BIOSTATISTIQUE 1

1.INTRODUCTION A LA STATISTIQUE

2.OBJECTIF DES MODULES DE BIOSTATIQUE

1. Le module biostatistique I : Statistiques descriptives
2. Le module biostatistique II : Statistiques inférentielles
3. Le module biostatistique III : Statistiques multifactorielles descriptive et inférentielle.

INTRODUCTION

l'ensemble des méthodes mathématiques permettant de :

1. Résumer quantitativement l'information recueillie sur un ensemble d'éléments au moyen d'une investigation exhaustive.

C'est la statistique descriptive,

2. Généraliser à de grands ensembles d'éléments les conclusions tirées des résultats obtenus avec des ensembles beaucoup plus restreints appelés échantillons.

C'est la statistique inférentielle ou probabiliste

Les statistiques ont pour origine le besoin des États pour gérer rationnellement leurs ressources.

Pour cela, il était nécessaire après collecte d'informations (nécessité de techniques de quantification ; production de données nombreuses, organisées en tableaux) de **disposer de méthodes** permettant de définir **les variations**, **les évolutions**, **les ressemblances** ou **les différences** entre régions, entre années, entre catégories.

Exemple de problèmes :

Dénombrement des populations humaines : recensements

Dénombrement des terres et leur répartition.

Calcul et répartition des impôts.

La statistique vise à **décrire**, à **résumer** et à **interpréter** des phénomènes dont le caractère essentiel est **la variabilité**. Elle fournit de la manière la plus rigoureuse possible des éléments d'appréciation utiles à **l'explication ou à la prévision** de ces phénomènes, mais elle n'explique ni ne prévoit aucun d'entre eux (Vigneron 1997).

La méthode statistique permet également d'éprouver la validité de résultats (obtenus, mesurés, collectés) en fonction même de leur variabilité, dans les domaines où les variations sont la règle, c'est-à-dire les domaines de la biologie ***sensu lato***, dans celui des sciences de l'environnement également.

La méthode statistique fournit de ce fait à tous les personnels confrontés à l'interprétation de résultats d'observation ou d'expérimentation, un outil d'interprétation adapté aux conditions particulières de leur domaine d'activité.

OBJECTIF DU MODULE DE BIOSTATIQUE

1. d'acquérir et de parfaire la connaissance des principales notions relatives à l'utilisation des méthodes statistiques,

2 de résoudre des questions empiriques par l'utilisation des tests statistiques,

3. de maîtriser et de compléter les notions de bases des statistiques en vue de les appliquer à des exemples spécifiques aux sciences biologiques, prises dans leur sens général (biologie, médecine, pharmacie, écologie…)

4. d'appliquer ces notions et méthodes sur des données biologiques à partir de logiciels simples

5. d'utiliser des logiciels de statistique et d'apprendre la lecture de leurs résultats.

Les statistiques constituent, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence aux biologistes, en voici, à titre d'exemples quelques unes :

-Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?

-Quelle est la fiabilité d'une mesure ou d'une observation ?

-Quel est le risque ou l'avantage d'un traitement?

-Les conditions expérimentales A sont-elles plus efficaces que celles des conditions de B ?

-Les effets de la variable A sont-ils les mêmes ou diffèrent-ils des effets de la variable B ?

NOTIONS DE BASE ET TERMINOLOGIE

1. Ensemble / Population / Echantillon / Elément / Individu

-L'**ensemble** en statistique, est la collection (finie ou infinie) d'unités, ou d'éléments, sur laquelle porte l'observation. Pour que cet ensemble soit correctement défini, il faut lui donner une définition précise de façon à ce que deux personnes différentes aboutissent toujours à la même liste d'éléments. L'ensemble des éléments observés sera appelé **E**.

-Les **éléments** sont les objets constitutifs de l'ensemble. Ce sont des objets déterminés dont l'appartenance à tel ou tel ensemble E est sans ambiguïté.

Les éléments peuvent être désignés par leur position dans le tableau de données : **1** pour le premier, **i** pour un élément quelconque, **n** pour le dernier élément, **N** pour la somme des éléments constituant l'ensemble.

***Exemple** :

Élément : membre d'une population statistique (prélèvement d'eau, individu...)

La **population** correspond à l'ensemble des **individus** sur lequel porte l'étude ou la prévision, (il est généralement difficile de l'étudier dans sa totalité), et l'**échantillon** représente la fraction de cette population qui est réellement observée ou étudiée :

- Population-cible : ensemble des éléments visés, en principe, par l'échantillonnage.

*** Question**

Quelle est la population-cible ? Il s'agit là de la population sur laquelle on aimerait bien que les conclusions de l'étude portent.

- **Population statistique** : ensemble des éléments effectivement représentés par l'échantillonnage. Les éléments qui la composent se caractérisent par au moins une caractéristique commune et exclusive qui permet de les distinguer sans ambiguïté.

*** Question**

Quelle est la population statistique ? Il faut mentionner la ou les caractéristiques qui permettent de la distinguer de tout autre population statistique.

Population biologique: ensemble des individus d'une même espèce habitant un lieu donné à un moment donné. Notion qui relève davantage de la biologie que de la statistique.

*** Question**

Quelle est la population biologique ? Il faut spécifier le temps et le lieu.

-**Communauté** : ensemble des individus de diverses espèces retrouvés dans un espace et un temps donnés. Notion qui relève davantage de la biologie que de la statistique.

-*Quelle est la communauté ? Il faut spécifier le temps et le lieu.*

-**La notion d'individu** est très large : les éléments d'un échantillon ou d'une population sont appelés généralement des individus, cependant cette notion peut être remplacé par plusieurs dénominations: unité statistique, sujet, objet, élément, observation, mesure, doses,… toutefois, dès que la dénomination est choisi aucune ambiguïté ne doit persistée.

2. Recensement / Echantillonnage

1.Le recensement : qui consiste généralement en un recueil d'informations auprès de tous les individus d'une population (ce qui est très difficile dans le cas de la Biostatistique, mais plus facile dans des études démographique). Il est plus adapté à l'étude des populations. Il consiste en un dénombrement de toutes les personnes ou individus ou attributs d'une population dans sa totalité.

Exemples : population d'un pays ; pollution mondiale ; animaux en voie de disparition ; génome humain ; ….

2.L'échantillonnage : qui consiste généralement en un recueil d'informations auprès de quelques individus ou partie d'une population « **l'échantillon** », (ce qui est généralement le cas en Biostatistique). Parfois l'échantillonnage se fait par sondage (cas en géologie (tremblement de terre), en médecine)

•**Échantillon** : fragment d'un ensemble prélevé pour juger de cet ensemble. Fraction de la population statistique sur laquelle des mesures sont faites pour connaître les propriétés de cette population.

*** Question**

- *quel est l'échantillon ? Quel est son effectif ?*

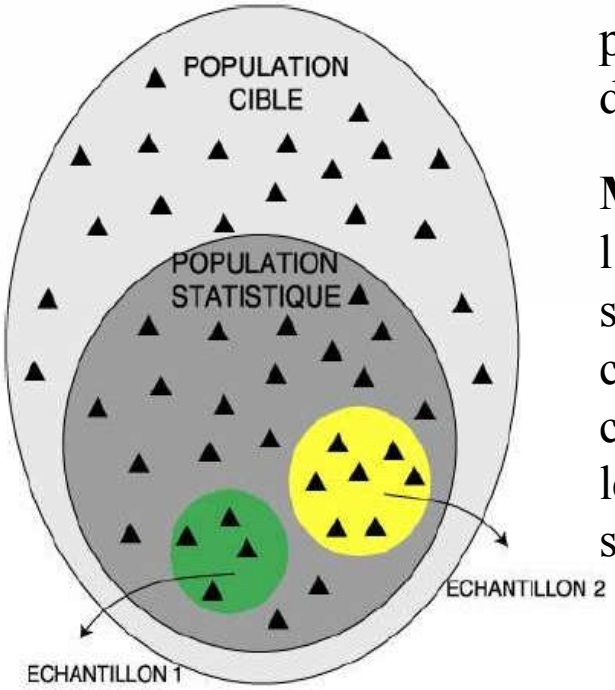


Figure 1 *: Populations et échantillons*

Le caractère

chaque élément de E a une modalité (caractère qualitatif) ou une valeur (caractère quantitatif) et une seule dans X. Ainsi le caractère peut être défini comme une des caractéristiques ou des attributs d'un individu,

Modalité / Mesure : la modalité (respectivement la mesure) est l'une des formes particulière d'un caractère. Les différentes situations où les éléments de E *peuvent se trouver* à l'égard d'un caractère qualitatif considéré, sont les différentes **modalités** du caractère qualitatif X. Dans le cas ou le caractère X est quantitatif, les différentes situations où les éléments de E peuvent se trouver sont des **mesures**.

LA VARIABLE STATISTIQUE

La variable statistique, chaque attribut (ou caractère ou caractéristique) a des modalités, ou peut s'exprimer selon une mesure, celles-ci varient d'un individu à l'autre ou d'un groupe d'individus à un autre groupe d'individus. La variable statistique est le nom que l'on donne à ces caractères (attributs, caractéristiques).

Explicitation de variable en biologie : caractéristique mesurable ou observable sur un élément (variable propre) ou dans son environnement (variable associée).

1.Variable quantitatif : c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut mesurer. Les différentes situations où peuvent se trouver les éléments sont des *mesures*; elles sont ordonnables et la moyenne a une signification, on distingue deux types:

a - Variable discrète ou discontinue : c'est un caractère quantitatif, un tel caractère ne prend qu'un nombre fini de valeurs (valeur entière dénombrable et sans aucune valeur intermédiaire). Les différentes situations où peuvent se trouver les éléments (observations, mesures, valeurs,…) sont des nombres isolés dont la liste peut être établie a priori. Exemple: (nombre d'enfant, nombre de pétales d'une fleur, nombre de dents,..) : (1 ; 2 ; 3 ; 4 ; 5 ;….10 ; 11 ;…)

b - Variable continue : c'est un caractère quantitatif, un tel caractère peut, théoriquement, prendre toutes les valeurs d'un intervalle de l'ensemble des nombre réelles. Toutes les valeurs ne sont pas dénombrables et ne peuvent pas être établit a priori ses valeurs sont alors regroupées en **classes** (taille, temps, poids, vitesse, glycémie, altitude, surfaces,….) (1,60 m ; 1,61 m ; 1,62 m ;…..)

Variable qualitative : c'est un caractère qualitatif, dans ce type de variable les modalités ne sont pas quantifiables (pas mesurable, couleur des yeux, douleurs).

. Exemple : type de relief avec trois modalités (plaine, montagne, plateau).

