Statistiques Descriptives

A. DAOUI

7 octobre 2013

Présentation des données : Distribution des fréquences

cas d'un caractère qalitatif

.

Pour l'élaboration du tableau, il faut distinguer les diverses modalités du caractère étudié. En pointant une à une les données, on détermine le nombre d'individus associés à chaque modalité ; ce nombre est appelé "fréquence" ou "effectif". Si le caractère présente k modalités ditinctes, leur fréquence respective sera notée n_1, n_2,..., n_k.

La relation $n_1 + n_2 + ... + n_k = n$, où n est le nombre total des individus concernés doit toujours être satisfaite.

Selon le besoin, on peut aussi donner les "fréquences relatives" $f_i = \frac{n_i}{n}$, ou encore les fréquences en pourcentage en multipliant les rapprts obtenus par 100.

Exemple

Les données suivantes proviennent d'une étude visant à déterminer le degré de satisfaction d'un groupe de personnes concernant leurs dernières vacances :

Très Elevé	Elevé	Bas	Elevé	Elevé	Modére	Elev
Elevé	Modére	Très Elevé	Elevé	Bas	Elevé	Elev
Très Elevé	Elevé	Elevé	Modére	Très Elevé	Elevé	Mod
Très Elevé	Bas	Très Elevé	Elevé	Elevé	Très Elevé	Très

Tableau-synthèse

degré de satisfaction	Pointage	Nbre de familles n_i	f
Bas	+++	3	
Modére	++++	4	
Elevé	++++++++++++++++++	15	
Très Elevé	++++++++++	10	
Total		$n = 32$	

Cas d'un caractère quantitatif discret

Un caractère quantitatif est dit discret si l'ensemble des valeurs qu'il peut prendre est fini ou dénombrable ; le plus souvent ces valeurs sont entières.

Exemple

La série suivante représente le nombre d'enfants par famille pris sur un échantillon de foyers d'une certaine région marocaine :

Nbre d'enfants	Effectif des familles n_i	fréquence relative $\frac{n_i}{n}$
1	5	$0,1$
2	15	$0,3$
3	17	$0,34$
4	8	$0,16$
5	3	$0,06$
6	2	$0,04$
Total	50	1

Cas d'un caractère quantitatif continu

Un caractère quantitatif est dit continu s'il peut théoriquement prendre n'importe quelle valeur dans un intervalle donné ou une réunion d'intervalles de nombres réels.

Exemple

La taille d'un individu, le poids d'un nouveau né.

Dans ce cas, la construction d'une distribution de fréquence devient un peu plus complexe. Le travail consiste essentiellement à constituer un certain nombre de classes, chacune représentant un intervalle précis de nombres. Dans l'élaboration de la distribution, on tiendra compte autant que possible des conditions suivantes :

- Le nombre de classes ne doit être ni petit ni grand (entre 5 et 15).
- Chaque valeur observée du caractère doit appartenir à une et une seule classe.
- Les classes doivent être contigües et d'égale largeur.

On commence par déterminer l'**étendue** de la série, $e =$ *plus grande valeur - plus petite valeur.*

Puis on détermine le nombre k de classes et leur amplitude commune $a = \frac{e}{k}$.

Notons que le centre c_i de la $i^{ème}$ classe est

$$c_i = \frac{\text{borne sup.} + \text{borne inf.}}{2}.$$

Exemple

La série suivante représente le poids réel, en grammes, d'un échantillon de 34 dorades d'élevage.

550	271	725	190	360	509	735
453	516	610	700	490	360	489
460	628	414	450	450	412	405
242	579	390	460	535	410	373
537	510	560	430	453	709	

On commence par calculer l'étendue : $e =$ *plus grande valeur - plus petite valeur* $= 735 - 190 = 545g$.
Si l'on veut construire une distribution de fréquences de 5 classes, on obtient :
$a = \frac{e}{k} = \frac{545}{5} = 109g$
et l'amplitude appropriée est $a = 110g$

