Chapitre 1 : DISTRIBUTION ET REPARTITION

Une observation fait en général apparaître un ou plusieurs éléments d'un ensemble E : c'est le cas par exemple des numéros gagnants du loto (pris parmi 1, 2 … 49) comme des symptômes montrés par un patient (pris parmi irritation, nausée, migraine, fièvre …).

1.1 POPULATION

On s'intéresse ici à un ensemble E *fini* dit POPULATION, dont les éléments i sont dits INDIVIDUS, ces termes restant employés même si on n'observe pas des êtres humains.

Exemple : On parle de population pour des malades dont on surveille l'évolution, pour des rats de laboratoire dont on examine les réactions à un traitement, pour des gélules d'un certain médicament dont on contrôle la conservation (et même pour les étoiles d'une galaxie). En revanche, les différents prélèvements que l'on peut faire dans un tube de plasma (pour doser le cholestérol) ne constituent pas une population parce que l'on ne peut pas *a priori* les discerner (mais les *molécules* de cholestérol constituent une population).

§

L'ensemble E étant fini, on peut en observer tous les éléments.

Rappel : La partie *vide*, qui ne contient aucun élément, se note $\varnothing$ et la partie *pleine* correspond à l'ensemble E tout entier. Pour 2 parties A et B l'*union* ($A \cup B$) est formée par les éléments aussi bien de A que de B et l'*intersection* ($A \cap B$) est formée par les éléments communs à A et B. On dit enfin que A et B sont *disjoints* si leur intersection est vide

$$A \cap B = \varnothing$$

et que la partie A est *incluse* dans B si tout élément de A est aussi un élément de B, ce que l'on écrit

$$A \subseteq B$$

et ce qui permet de définir le *complément* de A *dans* B ($[_B A$) par les éléments de B qui ne sont pas dans A.

§

Un CARACTERE pouvant être associé à un individu i, comme par exemple le sexe ou l'âge pour un être humain, une partie peut être définie par un *caractère* commun à différents éléments de E, comme par exemple l'ensemble des femmes est défini par le caractère féminin dans une population d'individus. La partie *vide* correspond à un caractère IMPOSSIBLE à observer et une partie ne contenant qu'un seul résultat (dite *singleton*) est appelée un caractère ELEMENTAIRE ou FONDAMENTAL.

Exemple : Si l'ensemble E précédent peut se présenter par d'abord 10 femmes puis 15 hommes, les caractères A « femme » et B « homme » s'écrivent comme les parties

$$A = \{i_1, i_2 \dots i_{10}\} \text{ et } B = \{i_{11}, i_{12} \dots i_{25}\}$$

Le caractère « Numéro d'individu 13 » est *élémentaire* et le caractère C « Homme avec cancer de l'utérus » est *impossible*

$$C = \varnothing$$

§

Un caractère, étant une partie d'ensemble, peut se faire appliquer les notions de *complément* et d'*union*.

Le mot *complément* sans mention « dans … » sous-entend « dans E » : le *complément* de A, noté $\overline{A}$ ou A^c, qui a ses éléments hors de A, définit le CONTRAIRE de A.

Exemple : Le caractère A « femme » a pour complément le caractère

$B = \overline{A}$ = « homme »

et le complément de B est

$\overline{B}$ = « femme » = A

§

D'une façon générale, le complément du complément de A est A lui-même

$\overline{A}^c = A$

et un caractère et son complément forment une ALTERNATIVE (un individu i est soit dans A soit dans $\overline{A}$) ou une DISJONCTION (dite parfois *exclusive* pour rappeler qu'un cas exclut l'autre).

L'*union* ($A \cup B$), formée par les éléments de A et de B, correspond à l'un au moins des caractères A et B et définit ainsi la DISJONCTION INCLUSIVE des caractères A et B.

Exemple : L'union des caractères A « Sexe féminin » et B « Age d'au moins 55 ans » désigne les femmes de n'importe quel âge, ainsi que les hommes d'au moins 55 ans.

§

Si on considère A et son *contraire* B, on est sûr d'observer A ou B, mais jamais simultanément : on est devant une *alternative* (« de deux choses l'une ») et l'union de A et B est l'ensemble E.