Dans l'échelle ordinale (de rangement); **caractère ordinal (caractères qui peuvent être** exprimés sur une échelle ordinale)

Dans l'échelle nominale, les nombres ou symboles identifient les groupes auxquels divers objets appartiennent. C'est le cas des numéros d'immatriculation des voitures ou de sécurité

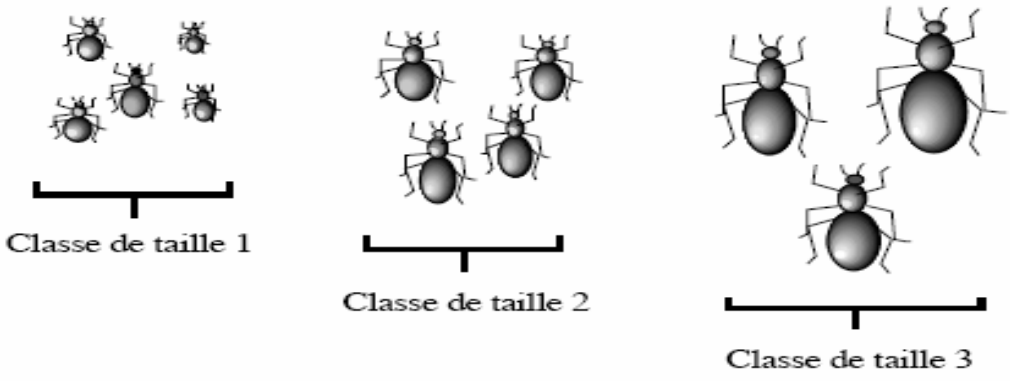
Binaires: 1 - 0 présent - absent

Descripteur "espèce" — Description "espèce présente": 1 ; Description "espèce absente": 0

	Relevé 1	Relevé 2	Relevé 3
Esp. 1	1	0	1
Esp. 2	1	1	0
Esp. 3	0	0	1

Multiples: - **non-ordonnés**, nominaux : ex. couleurs, type de sol...

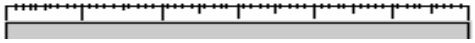
- **ordonnés**: - **semi-quantitatifs**, ordinaux, de rang, : ex. classes de taille (0-10 cm, 10-50 cm, plus de 50 cm...), rang dans une course.



- **quantitatifs** : - **discontinus** (ex.: nombre de personnes dans cette salle, nb. d'individus par espèce...

	Relevé 1	Relevé 2	Relevé 3
Esp. 1	12	0	18
Esp. 2	3	56	0
Esp. 3	0	0	1

- **continus** (ex.: température, longueur, ...)



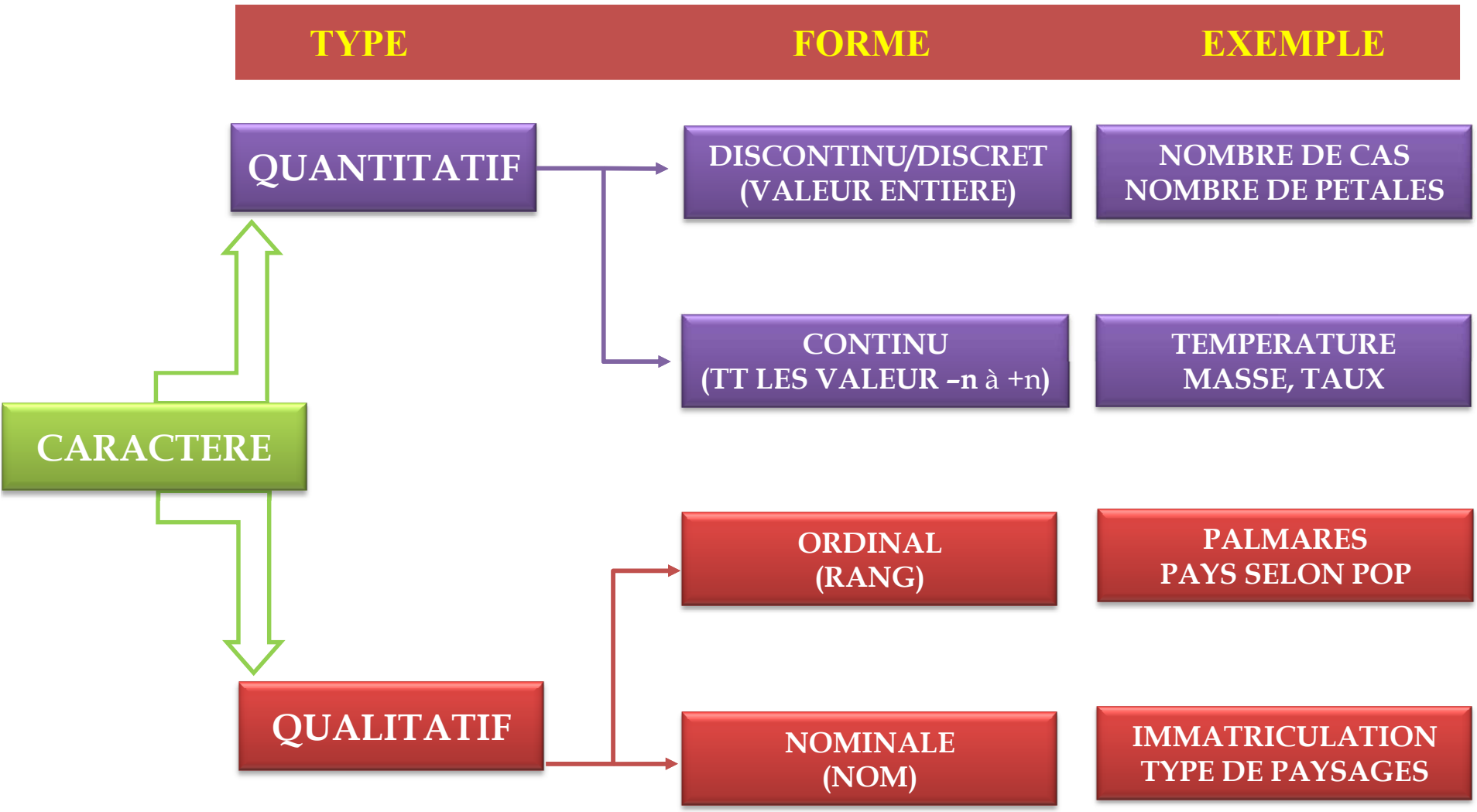


Fig. *Typologie des caractères pour une approche statistique*

6. Variables dépendantes et indépendantes

En statistique on adopte encore une autre dichotomie pour le concept de variable en parlant de variables indépendantes et de variables dépendantes.

1.Les variables indépendantes sont celles qui sont manipulées par l'expérimentateur (l'appartenance au groupe et nous contrôlons les traitements appliqués aux différents groupes).

2.Les variables dépendantes sont celles qui sont mesurés, référencés, exemple de données (survie, résistances, tolérance, performance, …).

Fondamentalement, une étude porte sur les variables indépendantes et les résultats de l'étude (les données) sont les variables dépendantes.

4.3.7. La variabilité et l'incertain en biologie

Toutes les questions, proprement biologique en relation avec les statistiques, reflètent une propriété fondamentale des systèmes biologiques qui est leur variabilité. Cette variabilité est la somme d'une variabilité expérimentale (liée au protocole de mesure) et d'une variabilité proprement biologique. On peut ainsi décomposer la variabilité d'une grandeur mesurée en deux grandes composantes :

Variabilité Totale = Variabilité Biologique + Variabilité Métrologique

4.3.7.1 La variabilité biologique

Elle peut être décomposée en deux termes :

d'une part la **variabilité intra-individuelle**, qui fait que la même grandeur mesurée chez un sujet donné peut être soumise à des variations aléatoires

d'autre part la **variabilité interindividuelle** qui fait que cette même grandeur varie d'un individu à l'autre.

Variabilité Biologique = Variabilité intra-individuelle + Variabilité interindividuelle

4.3.7.2 La variabilité métrologique

Elle peut être elle aussi décomposée en deux termes : d'une part les conditions expérimentales dont les variations entraînent un facteur d'aléas ; et d'autre part les erreurs induites par l'appareil de mesure utilisé.

Variabilité Métrologique = Variabilité Expérimentale + Variabilité instrumentale

TABLEAU STATISTIQUE

Définition

- Un tableau statistique donne pour chaque valeur (ou modalité) de la variable, l'effectif correspondant (c-a-d le nombre de fois où l'on observe la modalité).
- Il intègre en général également la fréquence de chaque modalité ainsi que l'effectif ou la fréquence cumulée (lorsqu'il s'agit d'une variable quantitative !).

- Soit p le nombre de modalités que l'on note en général x_i $(i = 1, \ldots, p)$
- Effectif : n_i ; Effectif total : $\boxed{n = \sum_{i=1}^{p} n_i}$.
- Fréquence : $\boxed{f_i = \frac{n_i}{n},\ f_i \in [0,1]}$ que l'on exprime en %.
- Fréquence cumulée : $\boxed{F_i = \sum_{j=1}^{i} f_j = \frac{1}{n} \sum_{j=1}^{i} n_j}$ en %.

TABLEAUX DES DONNEES

Tableau élémentaire : c'est un tableau à simple entrée où les lignes correspondent aux éléments de l'ensemble étudié et les colonnes aux caractères (ou variables) décrivant ces éléments (Tableau 1 (1.1 et 1.2)).

Observations	***Variables***			
	Variable 1	**Variable 2**	**Variable …**	**Variable p**
Observation 1				
Observation 2				
Observation …				
Observation n				

a : tableau de mesures

	A	B	C
indiv. 1	3	110,5	55,22
indiv. 2	1	109,5	53
indiv. 3	2	108,7	57
indiv. 4	4	107,3	62,8
indiv. 5	0,5	102,1	61,2

b : tableau d'effectif (ou de %)

	A	B	C
indiv. 1	12	125	5
indiv. 2	11	130	6
indiv. 3	9	120	5
indiv. 4	13	110	4
indiv. 5	11	115	5

c : tableau de présence/absence

	A	B	C
indiv. 1	x	x	
indiv. 2	x		x
indiv. 3	x		x
indiv. 4		x	
indiv. 5		x	

d : tableau de contingence

	A	B	C
A	-		
B	1	-	
C	2	0	-

nombre d'objets possédant les caractères pris deux à deux

e : tableau de Burt

	B1	B2	B3
A1	3	2	0
A2	1	2	1
A3	2	0	4

tableau de contingence particulier, issu d'un tableau disjonctif complet et croisant les modalités des variables entre elles

f : tableau de similarité

	indiv. 1	indiv. 2	indiv. 3
indiv. 1	-		
indiv. 2	1	-	
indiv. 3	0,6	0	-

calcul d'un indice de similarité entre individus

STATISTIQUES DESCRIPTIVES

L'objectif poursuivi dans une telle analyse est de 3 ordres :

-tout d'abord, obtenir un contrôle des données et éliminer les données aberrantes,

-ensuite, résumer les données (opération de réduction) sous forme graphique ou numérique,

-enfin, étudier les particularités de ces données.

Ce qui permettra éventuellement de choisir des méthodes plus complexes.

Les méthodes descriptives se classent en deux catégories qui souvent sont complémentaires :

la description numérique et la **description graphique**.