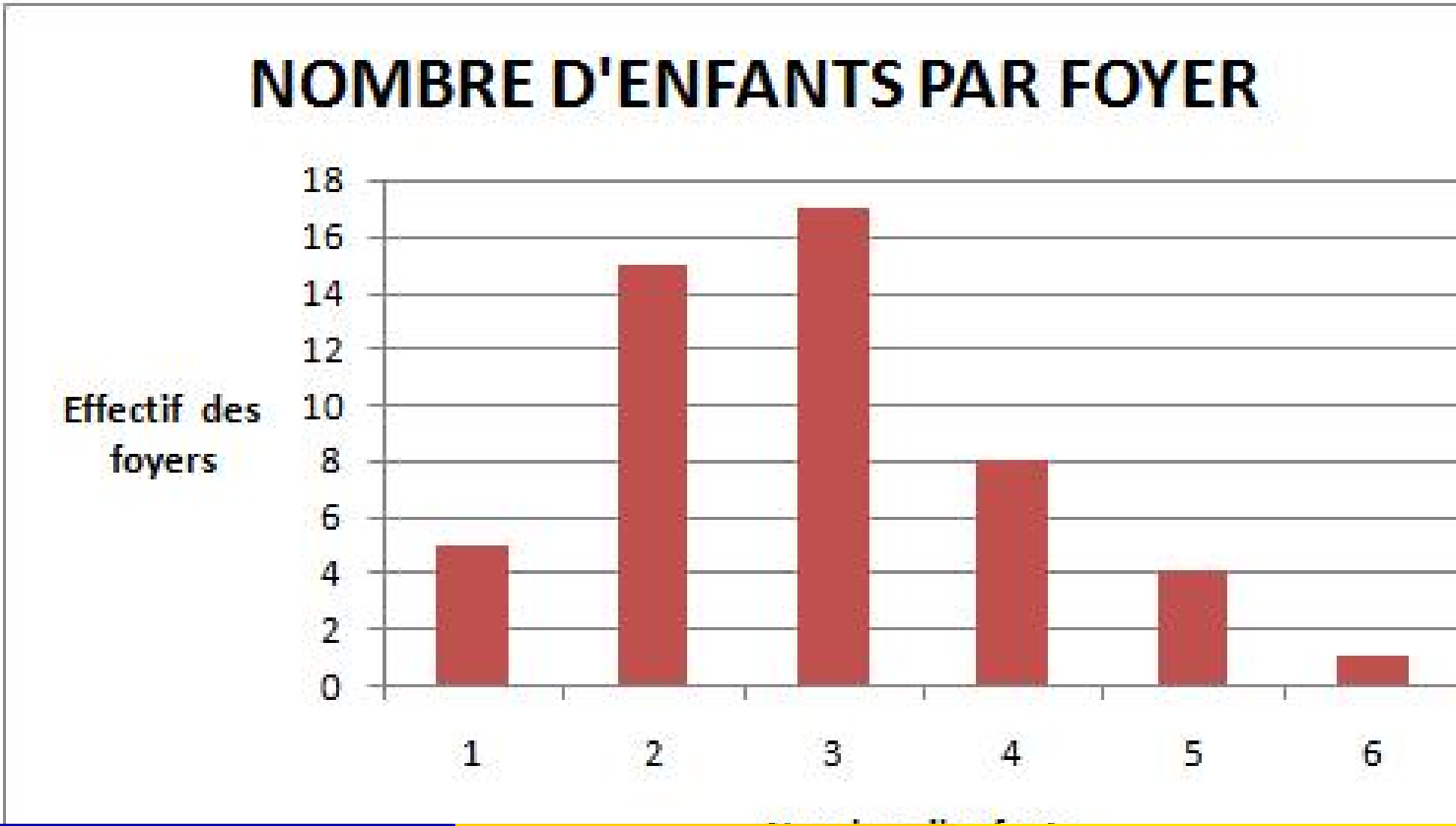
Tableau-synthèse

Poids : P (en g)	Fréquence n_i	Fréquence cumulée N_i	Centre de la class
$[190; 300[$	3	3	245
$[300; 410[$	5	8	355
$[410; 520[$	15	23	465
$[520; 630[$	7	30	575
$[630; 740[$	4	34	685
TOTAL	34		

Il est très adéquat dans la cas d'un caractère *qualitatif* ou *quantitatif discret.* Ce diagramme est constitué par la juxtaposition de bandes verticales ; la hauteur est proportionnelle à la fréquence de la modalité ou de la valeur du caractère observé.