Les deux opérations - complément et union - permettant de structurer un ensemble de caractères définis sur E, on doit considérer l'*intersection* de deux états A et B ($A \cap B$), qui est complément d'une union particulière

$(\overline{A} \cup \overline{B})^c = \overline{A}^c \cap \overline{B}^c = A \cap B$

et qui correspond à des caractères SIMULTANES, ou encore à la CONJONCTION des caractères A et B.

Exemple : L'intersection des caractères A « Sexe féminin » et B « Age d'au moins 55 ans » désigne les femmes d'au moins 55 ans.

§

Deux parties A et B *disjointes* n'ont en commun aucun cas : les caractères correspondants ne peuvent alors coexister (ils sont en *disjonction*) et sont dits pour cette raison INCOMPATIBLES.

Exemple : Les caractères « femme » et « homme » sont incompatibles.

§

Enfin, si la partie A est *incluse* dans B ($A \subseteq B$), tout individu ayant le caractère A possède aussi le caractère B, donc le caractère A IMPLIQUE le caractère B.

1.2 VARIABLE

On s'intéresse, dans une population, à différents caractères des individus : dans une étude portant sur des animaux d'expérimentation on observera le sexe, l'âge, le poids ...

Un caractère variant d'un individu à l'autre on considère ses différentes possibilités comme les valeurs d'une VARIABLE, qui est dite de nature LEXICALE (ou NOMINALE) s'il s'agit de texte et NUMERIQUE s'il s'agit de nombres. La description statistique d'un groupe d'individus requiert alors d'étudier la DISTRIBUTION des valeurs de la variable (c'est-à-dire les fréquences associées à des CLASSES de valeurs, une classe pouvant ne contenir qu'une valeur).

- Types de variable

Une variable est dite QUALITATIVE (ou de type QUALITATIF) si elle correspond à un *état*, comme le sexe, l'état civil, la couleur des cheveux ... Elle peut être à 2 modalités (binaire ou dichotomique) : sexe (masculin ou féminin), survie (vivant ou mort) ou à plus de 2 modalités : l'état civil (célibataire, concubin, marié, séparé, divorcé, veuf).

Une variable est dite QUANTITATIVE si elle représente une *quantité* ou un *montant* ; elle est
- *discrète* ou *discontinue* si elle ne peut prendre que des valeurs entières (nombre d'enfants d'une famille ...) ou, plus généralement, si elle ne peut passer que par « saut » d'une valeur observable à la plus proche (niveau d'énergie d'un électron) ;
- *continue* si elle peut théoriquement prendre toute valeur numérique (entière ou non) dans un intervalle de nombres réels (taille, glycémie ...).

Une variable est dite ORDINALE s'il existe un ordre *objectif* sur ses valeurs, donc toute variable quantitative est ordinale. Une variable qualitative ordinale dont les modalités sont *numériques* est dite SEMI-QUANTITATIVE, et une variable qualitative non ordinale est dite QUALITATIVE PURE.

Exemples : - La variable « sexe » est qualitative à 2 modalités (masculin ou féminin). On note [Sexe = féminin] l'ensemble des individus pour qui la variable « sexe » a la valeur « féminin ».

- La variable « douleur » à 3 modalités (absente, modérée ou forte) est qualitative ordinale, et serait dite semi-quantitative si on posait 0 = absente, 1 = modérée et 2 = forte.

- La variable « nombre de rémissions d'un cancer » est quantitative discrète, avec les valeurs 0, 1, 2 ...

- La variable « âge » est quantitative continue : elle peut prendre toute valeur réelle positive.

§

Une variable peut donc être soit *qualitative pure* ou *ordinale* (éventuellement *semi-quantitative*) soit *quantitative discrète* ou *continue*.

Une variable est dite CATEGORISEE si ses valeurs sont mises en *classes*, ce qui est toujours possible.

Exemples : On peut mettre en une même classe « concubin » et « marié » pour l'état civil, « absente » et « modérée » pour la douleur, les valeurs au moins égales à 2 pour le nombre de rémissions, et celles comprises entre 30 et 45 ans pour l'âge : dans ce dernier cas on note $[30 < \text{Age} \leq 45]$ l'ensemble des individus concernés.

1.3 DESCRIPTION PAR VALEURS ET FREQUENCES

Pour présenter les valeurs d'une variable, observables chez N individus, on commence par les ordonner, et éventuellement les mettre en classes, pour établir un TABLEAU DE FREQUENCES (généralement complété par un *graphique*) tenant compte du type qualitatif ou quantitatif de la variable.