Fréquences absolues, relatives et cumulées (voir tableau exemple)

Désigné par « **F** » ou « **f** » La notion de fréquence peut être exprimée de plusieurs manières :

*Fréquence absolue (effectif)

*Fréquence relative (ou fréquence)

*Fréquences cumulées

	Variables				
Exemples de Fréquences	**x_1**	**x_2**	**x_3**	**x_4**	**Total**
Effectif ou Fréquence absolue (n_i)	**8**	**2**	**9**	**3**	**22**
Fréquence absolue cumulée croissante	**8**	8+2=**10**	10+9=**19**	19+3=**22**	
Fréquence absolue cumulée décroissante	**22**	22-8=**14**	14-2=**12**	12-9=**3**	
Fréquence relative (f_i)	**8/22**	**2/22**	**9/22**	**3/22**	22/22 = **1**
Fréquence relative cumulée croissante	**8/22**	8/22+2/22=**10/22**	**19/22**	**22/22**	
Fréquence relative cumulée décroissante ou fréquence cumulée décroissante	22/22 = **1**	22/22-8/22 =**14/22**	(14-2)/22 =**12/20**	(12-9)/22 =**3/22**	

Caractères quantitatifs discrets

Dans le cas d'un **caractère quantitatif discret,** l'établissement de la distribution des données observées associées avec leurs fréquences est immédiat.

Exemple :

La **cécidomyie** du hêtre provoque sur les feuilles de cet arbre des galles dont ***la distribution de fréquences observées*** est la suivante :

x_i **:** nbr de galles/Feuille	0	1	2	3	4	5	6	7	8	9	10
nbr de feuilles portant $x_{i\ galles}$	182	98	46	28	12	5	2	1	0	1	0
f_i : fréq. relative	0,485	0,261	0,123	0,075	0,032	0,013	0,005	0,003	0	0,003	0
f_i ***cum.*** : fréq. Relative	0,485	0,746	0,869	0,944	0,976	0,989	0,994	0,997	0,997	1	1

La taille de l'échantillon étudié est $n = 375$ feuilles

Caractères quantitatifs continus

(**Mots clés** : Nombre de classes, intervalle entre classe (amplitude), étendu de la variable X)

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**.

La règle de STURGE: $Nombre\,de\,classe = 1 + (3.3 \log N)$

La règle de YULE : $Nombre\,de\,classe = 2.5 * \sqrt[4]{N}$

L'**intervalle** entre chaque classe est obtenu ensuite de la manière suivante

Intervalle de classe = (*X* max - *X* min) / Nombre de classes

Dans le cadre de l'étude de la population de *Gélinottes huppées* (Bonasa *umbellus*), les valeurs de la longueur des plumes principales peuvent être réparties de la façon suivante :

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

Définition du nombre de classes

$$\text{Règle de Sturge} : 1 + (3{,}3 \log 50) = 6.60$$

$$\text{Règle de Yule} : 2.5 \sqrt[4]{50} = 6.64$$

Les deux valeurs sont très peu différentes

Définition de l'intervalle de classe

$$IC = \frac{174 - 140}{6.6} = 5.15\,mm \ldots \text{.que l'on arrondit à 5 mm par commodité}$$

Caractère *X* : x_i l:ongueur de la rectrice bornes des classes en mm	**[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_io:mnbre d'individu par classe	**1**	**1**	**9**	**17**	**16**	**3**	**3**
FR	**0,02**	**0,02**	**0,18**	**0,34**	**0,32**	**0,06**	**0,06**
FRCC	**0,02**	**0,04**	**0,22**	**0,56**	**0,88**	**0,94**	**1**

Représentations graphiques et statistique descriptive

Les représentations graphiques sont très importantes en statistique descriptive. Elles ont l'avantage de **renseigner immédiatement** sur l'allure générale de la distribution. Elles **facilitent l'interprétation** des données recueillies.

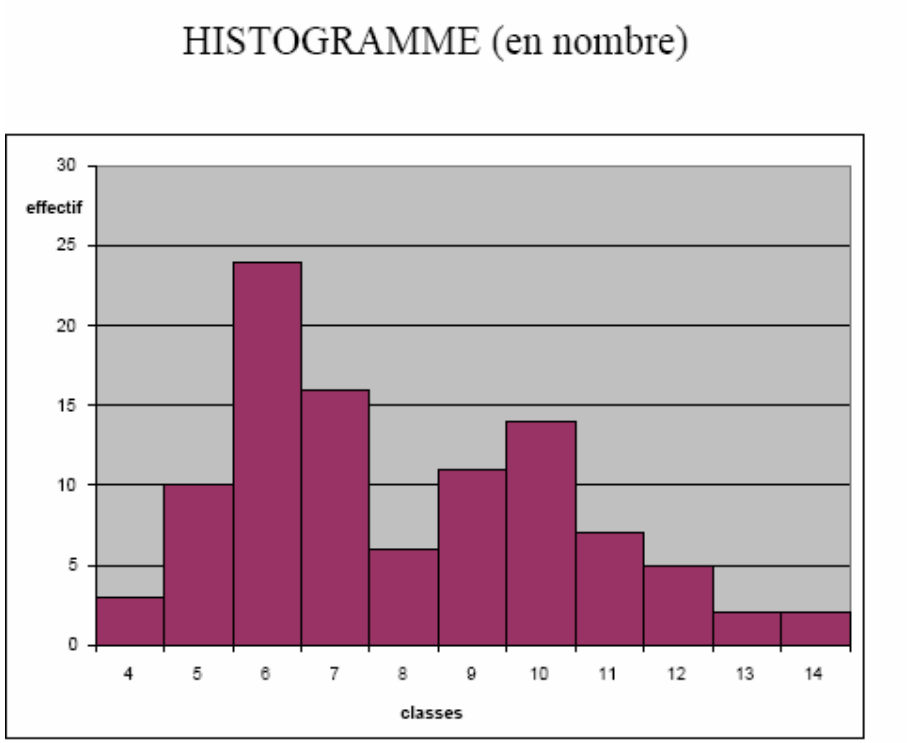
La représentation graphique des données montre la **forme générale** de la distribution et donne **une image** de la grandeur des nombres qui constituent les données.

L'histogramme

Définition : L'histogramme consiste à faire figurer les effectifs d'une variable par classe de valeur. Il est représenté quand la variable est quantitative continue par des ***rectangles*** dont la surface (et non la hauteur) est proportionnelle aux effectifs.

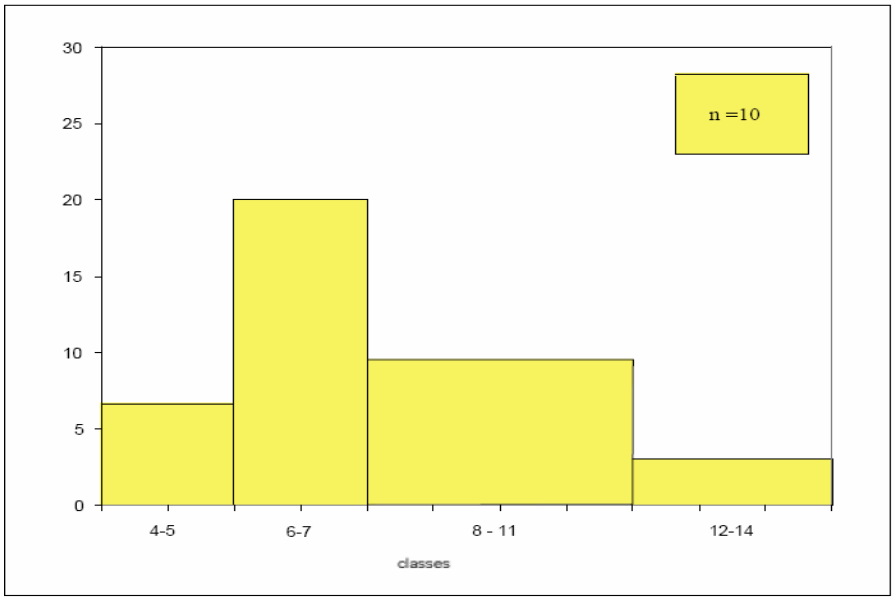
Application

classes (mettre l'unité)	effectif (en nombre)
4	3
5	10
6	24
7	16
8	6
9	11
10	14
11	7
12	5
13	2



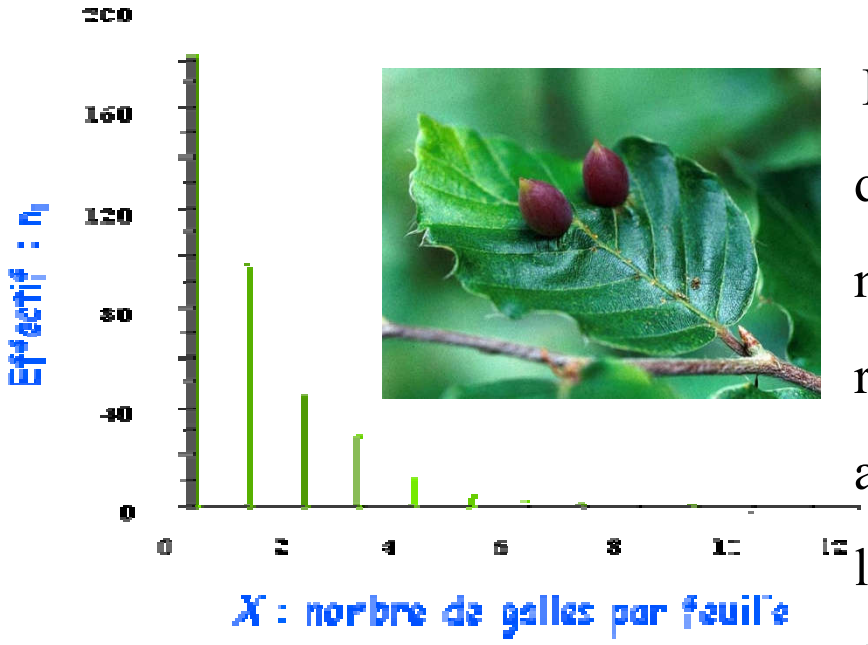
Les classes peuvent être définies d'intervalles égaux ou non.

Dans ce dernier cas, seule la surface sera proportionnelle à l'effectif (et non la hauteur)