Exemple

Cas du nombre d'enfants par famille

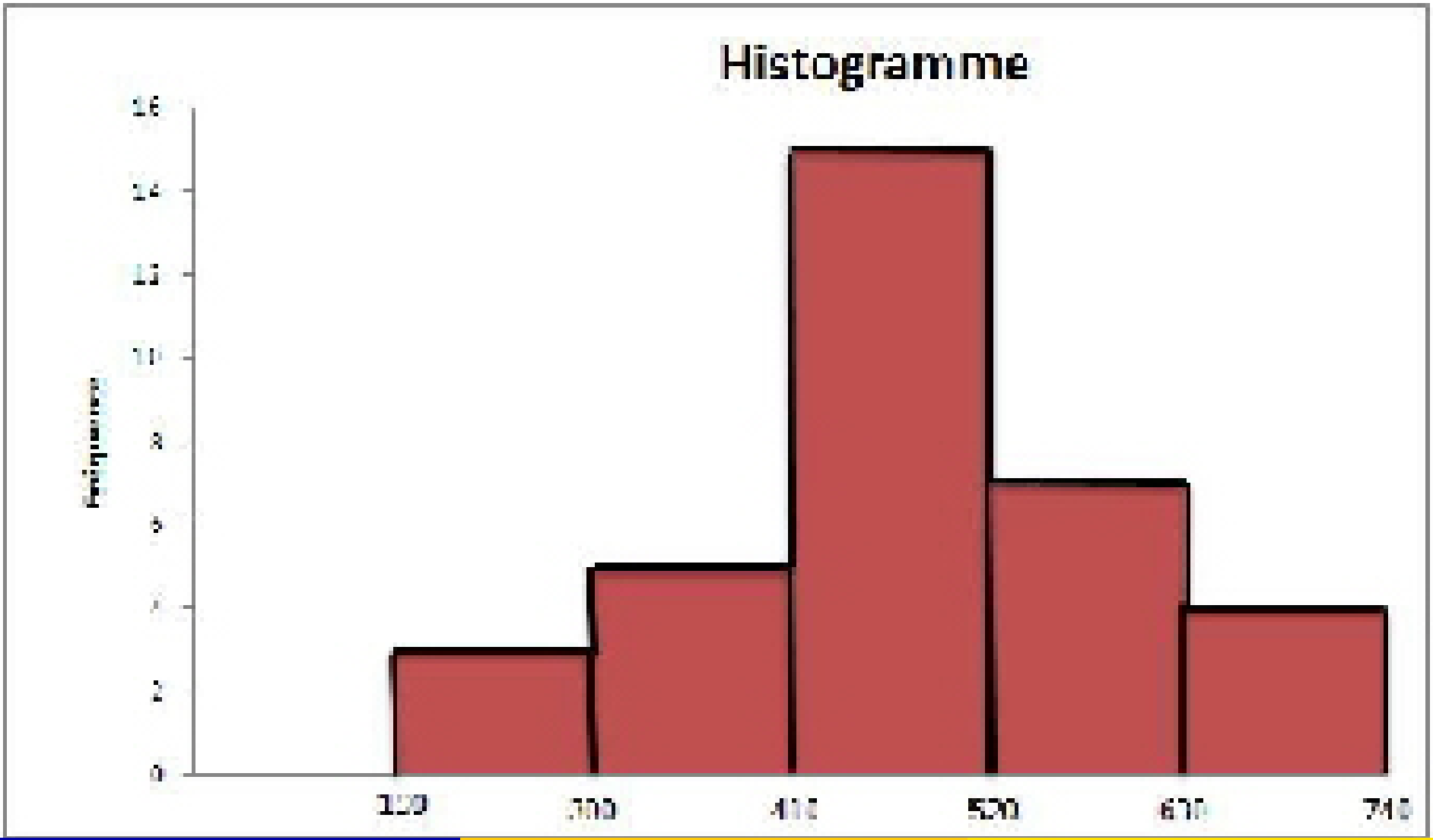


Histogramme

L'histogramme convient particulièrement pour représenter un caractère quantitatif continu ;il est constitué par la juxtaposition de bandes rectangulaires verticales et adjacentes ; la hauteur de chaque bande est proportionnelle à la fréquence correspondant à la modalité qu'elle représente.