- Effectif simple ou cumulé

Soit X la variable étudiée : on parle d'EFFECTIF SIMPLE pour les individus ayant une certaine *valeur* de X, ou ayant leur valeur de X dans une certaine *classe* (cas particulier de X *discrète* ou *continue* si on s'intéresse à une classe qui est un intervalle [b, a] et à l'effectif des individus tels que $b < X \leq a$).

- Fréquence absolue ou relative

Soit la modalité x_j (ou la j-ème classe de valeurs) de X.

L'*effectif* correspondant n_j est dit FREQUENCE ABSOLUE par opposition à la FREQUENCE RELATIVE notée Fréq(X = x_j) qui est la *proportion*

$$\varphi_j = \frac{n_j}{N}$$ (exprimée souvent en pourcentage en multipliant par 100)

rapportée à l'effectif total N qui est égal à la somme des effectifs simples

$$N = n_1 + n_2 \ \ldots + n_j \ldots + n_k = \sum_{j=1}^{k} n_j$$

Si on regroupe des valeurs ou des classes *consécutives* le total obtenu en additionnant les effectifs est dit EFFECTIF CUMULE.

On associe à l'*effectif cumulé*

$N_j = n_1 + n_2 \ \ldots + n_j$

la *fréquence cumulée relative* définie par la proportion calculée sur l'effectif total N

$$\Phi_j = \frac{N_j}{N}$$ (exprimée souvent en pourcentage en multipliant par 100)

On obtient alors un tableau de fréquences en présentant les fréquences associées à chaque valeur (ou chaque classe de valeurs) de la variable.

Exemple : Pour la variable X « résultat du baccalauréat » on observe k=3 modalités sur 9152 individus :

	x_1	x_2	x_3
Résultat	Admis à l'écrit	Admis à l'oral	Refusé
	n_1	n_2	n_3
Candidats	6864	1830	458

Pour les admis à l'écrit la fréquence absolue est l'effectif 6864 et la fréquence relative

6864/9152 = 0,750 = 75,0%

tandis que la fréquence des refusés vaut en relatif

458/9152 = 0,050 = 5,0%.

L'effectif cumulé des admis est

6864 + 1830 = 8694

et la fréquence cumulée relative vaut 8694/9152 = 95,0%.

Remarques : - La somme des fréquences relatives *simples* est égale à l'unité :

$$\sum_{j=1}^{k} \varphi_j = \frac{\sum_{j=1}^{k} n_j}{N} = 1 \quad \text{ou} \quad \Sigma \text{ Fréq}(X = x_j) = 1$$

- Pour une variable X *binaire* ne prenant que les k=2 valeurs $x_1 = 0$ et $x_2 = 1$ on convient de privilégier la valeur 1 en posant la fréquence simple $\psi = \varphi_1$, ce qui donne l'autre fréquence simple $\varphi_0 = 1 - \psi$.

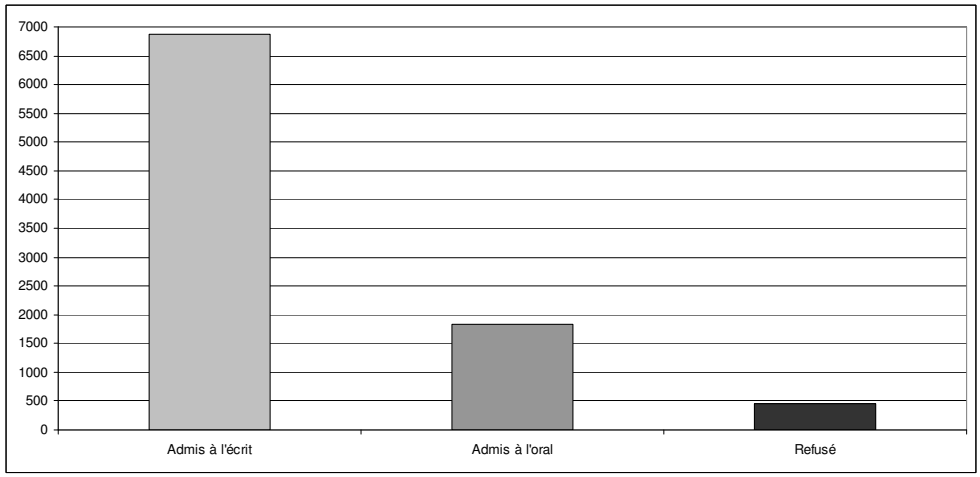
1.4 DISTRIBUTION (VARIABLE QUALITATIVE)

On appelle DISTRIBUTION (STATISTIQUE) la donnée de chaque valeur (ou de chaque classe de valeurs) de X et de sa fréquence : on présente graphiquement une *distribution* avec un diagramme particulier nommé HISTOGRAMME.