Caractères quantitatifs discrets

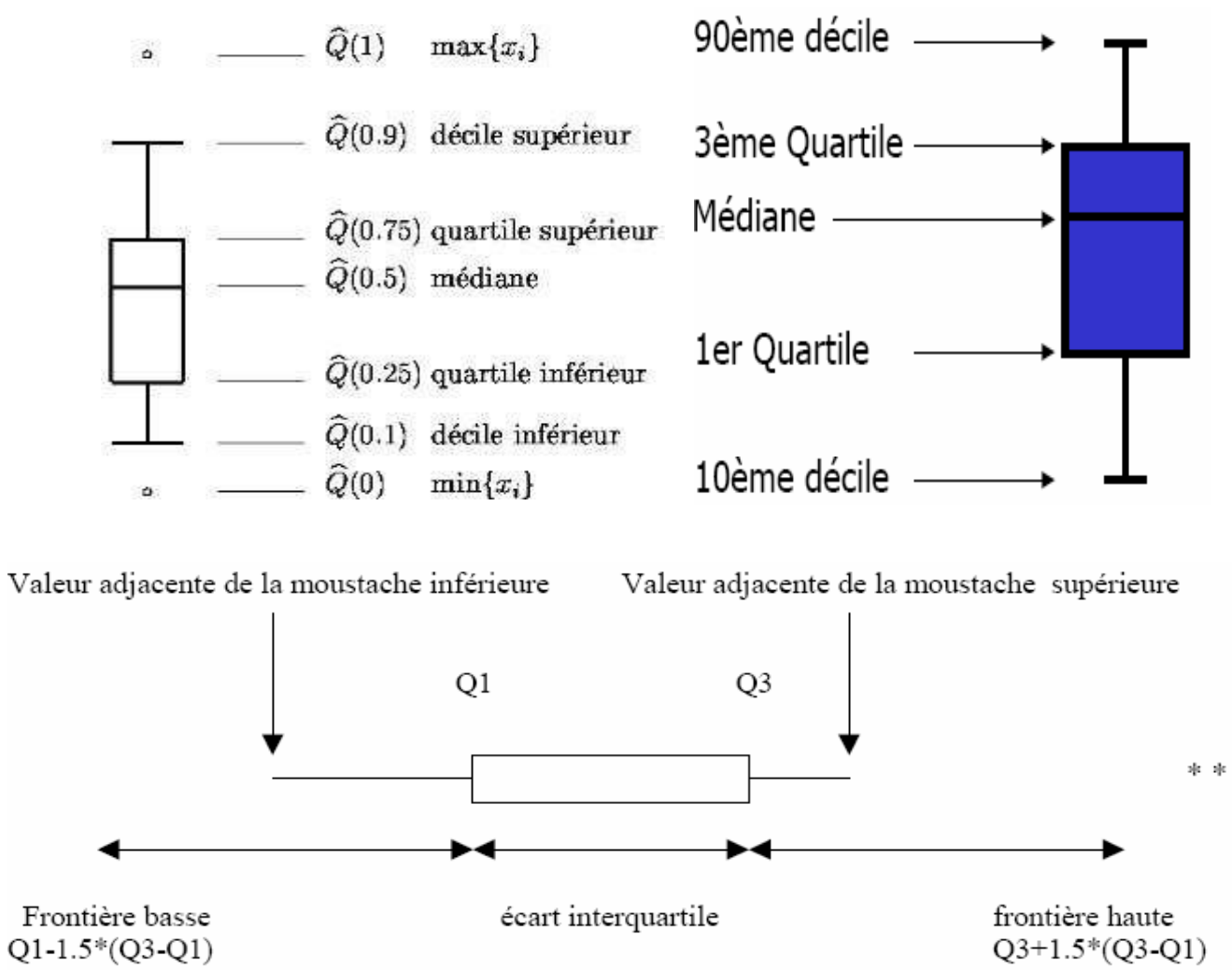
Pour les caractères quantitatifs discrets, la représentation graphique est le **diagramme en bâtons** où la hauteur des bâtons correspond à l'effectif n_i associé à chaque modalité du caractère x_i.



Dans l'exemple de la **cécidomyie** du hêtre, la distribution des fréquences observées du nombre de galles par feuille peut être représentée par **un diagramme en bâtons** avec en ordonnée les **effectifs** n_i et en abscisse les différentes **modalités** de la variable étudiée.

Barre à moustache - Box Plot

Une "boîte à moustaches" (traduction française du terme "Box and Whiskers Plot", ou en abrégé "Box Plot") est une représentation graphique de quelques paramètres de distribution d'une variable, inventée par Tukey en 1977. C'est une représentation graphique d'une variable quantitative qui permet d'appréhender (résumer une distribution empirique) la dispersion d'un échantillon.



Exemples de représentations graphiques: boîtes à moustache (box-plot)

Paramètres utilises dans la statistique descriptive

Trois aspects sont essentiels à l'interprétation d'une distribution :

-**Paramètre de position** : le centre de la distribution et la répartition autour d'une valeur centrale (moyenne, mode, médiane, quantiles, ..)

-**Paramètre de dispersion ou d'étendue** : les valeurs sont-elles dispersées ou concentrées ?

-**Paramètre de forme** : la forme de la distribution : la symétrie, l'aplatissement

Paramètre de position et valeurs centrales

Le but des valeurs centrales est de résumer en une seule valeur l'ensemble des valeurs d'une distribution statistique. Il existe quatre valeurs de positions :

- Le mode (Mo),
- La moyenne ($\overline{X}$ ou μ)
- La médiane ou le médian (Me ou Md)
- Les fractiles (Quantiles) (Qn)

Parmi ces valeurs les trois premières sont des valeurs de position centrales

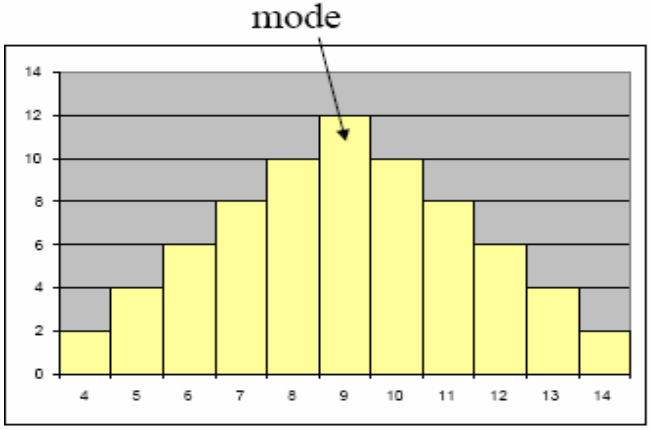
Le mode

Le mode, ou valeur dominante, est la valeur la plus fréquente d'une distribution.

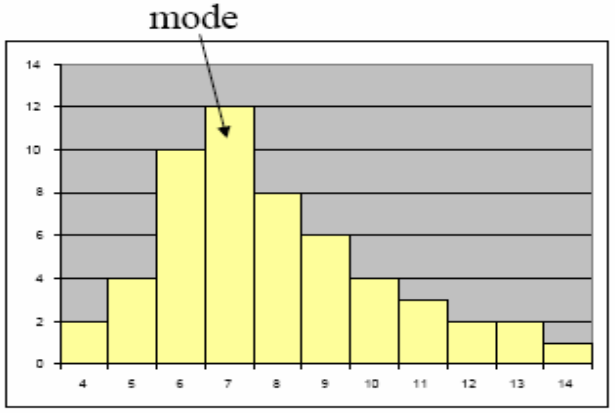
Cette valeur se calcule toujours à partir d'un dénombrement des modalités du caractère. Il faut donc distinguer le cas des caractères discrets et des caractères continus (voir notions de bases).

***Caractère qualitatif et caractère discret** : le mode est la modalité ou la valeur qui a **la fréquence simple la plus élevée** (ou l'effectif le plus élevé, ce qui revient au même).

***Caractère quantitatif continu :** Le mode est alors le centre de la classe modale, c'est à dire de **la classe qui a la fréquence moyenne la plus élevée**.



Distribution unimodale symétrique



Distribution unimodale asymétrique

Le mode

Application : Cas de calcul des modes

-Données rangées : le mode est la valeur de la donnée qui apparaît le plus fréquemment (celle qui a le plus d'occurrences) :

140 ; 141 ; 144 ; 144 ; 148 ; 148 ; **152** ; **152** ;**152** ; 154 ; 155 ; 158 ; 158 ; 161 ; 170 ; 172

Le mode est 152 car il possède le plus grand nombre d'occurrences (il est référencé 3 fois)

Données condensées : le mode est la valeur de la donnée qui possède la fréquence

Modalités xi (age en années)	**14**	**16**	**18**	**21**	**22**	**24**	**25**	**Total**
Fréquences absolues	5	**12**	10	8	11	7	3	56
Fréquences relatives	0,089	**0,214**	0,179	0,143	0,196	0,125	0,054	1,000

Dans cette série statistique, le mode est égal à Mo = 16 ans

Cas 3 : Données groupées en classes : la classe modale est la classe ayant la plus haute fréquence (relative ou absolue).

Dans le tableau des classes relatives à la longueur de la rectrice de Bonasa *umbellus* , la classe modale est [155mm-160mm[. Il est possible de calculer de façon plus précise le mode en appliquant la formule suivante :

$$M_0 = bm_0 + (\frac{\Delta_1}{\Delta_1 + \Delta_2})Lm_0$$

Δ1 = différence entre l'effectif de la classe modale et l'effectif de la classe précédente.
Δ2 = différence entre l'effectif de la classe modale et l'effectif de la classe qui suit.

bmo : Borne inférieur de la classe modale

Lmo : largeur de la classe modale

Caractère *X* : x_i l:ongueur de la rectrice bornes des classes en mm	**[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_io:mnbre d'individu par classe	**1**	**1**	**9**	**17**	**16**	**3**	**3**
FR	**0,02**	**0,02**	**0,18**	**0,34**	**0,32**	**0,06**	**0,06**
FRCC	**0,02**	**0,04**	**0,22**	**0,56**	**0,88**	**0,94**	**1**

$\Delta\mathbf{1} = (17\text{-}9) = 8$; $\Delta\mathbf{2} = (17\text{-}16) = 1$; **bmo** = 155 ; **Lmo** = 5

$$M_0 = 155 + (\frac{8}{8+1}) * 5 = 159mm$$

Formalisation mathématique de la moyenne arithmétique

La moyenne arithmétique, noté $\overline{X}$ ou μ, est la mesure la plus commune de tendance centrale, elle se définit comme la somme des scores divisée par le nombre de scores.

Elle est calculée pour les caractères quantitatifs.

*** Calcul à partir du tableau élémentaire :**

La moyenne est la somme des valeurs divisée par le nombre d'éléments : $\overline{X} = \frac{\sum x}{N}$

Elle est calculée pour les caractères quantitatifs.

*** Calcul à partir du tableau de dénombrement :**

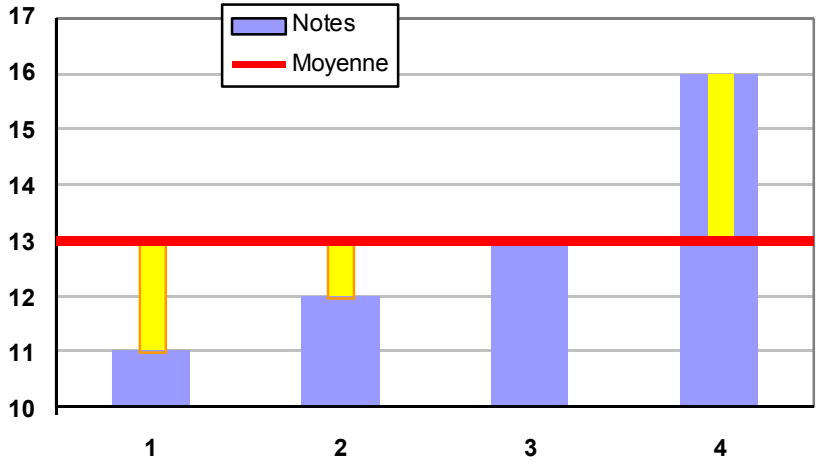
On effectue une moyenne pondérée en assimilant chaque classe j à son centre X_j et en pondérant par l'effectif n_j de la classe.

$$\overline{X} = \frac{\sum_{j=1}^{k} x_j * n_j}{N}$$

Exemple avec illustration

Soit les valeurs de quatre notes : 10, 12, 13 et 16, la moyenne arithmétique est:

(11 + 12 + 13 + 16) / 4 = 13



La moyenne arithmétique donne une valeur telle que la somme des écarts (rectangles jaunes) est nulle

Exemple

Soit la série statistique suivante :

Valeurs	0	1	2	3	4
effectifs	1	2	1	4	2

$$\overline{X} = \frac{(0*1)+(1*2)+(2*1)+(3*4)+(4*2)}{10} = \frac{24}{10} = 2.4$$

Remarque :

Si les données ont été regroupées en classes, on ne peut calculer la valeur exacte de la moyenne. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu.