Exemple

Cas du poids des dorades d'élevage :



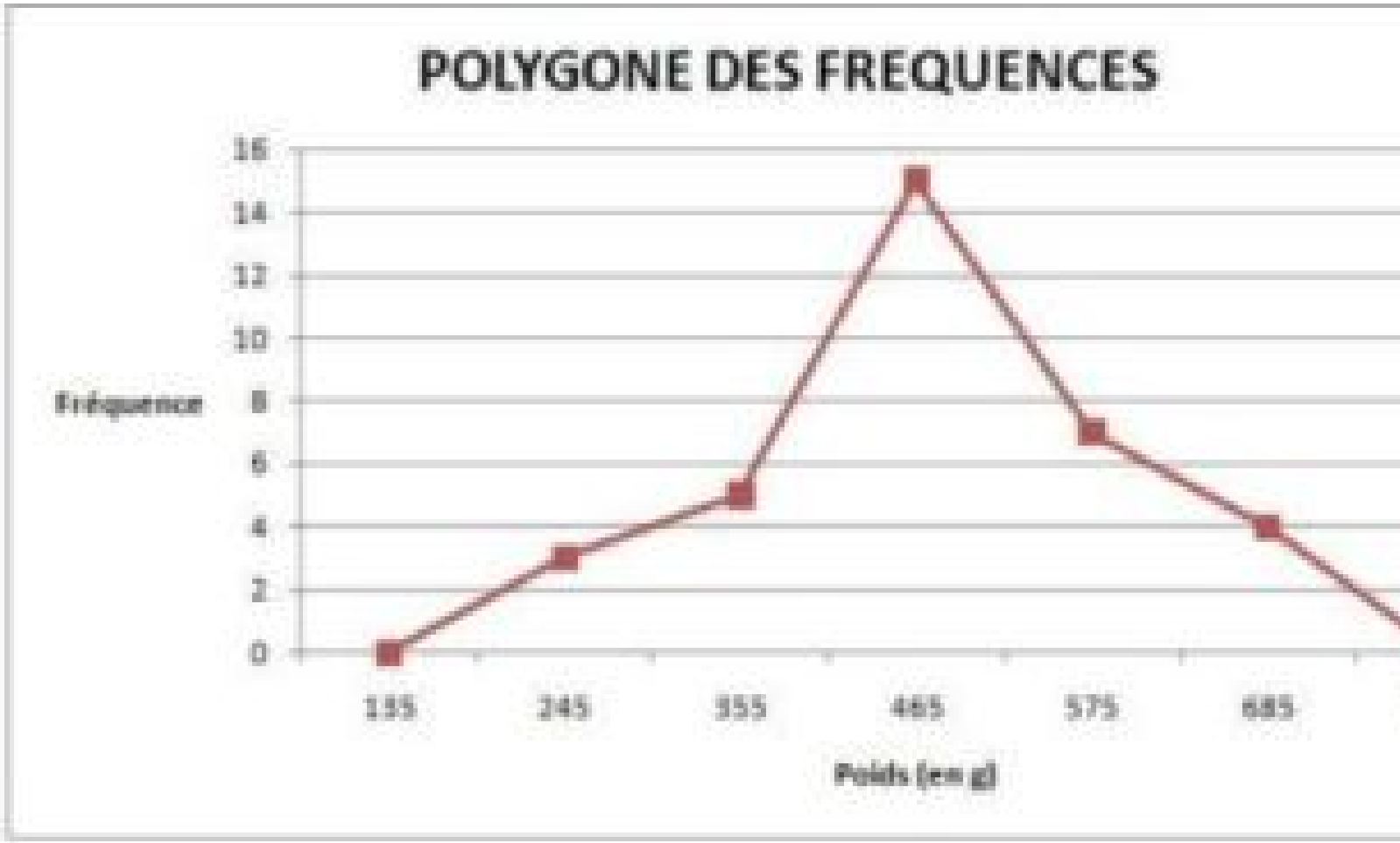
Polygône des fréquences

Pour construire ce graphique, il suffit de placer sur l'axe des abscisses les différents centres c_i des classes, et sur l'axe des ordonnées les fréquences n_i ou f_i. Par la suite on place les points $(c_i; n_i)$ et on les relie par des segments de droites.