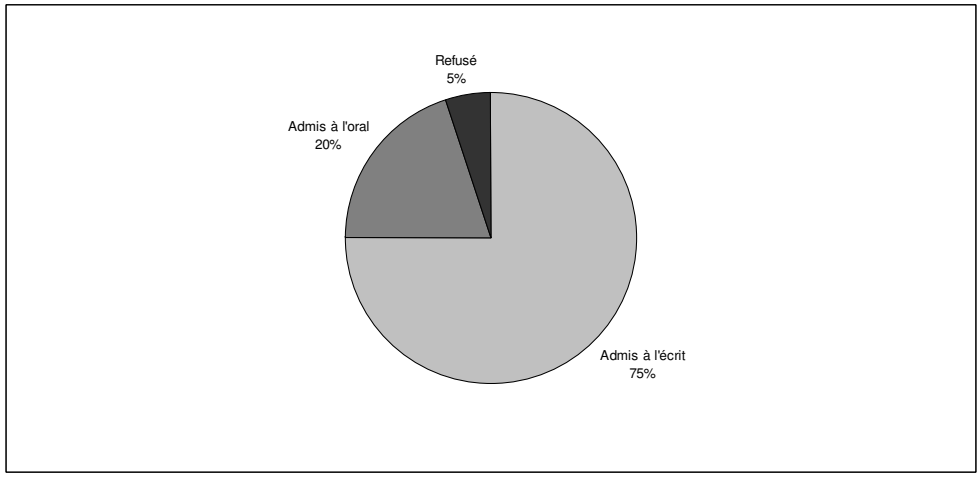
[variable qualitative] - Histogramme en barres

L'*histogramme,* pour une variable qualitative, prend principalement la forme du *diagramme en barres* : on trace des rectangles non contigus (les « barres ») dont chacun a une hauteur proportionnelle à la fréquence à représenter.

Exemple : Résultats d'une session de baccalauréat en France.



On présente parfois les barres « réduites à des bâtons », qu'il faut alors voir comme des barres minces, ou « couchées », mais le diagramme devient moins parlant. On utilise aussi le diagramme circulaire : on représente sur un disque chaque modalité (ou chaque classe) par un secteur circulaire d'angle proportionnel (donc d'aire proportionnelle) à sa fréquence.



1.5 REPARTITION (VARIABLE ORDINALE)

Si la variable considérée est ordinale on peut en *ordonner* les k modalités

$x_1 < x_2 \ldots < x_j \ldots < x_k$

et calculer pour chaque x_j la *fréquence cumulée absolue* N_j (l'*effectif cumulé*) en comptant les individus depuis la première modalité jusqu'à celle d'indice j incluse :

$$N_j = n_1 + n_2 + n_3 + ... + n_j = \sum_{i=1}^{j} n_i$$

puis la *fréquence cumulée relative* Φ_i qui est somme de fréquences simples

$$\Phi_j = \frac{n_1 + n_2 + ... + n_j}{N} = \sum_{i=1}^{j} \frac{n_i}{N} = \sum_{i=1}^{j} \varphi_i = \varphi_1 + \varphi_2 \ldots + \varphi_j$$

et vérifie pour la k-ème modalité

$$\Phi_k = \sum_{i=1}^{k} \varphi_i = 1$$

La *dernière* fréquence cumulée vaut donc 1 (ou 100 %), de même que le dernier effectif cumulé est égal au nombre total N d'individus de la population étudiée.