Dans les séries statistiques suivantes déterminer les moyennes

a) Tableau de fréquences

Valeurs	12	13	14	15	16
fréquences	0,05	0,17	0,43	0,30	0,05

$$\overline{X} = (12*0.05) + (13*0.17) + (14*0.43) + (15*0.30) + (16*0.05) = 14.13$$

b) Données réparties en classes

Classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
Effectifs	7	12	14	2

$$\overline{X} = \frac{(2.5*7) + (7.5*12) + (12.5*14) + (17.5*2)}{35} = 9.0714$$

Cas général : Soit ∞ un réel quelconque :

- Si l'on ajoute ∞ à toutes les données, la moyenne augmente d'∞
- Si on multiplie toutes les données par ∞, la moyenne est multipliée par ∞
- Si on divise toutes les données par ∞, la moyenne est divisée par ∞

Ex : Soit la série : 10, 12, 14. $\overline{X} = 12$

Ajoutons 2 : la nouvelle série est : 12, 14, 16 $\overline{X} = 14$

Divisons par 2 : la nouvelle série est : 5, 6, 7. $\overline{X} = 6$

La médiane et la classe médiane

On appelle **médiane** la valeur "du milieu". On dit qu'elle partage la série statistique en deux moitiés : il y a autant de valeurs en dessous qu'au dessus. (C'est la donnée qui permet de diviser une série ordonnée d'une façon croissante en 2 parties égales (50%, 50%). La médiane ne peut être calculée que pour les caractères quantitatifs.

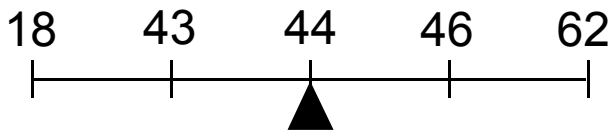
2.Médiane, pour les données rangées : Les valeurs du caractère X étant classées par ordre croissant, la médiane est la valeur du caractère qui partage l'ensemble décrit par X en deux sous ensembles d'effectifs égaux : 50 % des éléments ont des valeurs de X supérieures à $X_{méd}$ et 50% prennent des valeurs inférieures.

- Méthode

Soit une série statistique d'effectif total *n*, rangée par ordre croissant. Pour déterminer son rang, il y a 2 cas :

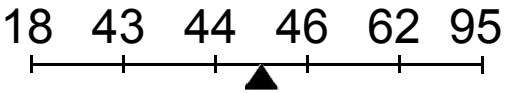
Médiane

•si n est impair : la médiane est la valeur de rang $\frac{n+1}{2}$



méd = 44

•si n est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre



méd = 45

Cas de données discrètes "en vrac" ***10, 7, 12, 18, 16, 15, 5, 11, 11, 20, 15, 11, 18, 14***

Ordonnons la série par ordre croissant : 5, 7, 10, 11, 11, 11, 12, 14, 15, 15, 16, 18, 18, 20

Il y a 14 termes or la valeur de rang est 14+1/2 = 7.5.

La médiane est donc la demi somme des $7^{ème}$ et $8^{ème}$ termes : médiane 12+14/2 = 13

Médiane

Médiane, pour les données condensées : La définition est la même, elle correspond dans ce cas à la première modalité ou valeur dont la fréquence relative cumulée dépasse 0,500 ou l'effectif cumulé dépasse les 50%.

Méthode :

Il faut calculer les fréquences ou les effectifs cumulés dès que celle-ci atteint respectivement 0.5 ou 50% il suffit de choisir le nombre à mi chemin entre la modalité ou valeur concernée et la suivante.

Médiane

Cas d'un tableau d'effectifs

On ordonne le tableau, et on cherche l'élément qui partage la distribution en deux parties égales: on repère l'élément qui a le rang (N+1)/2 pour le caractère X. Si la distribution a un nombre impair d'éléments on trouve une valeur unique qui est la médiane, si la distribution a un nombre pair d'éléments, on trouve deux valeurs qui déterminent un **intervalle médian** : on prend alors pour médiane le centre de cet intervalle médian.

Valeurs	1	2	**3**	4	5	6
Effectifs	6	11	**25**	19	15	5
effectifs cumulés	6	17	**42**	61	76	81
Fréquence	0,07	0,14	**0,31**	0,23	0,19	0,06
fréquences cumulées	0,07	0,21	**0,52**	0,75	0,94	1,00

L'effectif total est de 81 or la valeur de rang 81+1/2 = 41

La médiane est donc le 41ème terme : médiane = 3

Médiane, pour les données réparties par classes

Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera classe médiane, la classe qui la contient (et permet donc d'en donner un encadrement).

La classe médiane est la première classe où la fréquence cumulée est supérieure à 0,50 ou à 50%

Classe	[0 ; 2[	[2 ; 4[	[4 ; 6[	[6 ; 8]
Fréquence	10%	38%	45%	7%
Fréquence cumulée	10%	48%	93%	100%

48% des valeurs sont strictement inférieures à 4

Et 93% des valeurs sont strictement inférieures à 6 La classe médiane est donc la classe [4 ; 6[

On peut donc en déduire l'encadrement suivant $4 < méd < 6$

Méthode de calcul

Pour préciser la valeur de la médiane, il faut supposer que toutes les données sont réparties uniformément (c'est-à-dire que les données sont réparties sur un continuum).

On repère la classe qui contient la médiane, puis on réalise une interpolation linéaire pour estimer la valeur de celle-ci selon la formule suivante :

$$Md = Bmd + \left(\frac{0.5 - Fmd_{-1}}{Fmd} \right) * Lmd$$

Bmd : Borne inférieure de la classe médiane

Fmd-1 : Fréquence relative cumulée de la classe qui précède la classe médiane.

Fmd : Fréquence relative de la classe médiane.

Lmd : largeur, amplitude des classes

Application pour l'exemple précédent : $Md = 4 + \left(\frac{0.5 - 0.48}{0.45} \right) * 2 = 4.088$

Quantiles : Mesures de position statistique en référence à la médiane

Il a été vu précédemment que la médiane partage la distribution des fréquences en 2 parties égales. Il est possible de partager une distribution de fréquence en 4 parties égales (quartiles), en 10 parties égales (déciles), en 100 parties égales (centiles), en n parties égales….

1.Définition des quantiles : on appelle quantiles les valeurs du caractère qui définissent les bornes d'une partition en classes d'effectifs égaux.

2.Les quartiles sont les trois valeurs qui permettent de découper la distribution en quatre classes d'effectifs égaux. On les notes X_{Q1}, X_{Q2} et X_{Q3}

Représentation des quartiles

Partition du caractère	X_{min} intervalle interquartile 1	X_{Q1} ¼ Quartile inférieur	intervalle interquartile 2	X_{Q2} ½ Médiane	intervalle interquartile 3	X_{Q3} ¾ Quartile supérieur	X_{max} intervalle interquartile 4
Fréquence des éléments	**25%**		**25%**		**25%**		**25%**

-**Q1 :** quartile inférieur, 25% des valeurs de la variable lui sont inférieures et 75% lui sont supérieures

-**Q2 :** médiane, 50% des valeurs de la variable lui sont inférieures et 50% lui sont supérieures

-**Q1 :** quartile supérieur, 75% des valeurs de la variable lui sont inférieures et 25% lui sont supérieures

Remarque : X_{Q2} est égal à la médiane.

Les **déciles** sont les 9 valeurs de X qui permettent de découper la distribution en dix classes d'effectifs égaux. 0n les note $X_{d1}...X_{d9}$.

Représentation des déciles

Partition du caractère	xmin **Int-1**	**xd1** 1/10	**Int-2**	**xd2** 1/20	**Int-3**	**xd3** 1/30	…..	**xd8** 1/20	**Int-9**	**xd9** 9/10	xmax **Int-10**
Fréquence des éléments	**10%**		**10%**		**10%**				**10%**		**10%**

Int-(intervalle interdécile)

-**D1 :** décile inférieur, 10% des valeurs de la variable lui sont inférieures et 90% lui sont supérieures

-**D2 :** 20% des valeurs de la variable lui sont inférieures et 80% lui sont supérieures

-**D3 :** 30% des valeurs de la variable lui sont inférieures et 70% lui sont supérieures

- D4 :….

-**D5 :** médiane, 50% des valeurs de la variable lui sont inférieures et 50% lui sont supérieures

-**D9 :** décile supérieur, 90% des valeurs de la variable lui sont inférieures et 10% lui sont supérieures

Les **centiles** sont les 99 valeurs de X qui permettent de découper la distribution en 100 classes d'effectifs égaux. 0n les note $X_{c1}...X_{c99}$.

Remarques

Les différentes mesures de position (quartile, décile,…) ne sont que des cas particuliers des centiles.

Les centiles sont donc très utiles pour déterminer les valeurs des autres mesures de positions

Q1= C25 = 25ème centile
Q2= C50 = D50 = 50ème centile = Médiane
Q3 = C75 = 75ème centile
D1= C10 = 10ème centile
D2= C20 = 20ème centile
…
D9 = C90 = 90ème centile

Calculs des quantiles

Nous nous limiterons aux cas des centiles car nous pouvons facilement faire des correspondances avec les autres mesures de positions.

Détermination des valeurs de la variable à partir d'un rang centile données. Cα **rang du centile** (le rang est donnée, quelle est la valeur de la variable correspondant à ce rang ?)

Cas des données rangées :

Cα : rang du centile : Il correspond à la donnée dont le rang est l'entier qui suit $\frac{N\alpha}{100}$

N : nombre total de valeurs dans la série statistique

α : le rang du centile

Exemples :

Soit la série statistique suivante :

58 ; 59 ; 64 ; 64 ; 64 ; 68 ; 71 ; 71 ; 79 ; 82 ; 82 ; 85 ; 92 ; 92 ; 92 ; 95

-Trouver les centiles suivants : C15 ; C40

-Trouver les quartiles : Q2 et Q3

-Pour centile C15

α **=15, le rang de la donnée est déterminé par la formule** $\frac{N\alpha}{100} = \frac{16*15}{100} = 2.4$

La valeur n'est pas un entier, le rang est donc le premier entier suivant 2,4 ainsi C15 correspond au rang 3, ce dernier correspond à la valeur : 64

Pour centile C40 (qui correspond au décile 4)

α **=20 le rang de la donnée est déterminé par la formule** $\frac{N\alpha}{100} = \frac{16*20}{100} = 6.4$

La valeur n'est pas un entier, le rang est donc le premier entier suivant 6,4 ainsi C40 (ou D4) correspond au rang 7, ce dernier correspond à la valeur : 71

$$\text{Pour quartile } Q_2 : \frac{N50}{100} = \frac{16*50}{100} = 8$$

Q2 correspond à la moyenne des valeurs du au rang 8 (qui correspond à la valeur 71) et le rang 9 (qui correspond à la valeur 79)

$$Q_2 = \frac{71+79}{2} = 75$$

$$\text{Pour quartile } Q_3 : \frac{N75}{100} = \frac{16*75}{100} = 12$$

Q3 correspond à la moyenne des valeurs du au rang 12 (qui correspond à la valeur 85) et le rang 13 (qui correspond à la valeur 92)

$$Q_3 = \frac{85+92}{2} = 88.5$$

Cas des données condensées :

La méthode est identique à la précédente, mais il est aussi possible de travailler avec les fréquences relatives.

