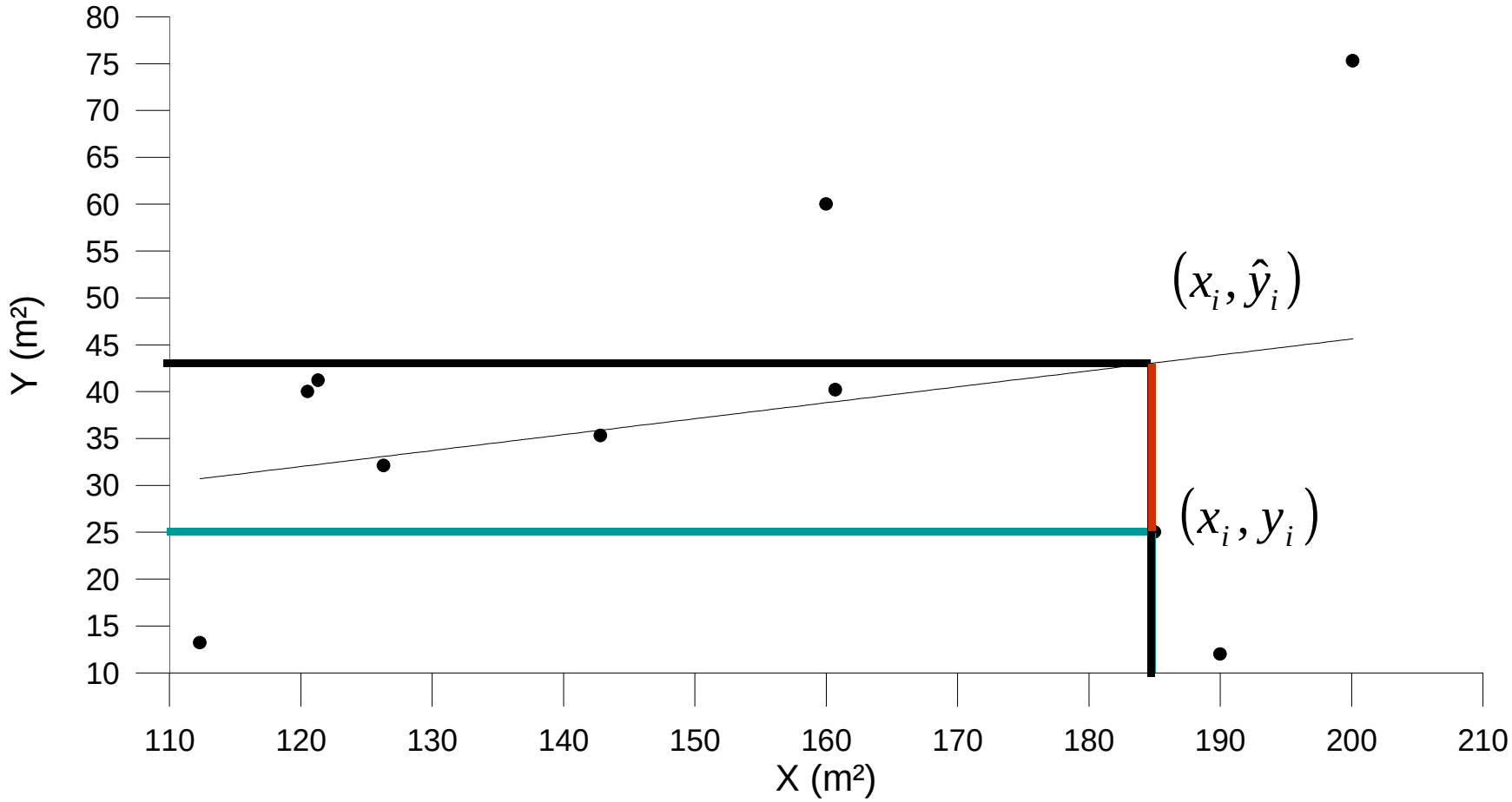
Chercher une relation de corrélation entre deux variables statistiques X et Y.

Quelles sont nos données:

On connaît l'ensemble des couples (x_i,y_i) $i=1,..,n$

Comment faire:

Utiliser la méthode des moindres carrés



$$er_i = y_i - \hat{y}_i$$ Soit très petite

En considérant une régression linéaire:
La méthode des moindre carrés vous permet d'écrire la relation:

$$y=ax+b$$

Et de ce fait vous permet de déterminer les coefficients a et b (la démonstration sera donnée au niveau du cours)

Une corrélation n'a de valeurs que s'il y'a détermination du coefficient de corrélation

•4.4 Covariance

$$Cov(X,Y)=\sum_{i=1}^{i=m}\sum_{j=1}^{j=k} f_{ij}\left(x_i-\overline{\overline{X}}\right)\left(y_j-\overline{\overline{Y}}\right)$$

$$Cov(X,Y)=\sum_{i=1}^{i=n}\frac{1}{n}\left(x_i-\overline{\overline{X}}\right)\left(y_i-\overline{\overline{Y}}\right)$$