La différence entre deux fréquences cumulées *consécutives* est une fréquence simple :

$\Phi_{j+1} - \Phi_j = \varphi_1 + \varphi_2 \ldots + \varphi_j + \varphi_{j+1} - (\varphi_1 + \varphi_2 \ldots + \varphi_j) = \varphi_{j+1}$

On appelle FONCTION DE REPARTITION la fonction associant sa fréquence *cumulée* à chaque valeur x_j (ou à chaque borne supérieure x_j de classe de valeurs) de X :

$$\text{Fréq}(X \le x_j) = \Sigma\, \text{Fréq}(X = x_i) \quad \text{pour} \quad i = 1, 2 \ldots j$$

[variable ordinale] - Diagramme cumulatif en barres

Exemple : On peut considérer pour la variable ordinale « habillage » du score AGGIR (évaluant la dépendance d'une personne) la *répartition* obtenue à partir des effectifs cumulés correspondant à « fait seul », « fait accompagné » (dépendance nulle ou modérée) ou « non fait » (dépendance nulle, modérée ou totale) :

Habillage	Fait seul	Fait accompagné	Non fait	TOTAL
Nombre de patients	n_1	n_2	n_3	N
dans cet état	40	35	25	100
au pis dans cet état	40	40+35 = 75	75+25 = 100	

La variable étudiée étant *qualitative*, on représente sa répartition avec un *diagramme en barres* de hauteurs croissantes (40%, 75% et 100%).

§

Les effectifs n_j ou les fréquences simples φ_j pouvant s'obtenir par *différence* entre deux valeurs de la fonction de répartition, l'histogramme est parfois appelé diagramme « différentiel ».

- Fonction de répartition d'une variable quantitative

La *fonction de répartition* d'une variable quantitative X

$F(a) = \text{Fréq}(X \le a)$

croît de

$F(-\infty) = 0$ (parce qu'il n'existe pas de valeur inférieure à $-\infty$)

à

$F(+\infty) = 1$ (parce que toutes les valeurs sont inférieures à $+\infty$)

et permet de calculer toute fréquence associée à un intervalle sous la forme

Fréq($b < X \leq a$) = F(a) – F(b)

(parce qu'on obtient toutes les valeurs comprises entre b et a en prenant toutes les valeurs ne dépassant pas a puis en retirant toutes celles ne dépassant pas b).

On présente graphiquement la *fonction de répartition* en exprimant les effectifs cumulés N_j, ou les fréquences cumulées Φ_j, en fonction des valeurs observable x_j : on trace une LIGNE *discontinue* (si la variable est discrète) ou *continue* (si la variable est continue) sur un diagramme parfois appelé « intégral ».

1.6 DISTRIBUTION (VARIABLE QUANTITATIVE)

L'*histogramme,* pour une variable quantitative, repose sur la notion de DENSITE.

- Densité de fréquence

On appelle DENSITE DE FREQUENCE de la variable X une fonction f(a) définie

. si X est *discrète* par la fréquence simple :

$f(a) = \text{Fréquence}(X = a)$

. si X est *continue* à partir de la fonction F de répartition :

$$f(a) = \frac{\text{Fréquence}(b < X \leq a)}{a - b} = \frac{F(a) - F(b)}{a - b} \text{ pour b voisin de a}$$

et qui est donc dans chaque cas positive ou nulle :

$f(a) \geq 0$

La fonction f qui associe ainsi à *toute* valeur réelle a une densité (de fréquence) f(a) est dite FONCTION DE DISTRIBUTION : elle n'a de sens que si la variable est *quantitative*, et est représentée graphiquement par un histogramme en traçant des *bâtons* pour les valeurs isolées, et des *rectangles* pour les valeurs en intervalle.

1.7 PRESENTATION D'UNE VARIABLE DISCRETE

Une variable discrète ne prenant que des valeurs isolées (il est impossible de passer continûment de l'une à l'autre), la densité en une valeur est la fréquence de cette valeur et l'histogramme est généralement un diagramme en bâtons.

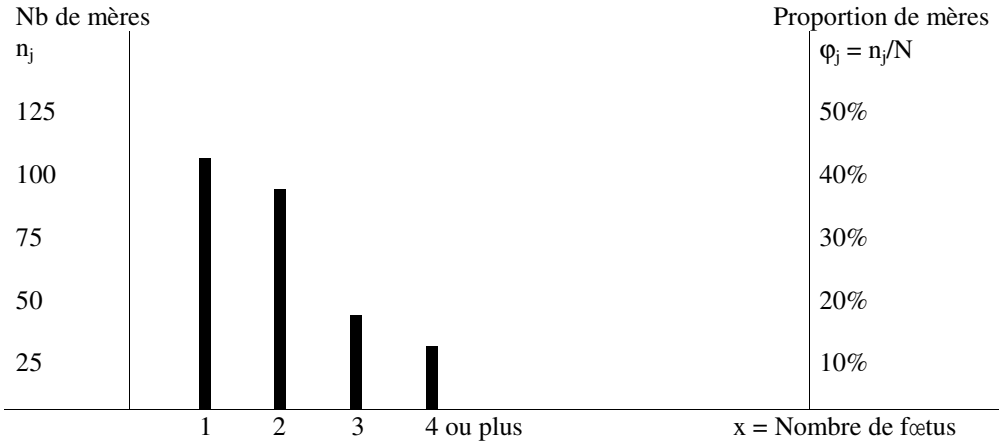
Exemple : Distribution du nombre X de fœtus par mère (« gémellité ») chez 250 mères ayant recouru à la procréation médicalement assistée (PMA).