Dans le cas de détermination avec les fréquences, Cα correspond à la première modalité dont la fréquence cumulée dépasse $\frac{\alpha}{100}$

Dans le cas où $\frac{\alpha}{100}$ est un entier, il suffira de choisir le nombre à mi-chemin entre la modalité concernée et la suivante

Calculons C69 pour des données condensées

Xi	Ni	eff cum	Fi	Fi (fr cum)
128	8	8	0,11	0,11
145	13	21	0,18	0,29
160	14	35	0,19	0,49
180	**16**	**51**	**0,22**	**0,71**
195	11	62	0,15	0,86
197	7	69	0,10	0,96
209	3	72	0,04	1,00
Somme	**72**			

Choisir C = 69

calcul avec les effectifs

C 69 = 49,68

calcul avec les fréquences

C 69 = 0,69

Pour le calcul avec les effectifs : la formule est la suivante : (N=72) $C_{69} : \frac{72*69}{100} = 49.68$

C69 correspond à la modalité occupant le rang 50 dans la distribution.

Elle correspond donc à la valeur 180

Pour le calcul avec les fréquences : la formule est la suivante : $C_{69} : \frac{\alpha}{100} = 0.69$

C69 correspond à la modalité dont la fréquence relative cumulée dépasse 0,69.

Dans la distribution, cette fréquence correspond à la valeur 180

Cas des données groupées en classes :

La classe contenant C α correspond à la première classe où la fréquence cumulée atteint ou dépasse $,\frac{\alpha}{100}$ par référence à la formule du calcul de la médiane (vue précédemment) il est possible d'écrire la formule suivante de Cα

$$C_\alpha = Bc_\alpha + \frac{(\frac{\alpha}{100} - Fc_{\alpha-1})}{Fc_\alpha} Lc_\alpha$$

Bcα : Borne inférieure de la classe contenant cα

Fcα -1 : Fréquence relative cumulé de la classe qui précède la classe contenant cα

Fcα : Fréquence relative de la classe contenant cα.

Lcα : largeur, amplitude de la classe contenant cα

Calculer C80 des classes suivantes

Classes (cm)	Mi	ni	eff cum	fi	Fi (freqcum)
[130-140[	135	12	12	0,12903	0,1290
[140-150[	145	20	32	0,21505	0,344
[150-160[	155	24	56	0,25806	0,602
[160-170[	**165**	**21**	**77**	**0,22581**	**0,828**
[170-180[	175	11	88	0,11828	0,946
[180-190[	185	5	93	0,05376	1,000
Somme		**93**		1,00000	

La classe contenant C_α (C80) est la première classe où $F_i > \frac{\alpha}{100} = \frac{80}{100} = 0.80$

C80 correspond à la classe [160-170[

Calcul de la valeur de la modalité correspondant à C80

$$\text{La classe contenant } C_\alpha = Bc_\alpha + \frac{(\frac{\alpha}{100} - Fc_{\alpha-1})}{Fc_\alpha} Lc_\alpha$$

Bc_α : Borne inférieure de la classe contenant $C_{80} = 160$ cm

$Fc_{\alpha\text{-}1}$: Fr - équence relative cumulé de la classe qui précède la classe contenant $C_{80} = 0.828$

Fc_α : Fréquence relative de la classe contenant $C_{80} = 0{,}22581$

Lc_α : largeur, amplitude de la classe contenant $C_{80} = 170 - 160 = 10$cm

$$C_{80} = 160 + \frac{(\frac{80}{100} - 0.602)}{0.22581} 10 = 168.761cm$$

Détermination du rang centile à partir d'une valeur donnée de la variable.

Cette détermination est le processus inverse ce qui consiste à recherche Cα pour une valeur connu Xi d'une série statistique X.

a) Cas des données rangées ou condensées

Il suffit de calculer simplement le pourcentage des données inférieures à la valeur (ou observation) donnée.

Exemple 1

Série ordonnée croissante

Dans les valeurs de la glycémie de la série statistique suivante, trouver le centile Cα de la valeur 0,96 :

0,6 ; 0,6 ; 0,65 ; 0,7 ; 0,72 ; 0,72 ; 0,72 ; 0,74 ; 0,75 ; 0,75 ; 0,76 ; 0,78 ; 0,78 ; 0,8 ; 0,8 0,83 ; 0,83 ; 0,84 ; 0,84 ; 0,84 ; 0,9 ; **0,96** ; 1,01 ; 1,02 ; 1,1 ; 1, 15 ; 1,16 ; 1,18 ; 1,2.

Il s'agit de trouver le pourcentage des données dont la valeur de la glycémie est **inférieure** à **0,96g/l** Cette valeur est à la 22 positions (22ème valeur de la série ordonnée de façon croissante), il y a 21 valeurs de la glycémie inférieures à 0,96g/l sur un total de 29 valeurs (N= 29),le pourcentage est donc: (21/29)*100 = 72.41% ainsi

le rang centile Cα de la valeur de la glycémie de 0,96g/l est de 72

(la valeur de 0,96g/l de glycémie correspond au centile C72)

Exemple 2

Tableau de distribution condensée

Dans le tableau de distribution des valeurs de la glycémie suivante trouver le centile Cα de la valeur 1,1g/l :

xi (g/l)	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	1	1,05	1,1	1,15	1,2
ni	8	12	24	26	32	32	28	26	21	24	20	18	11
ni (cum)	8	20	44	70	102	134	162	188	209	**233**	253	271	282

Il s'agit de trouver le pourcentage des données dont la valeur de la glycémie est inférieure à 1,1g/l :

Cette valeur est à la 253 positions (253 ème valeurs des effectifs cumulés), il y a 233 valeurs de la glycémie inférieures à 1,1g/l sur un total de 282 valeurs (N= 282), le pourcentage est donc de : (233/282)*100 = **82.62%**

le rang centile C α de la valeur de la glycémie de 1,1g/l est de 82

(au moins 82% des valeurs de la glycémie sont inférieures à 1,1g/l).

Exemple 3

Des données rangées en classes

Le rang centile C α d'une donnée (ou observation) est obtenu par la formule suivante:

$$RC_{\alpha} = 100 * \left[\left(\frac{x_r - b_r}{L_r} \right) F_r + F_{r-1} \right]$$

Xr : la donnée dont on recherche le rang centile C α

br : borne inférieure de la classe contenant Xr

Lr : largeur de la classe contenant Xr

Fr : fréquence relative de la classe contenant Xr

Fr-1 : fréquence relative cumulée de la classe qui précède la classe contenant Xr

Dans le tableau de distribution des valeurs de la glycémie suivante trouver le centile C α de la valeur 0.8g/l :

xi (g/l)	[0,6-0,7[	[0,7-0,8[	[0,8-0,9[	[0,9-1,0[	[1,0-1,1[	[1,1-1,2[	[1,2-1,3[	[1,3-1,4[	[1,4-1,5[	Somme
Ni	20	18	26	28	29	25	21	20	21	**208,00**
fi	0,10	0,09	0,13	0,13	0,14	0,12	0,10	0,10	0,10	**1,00**
ni:cum	20	38	64	92	121	146	167	187	208	
fi cum	0,10	**0,18**	0,31	0,44	0,58	0,70	0,80	0,90	1,00	

Pour la valeur de la glycémie de 0,81g/l

0,8g/l se situe dans la classe [0,8-0,9[, le rang centile de 0,8g/l est l'entier inférieur à :

$$RC_{\alpha} = 100 * \left[\left(\frac{0.81 - 0.8}{0.1} \right) 0.13 + 0.18 \right] = 19.3$$

Le rang centile de 0,81g/l est 19

ainsi au moins 19% des données sont inférieures à 0,81g/l

Avantages et inconvénients des différentes valeurs centrales :

Le statisticien Yule (XIXème siècle) a défini six propriétés souhaitables pour les valeurs centrales. Le tableau ci-dessous permet de montrer les avantages et inconvénients des trois valeurs centrales (Mode, Médiane, Moyenne arithmétique)

Propriétés	**Mode**	**Médiane**	**Moyenne**
1) est définie de façon objective	+	+	+
2) dépend de toutes les valeurs observées	-	-	+
3) a une signification concrète	+	+	-
4) est simple à calculer	+	+	+
5) est peu sensible aux fluctuations de l'échantillon	-	+	-
6) se prête au calcul algébrique	-	-	+

Paramètre de dispersion

Dispersion statistique : On appelle dispersion statistique, la tendance qu'ont les valeurs de la distribution d'un caractère à s'étaler, à se disperser, de part et d'autre d'une valeur centrale.

On distingue la dispersion absolue (mesurée dans l'unité de mesure du caractère) et la dispersion relative (mesurée par un nombre sans dimension).

Les paramètres de dispersion absolue

Les paramètres de dispersion absolue indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur centrale de référence. Un paramètre de dispersion absolue s'exprime toujours dans l'unité de mesure de la variable considérée. Les quatre paramètres de dispersion absolue les plus courants sont :

-l'étendue,

-l'intervalle inter quantile (écarts inter quantiles),

-l'écart absolu moyen

-l'écart type.

L'étendue de la variation: l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution :

Etendue de X = Xmax - Xmin

Plus l'étendu est grande plus les valeurs sont dispersées.

Exemple : l'étendue est donnée par la valeur minimale et la valeur maximale : dans le cas de l'exemple précédent il s'agit de la différence : 14 mm - 4 mm = 10 mm

a) cas de données rangées : L'étendu de la distribution de la série statistique :

0,5 g/l; 0,58 g/l; 0,65 g/l; 0,7 g/l; 0,72 g/l; …….; 1, 15 g/l; 1,16 g/l; 1,18g/l ; 1,2g/l. La plus grande valeur est: 1,2g/l, La plus petite valeur est :0,6g/l

L'étendue de la variation : 1,2-0,5 = 0,7

[0,6-0,7[	[0,7-0,8[	[0,8-0,9[	[0,9-1,0[	[1,0-1,1[	[1,1-1,2[	[1,2-1,3[	[1,3-1,4[	[1,4-1,5[

La dernière classe a comme borne supérieure = 1,5

La première classe a comme borne inférieure =0,6

L'étendue de la variation est : 1,5 – 0,6 = 0,9

L'intervalle interquartile

L'intervalle interquartile est l'étendue de la distribution sur laquelle se trouvent concentrée la moitié des éléments dont les valeurs de X sont les plus proches de la médiane. On exclut alors de la distribution les 25% des valeurs les plus faibles et les 25 % des valeurs les plus fortes de X. Cet intervalle se note:$(X_{q3}-X_{q1})$.

L'intervalle inter-décile est l'étendue de la distribution sur laquelle se trouvent concentrés 80% des éléments dont les valeurs de X sont les moins différentes de la médiane. On exclut alors de la distribution les 10 % des valeurs les plus faibles et les 10% des valeurs les plus fortes. Il se note $(X_{d9}-X_{d1})$.

Mesures de la dispersion statistique en utilisant l'écart semi-interquartile

Cet écart mesure la moitié de l'étendue de la moitié centrale des données Il est calculé selon la formule suivante :

$$Q = \frac{Q_3 - Q_1}{2}$$

L'intervalle interquartile

Cas 1 : Données rangées

Calcul des Quartiles (par méthode des centiles)					
	Centile	Rang	Rang Arrondi >	Quartiles	Valeurs
Q1 = Cn =>	25	31,75	32	**Q1 =**	52
Q3 = Cn =>	75	95,25	95	**Q3 =**	58
Q2 = Cn =>	50	63,5		**Q2 =**	56
Semi – Interquartile =		**3**			
	52,46	**< 50% des valeurs <**		58,46	

$$\text{L'intervalle interquartile} = Q_3 - Q_1 = 58 - 52 = 6$$

$$\text{L'intervalle semi-interquartile } Q = \frac{Q_3 - Q_1}{2} = \frac{58 - 52}{2} = 3$$

Moyenne – Q < 50% des valeurs < Moyenne + Q

Mesures de la dispersion

1.Ecart absolu moyen ou Ecart Moyen Absolu « EMA »:

Ce paramètre est la moyenne arithmétique de la valeur absolue des écarts à la moyenne.

a) Données rangées :

L'écart absolu moyen est la moyenne des distances mesurées positivement (en valeur absolue) entre les données et la moyenne.

$$EM_X = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{N}$$

Ecart Moyen Absolu

Mesures de la dispersion

Poids (kg)	65	66	67	68	68	69	70	70	71	71	71	72	73	74	74	75	75	75

N= 18 ; Moyenne = 70,77kg

$$EM_X = \frac{|65-70.77|+|66-70.77|.....+|75-70.77|}{18} = 0.143$$

L'écart absolu moyen est faible et les valeurs sont très concentrées autour de la moyenne

Ecart Moyen Absolu

Mesures de la dispersion

Données rangées : le calcul de EMA s'exprime

$$EM_X = \frac{\sum_{i=1}^{n} n_i \left| x_i - \bar{x} \right|}{N}$$

Données groupées en classes : le calcul de EMA s'exprime

$$EM_X = \frac{\sum_{i=1}^{n} n_i \left| m_i - \bar{x} \right|}{N}$$

Variance et écart-type

La variance et écart-type servent à évaluer la dispersion d'une distribution autour d'une valeur centrale, la moyenne.

Variance : La variance, notée $(\sigma_x)^2$ est la moyenne du carré des écarts à la moyenne.

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

La variance n'est pas un paramètre de dispersion absolue mais plutôt une mesure globale de la variation d'un caractère de part et d'autre de la moyenne arithmétique (quantité d'information). Pour obtenir un paramètre de dispersion absolue, on effectue la racine carrée de la variance, appelé **écart-type** et que l'on note **σX**

Mesures de la dispersion

La variance pour des données rangées ou groupées en classe devient :

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

Ou $\mathbf{n_i}$ désigne les effectifs de chaque donnée ou de chaque classe

$$(\delta_x)^2 = \frac{\sum_{i=1}^{n} n_i (m_i - \bar{x})^2}{N}$$

Mesures de la dispersion

L'écart-type est une mesure de dispersion par rapport à la moyenne qui intègre les valeurs algébriques des écarts à la moyenne et qui pourra, à ce titre être réintroduite dans des calculs algébriques ultérieurs.

Elle présente de plus l'avantage d'avoir une **signification probabiliste** que ne possède pas l'écart absolu moyen. La théorie des probabilités permet en effet d'estimer la chance qu'à une valeur d'être éloignée de la moyenne de plus d'un certain nombre d'écart-types.

b - Ecart-type : L'écart type, noté σx est la racine carré de la moyenne du carré des écarts à la moyenne, c'est à dire la racine carrée de la variance.

$$\delta_x = \sqrt{(\delta_x)^2}$$

Mesures de la dispersion

Lorsqu'une distribution est **gaussienne** (on dit aussi "**normale**") les probabilités de trouver les valeurs a une distance donnée de la moyenne sont les suivantes :

68.3 % des valeurs sont comprises entre $(\overline{x} - \delta_x)$ et $(\overline{x} + \delta_x)$

95.5 % des valeurs sont comprises entre $(\overline{x} - 2\delta_x)$ et $(\overline{x} + 2\delta_x)$

99.7 % des valeurs sont comprises entre $(\overline{x} - 3\delta_x)$ et $(\overline{x} + 3\delta_x)$

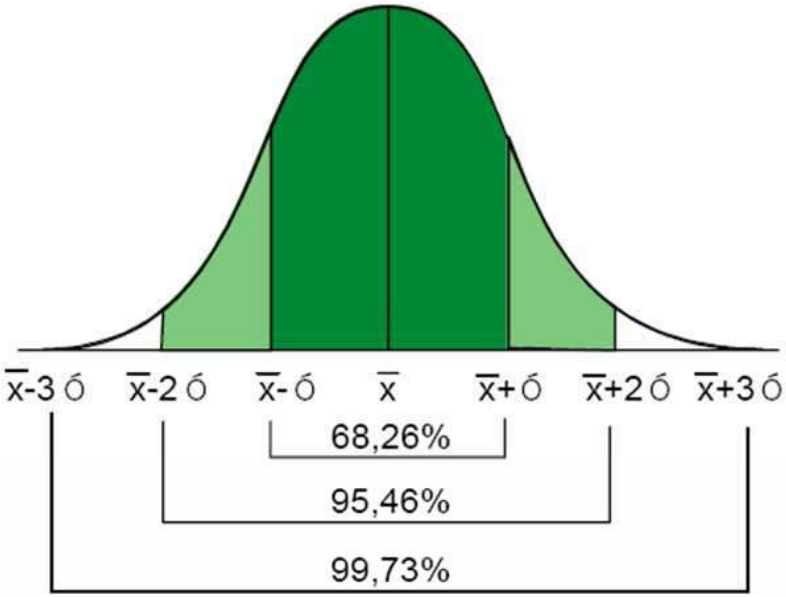


Fig *Représentation graphique d'une distribution « normale »*
(Loi de Gauss ou Loi Normale)

Les paramètres de dispersion relative

La comparaison des paramètres de dispersion absolue de deux caractères n'a de sens que si les deux caractères sont de même nature et de même ordre de grandeur. Dans le cas contraire, la comparaison n'est possible qu'en ayant recours à des mesures de **dispersion relative**, c'est à dire en effectuant le rapport entre un paramètre de dispersion absolue et la valeur centrale qui lui tient de référence.

Un paramètre de dispersion relative est une mesure de **l'écart relatif** des valeurs d'une distribution à une valeur centrale. C'est donc le rapport d'un paramètre de dispersion absolue divisé par une valeur centrale. On obtient un nombre sans dimension qui peut être exprimé en %.

Dispersion relative = Paramètre de dispersion absolue/Valeur centrale

-le coefficient interquartile relatif

$= (X_{q3}-X_{q1})/$ médiane X

-l'écart moyen relatif

= E.A.M. / $\bar{X}$

-le coefficient de variation

$= \sigma_x / \bar{X}$

Explication des paramètres de dispersion relative pour la variance et l'écart-type :

Ces deux mesures de dispersion (variance et écart-type) sont des grandeurs de même ordre de la variable étudiée : il s'agit d'une mesure de dispersion absolue or pour comparer des séries différentes, il faut éliminer l'unité de mesure afin d'obtenir une mesure de dispersion relative on utilise alors le coefficient de variation exprimé en % :

$$CV = \frac{\delta_x}{\bar{x}} * 100$$

Plus le coefficient de variation est faible, plus la dispersion est faible.

Paramètres de forme

Ces paramètres permettent de préciser la forme de la distribution expérimentale. Ils affinent la description de la distribution d'une variable et facilite la comparaison de plusieurs distributions expérimentales. Les paramètres de forme que nous aborderons sont :

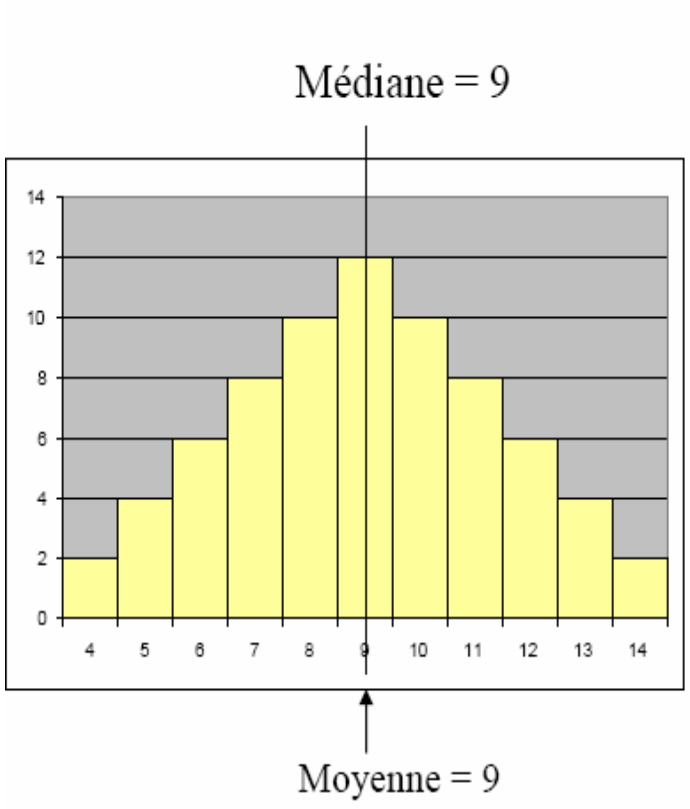
(1)le coefficient d'asymétrie il permet de nous renseigner sur la façon régulière ou non dont les observations se répartissent de part et d'autre d'une valeur centrale.

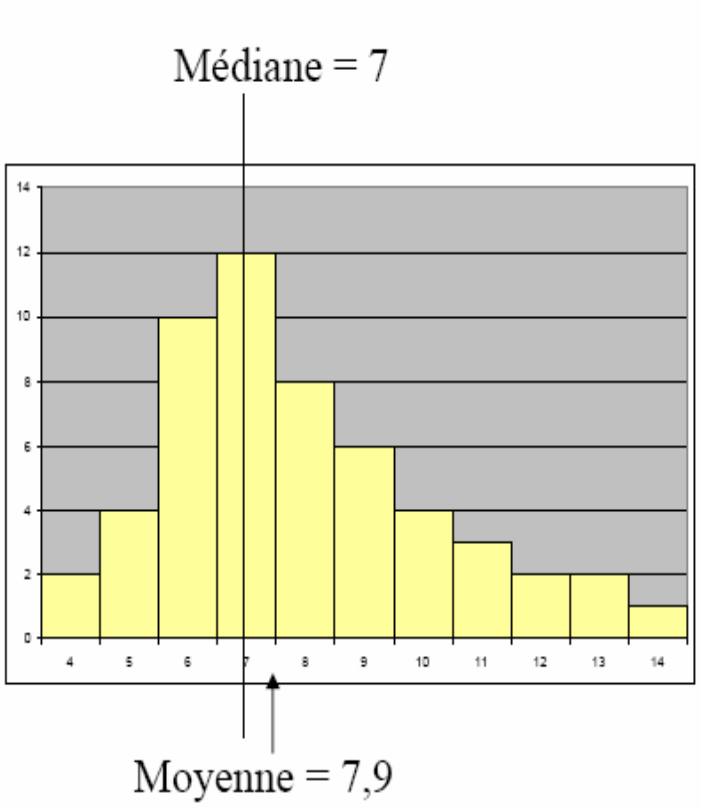
(2)le coefficient d'aplatissement dont l'objet et de faire apparaître si une faible variation de la variable entraîne ou non une forte variation des fréquences relatives.