Nb de fœtus (x_j)	Nb de mères (n_j)	Fréquence φ_j (%)	Effectif cumulé (N_j)	Fréquence cumulée (Φ_j)
1	103	41,2	103	41,2
2	89	35,6	103 + 89 = 192	76,8
3	37	14,8	192 + 37 = 229	91,6
4 ou plus	21	8,4	229 + 21 = 250	100,0
TOTAL	250	100,0		

[variable discrète] - Histogramme en bâtons

Si la variable a peu de valeurs, l'histogramme est représenté par des *verticales distinctes* (les « bâtons ») en plaçant sur l'axe horizontal les valeurs de la variable étudiée et sur l'axe vertical les effectifs n_j , ou mieux les fréquences relatives $\varphi_j = n_j/N$.

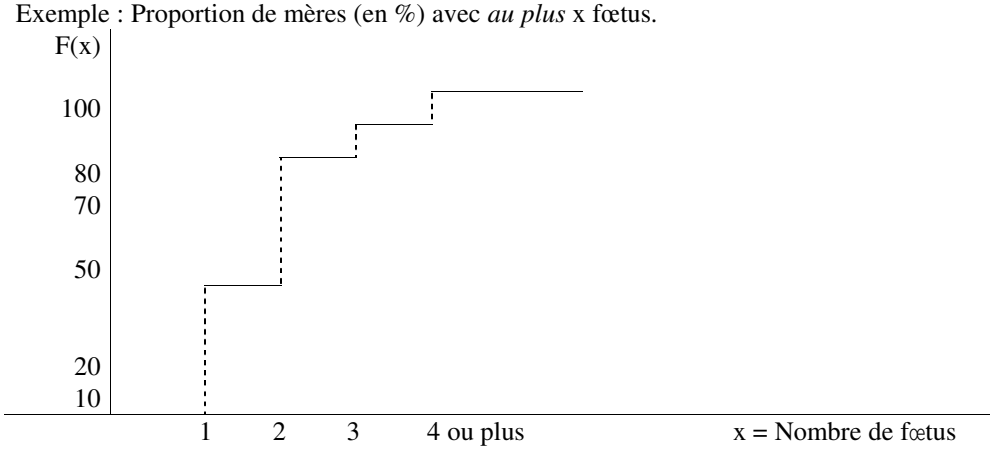
Exemple : Nombre de fœtus par mère ayant recouru à la PMA.



[variable discrète] – Diagramme cumulatif en escalier

La variable étant discrète un trait joignant deux points de la fonction de répartition est horizontal ou vertical : en effet, la fonction est constante entre deux valeurs possibles consécutives mais fait à chaque valeur possible x_j un saut égal à la fréquence simple correspondante φ_j, de sorte que l'on obtient une ligne discontinue en escalier.

Exemple : Proportion de mères (en %) avec *au plus* x fœtus.



1.8 PRESENTATION D'UNE VARIABLE CONTINUE

Pour présenter une variable *continue* il est nécessaire de la *catégoriser*, c'est-à-dire d'en mettre les valeurs en *classes* qui doivent être des intervalles *contigus* : pour cela on divise l'intervalle entre les valeurs extrêmes (limites inférieure et supérieure de la distribution) en fixant des LIMITES (ou des BORNES), en précisant si on inclut la borne inférieure (intervalle *ouvert* à droite) ou la borne supérieure (intervalle *fermé* à droite). On peut rendre les classes extrêmes plus larges pour éviter des effectifs trop faibles.

- Centre et largeur de classe

On appelle CENTRE de la classe [a , b] la moyenne arithmétique de ses limites

$$c = \frac{a+b}{2}$$

et LARGEUR (ou AMPLITUDE) de la classe la différence de ses limites

$$d = b - a$$