Coefficient d'asymétrie et de dérive

Le coefficient d'asymétrie renseigne sur l'asymétrie et éventuellement la dérive par rapport à une valeur centrale choisie. La distribution d'une variable est symétrique si les observations sont également dispersées de part et d'autre d'une valeur centrale. Ainsi, dans le cas de distributions symétriques, moyenne et médiane sont confondues, sinon elles sont distinctes.

Les coefficients d'asymétrie





Coefficient d’asymétrie

Ce coefficient mesure l'asymétrie d'une distribution, il renseigne sur une asymétrie négative (dissymétrie à gauche), ou une asymétrie positive (dissymétrie à droite), c'est-à-dire il précise si la répartition "penche" d'un côté ou de l'autre. Selon la valeur centrale choisie (mode, médiane ou moyenne arithmétique), il existe différentes manières de caractériser et de mesurer une dissymétrie.

Astuce :

-Dans le cas d'une dissymétrie positive on a généralement (partie droite plus longue que la partie gauche) : **Mo** (Mode) **< Md** (Médiane) **<**(Moyenne)

-Dans le cas d'une dissymétrie négative on a généralement (partie gauche plus longue que la partie droite) : **Mo** (Mode) **> Md** (Médiane) **>**(Moyenne)

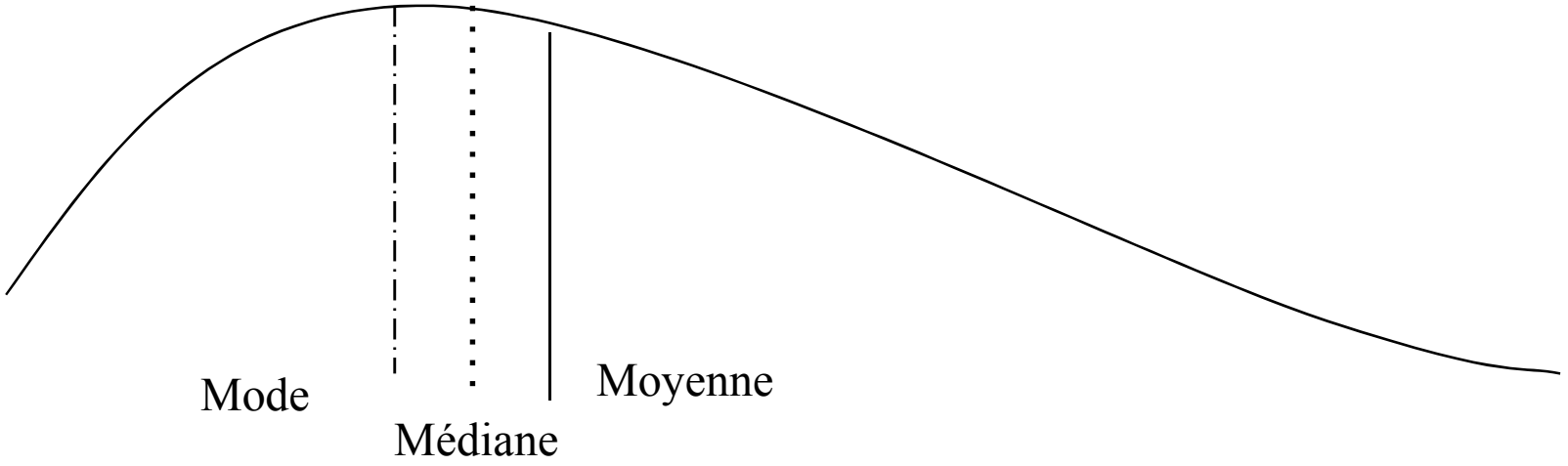


Fig. *Exemple de dissymétrie à droite (distribution étiré à droite et oblique à gauche)*

Les coefficients d'asymétrie de Yule, si la valeur centrale choisie est la médiane :

Yule propose une mesure de l'asymétrie en comparant l'étalement vers la gauche et l'étalement vers la droite, tous deux repérés par la position des quartiles (Q1, Médiane (Q) et Q3)

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$S = 0 \Leftrightarrow$ *symétrie parfaite*

$S > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite* $S <$

$S < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

Les coefficients d'asymétrie

Les coefficients d'asymétrie de Pearson, si les valeurs centrales choisies sont le mode et la moyenne. Pearson propose deux coefficients :

a)le premier coefficient d'asymétrie de Pearson analyse la position de deux valeurs centrales (le mode et la moyenne arithmétique) relativisée par la dispersion de la série :

$$P = \frac{\mu - \mathrm{mod}\, e}{\delta}$$

$p = 0 \Leftrightarrow$ *symétrie parfaite*

$p > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite* $p <$

$p < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

Remarque : ce coefficient est plutôt performant pour des distributions faiblement asymétriques

b) Le second coefficient d'asymétrie de Pearson (β_1) est plus élaboré : il s'appuie sur le calcul des moments centrés. Il s'écrit :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Où

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3$$

$$\mu_2 = m_2 - m_1^2$$

Avec

$$m_1 = \frac{\sum n_i x_i}{N}$$

$$m_2 = \frac{\sum n_i x_i^2}{N}$$

$$m_3 = \frac{\sum n_i x_i^3}{N}$$

$\beta_1 = 0 \Leftrightarrow$ *symétrie*

$\beta_1 > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite*

$\beta_1 < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

Les coefficients d'asymétrie

Les coefficients d'asymétrie de Fisher, si la valeur centrale choisie est la moyenne : Fisher propose un coefficient qui correspond à la racine carrée du coefficient β_1 de Pearson :

$$y_1 = \frac{\mu_3}{\delta^3} \quad \text{où} \quad S^3 = \sqrt{\mu_2^3}$$

$\gamma_1 = 0 \Leftrightarrow$ *symétrie*

$\gamma_1 > 0 \Leftrightarrow$ *oblique à gauche (ou étalement à droite)= dissymétrie à droite*

$\gamma_1 < 0 \Leftrightarrow$ *oblique à droite (ou étalement à gauche)= dissymétrie à gauche*

Coefficient de dérive

Le coefficient **d'asymétrie de Fisher** calculé ci-dessus correspond pour certains auteurs au **coefficient de dérive** « d » ainsi

$$d = y_1 = \frac{\mu_3}{\delta^3}$$

Les coefficients d et δ sont très sensibles aux fluctuations d'échantillonnage, il faudra disposer d'un grand nombre d'observations pour les utiliser.

Exemple

Les coefficients d'asymétrie

Les résultats d'une enquête des âges au niveau d'une entreprise nous permis d'obtenir le tableau suivant:

Classes	n_i	x_i	$n_i(x_i-\bar{x})^2$	$n_i(x_i-\bar{x})^3$	$n_i(x_i-\bar{x})^4$
50 – 60	8	55	613.0576	4904.461	
60 – 70	10	65	217.8576	2178.576	
70 – 80	16	75	22.6576	362.5216	
80 – 90	14	85	27.4576	384.4064	
90 – 100	10	95	232.2576	2322.576	
100 – 110	5	105	637.0576	3185.288	
110 – 120	2	115	1241.858	2483.715	
TOTAL	65			19.10	38.211

$S = 0.03$ $\beta_1 = 0.131$

$P = 0.3$ $y_1 = 0.336$

La distribution est donc légèrement oblique à gauche.

Coefficient d'aplatissement

Le coefficient d'aplatissement, par référence à la courbe de la loi normale, indique si la distribution de la variable est leptocurtique (pointue), mésocurtique (normale) et platycurtique (plat).

Coefficient d'aplatissement

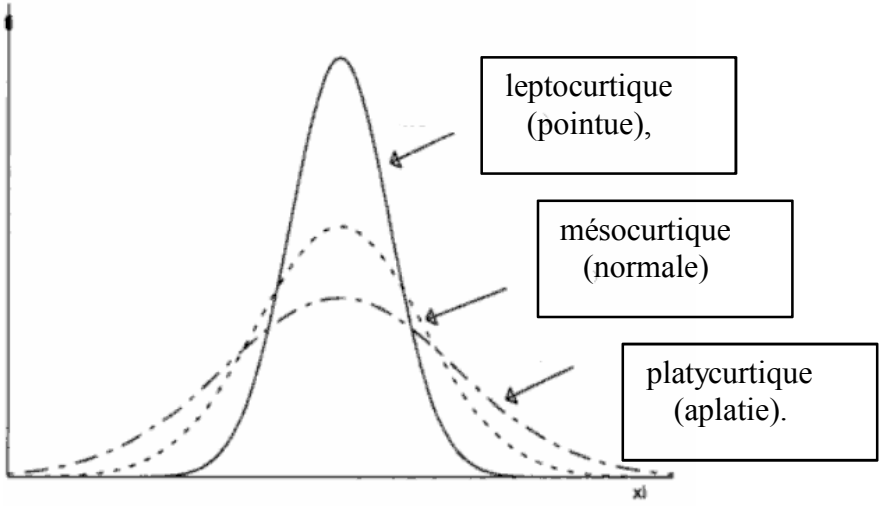


Fig: *Courbe avec coefficient d'aplatissement différent*

Coefficient d'aplatissement de Pearson

Ce coefficient est toujours supérieur ou égal à 1.

Plus ce coefficient est faible plus la répartition est aplatie (plus la courbe est platicurtique).

Plus il est grand, plus les observations sont plus regroupées autour de la moyenne.

β_2 prend la valeur 3 pour une distribution normale.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\delta_4}$$

Coefficient d'aplatissement

Coefficient d'aplatissement de Fisher

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\delta_4} - 3$$

$\gamma_2 = 0 \Leftrightarrow$ *distribution normale, l'aplatissement est le même que celui de la loi de Gauss réduite*

$\gamma_1 < 0 \Leftrightarrow$ *la distribution est plus aplatie (*platicurtique*)*

$\gamma_1 > 0 \Leftrightarrow$ *la distribution est moins aplatie (*leptocurtique*)*

Exemple

Coefficient d'aplatissement

x_i	f_i	$f_i x_i$	$f_i x_i^2$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$	$f_i(x_i - \bar{x})^4$
0	0.216	0	0	$(-1.2)^2$	0.311	-0.373	0.448
1	0.432	0.432	0.432	$(-0.2)^2$	0.017	-0.0035	0.00069
2	0.288	0.576	1.152	$(+0.8)^2$	0.184	+0.147	0.11796
3	0.064	0.192	0.576	$(+1.8)^2$	0.207	+0.373	0.6718
$\sum$	1	1.2	2.16		0.72	0.144	1.238
		m_1	m_2		μ_2	μ_3	μ_4

$\beta_1 = 0.05$

$\gamma_1 = 0.24$

La distribution est oblique à gauche

$\beta_2 = 2.39 (< 3)$

$\gamma_2 = -0.61$

La distribution est platicurtique