Exemple

Cas du poids des dorades d'élevage :

frequence

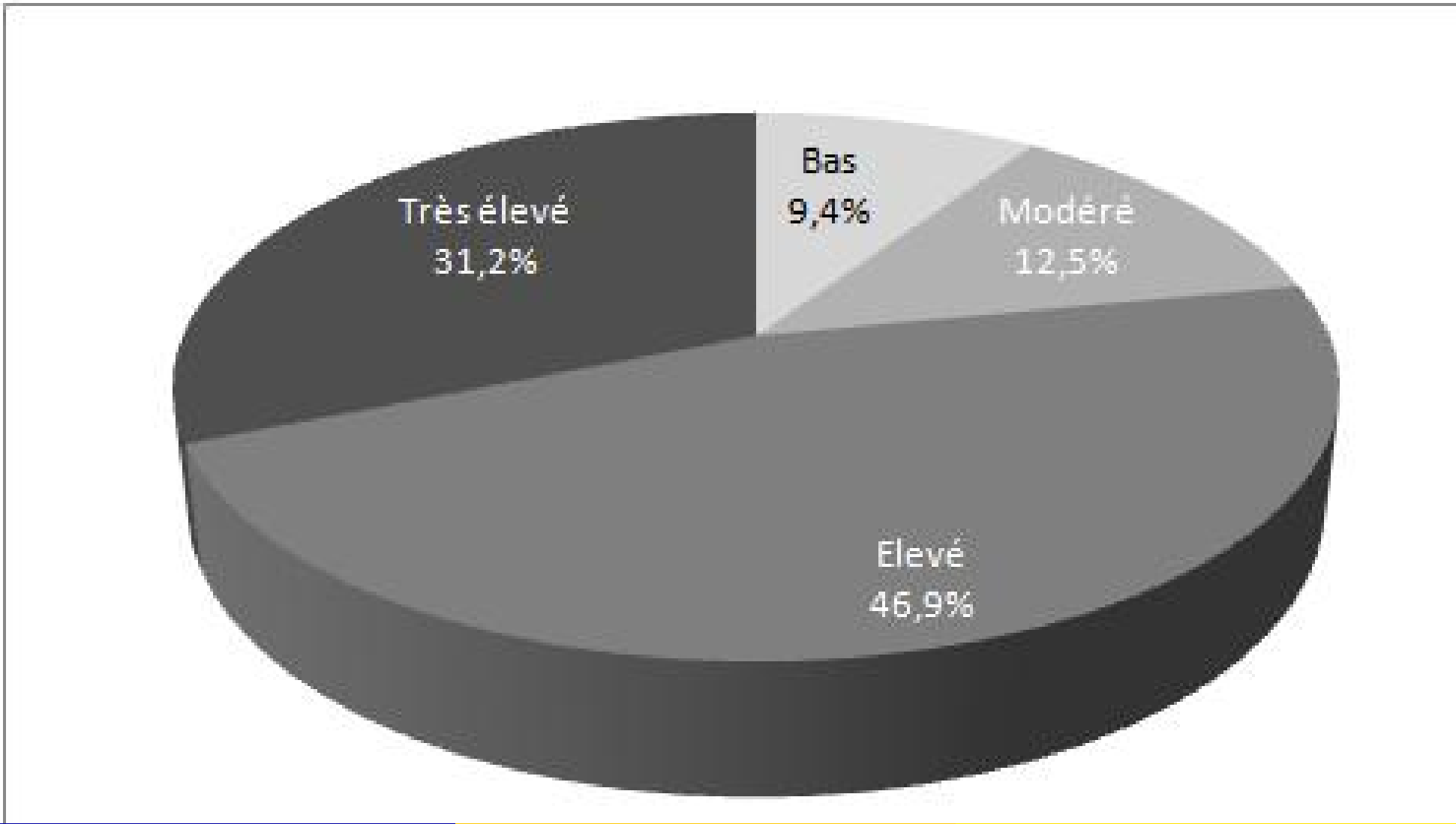


3.*jpg*

Ce diagramme est surtout utilisé dans la cas d'un caractère qualitatif. On divise le cercle en autant de seceurs circulaires que de classes. La mesure de l'angle d'un secteur est directement proportionnelle à à la fréquence relative correspondant à la modalité qu'il représente.

Exemple

Cas de degré de satisfaction des dernières vacances :

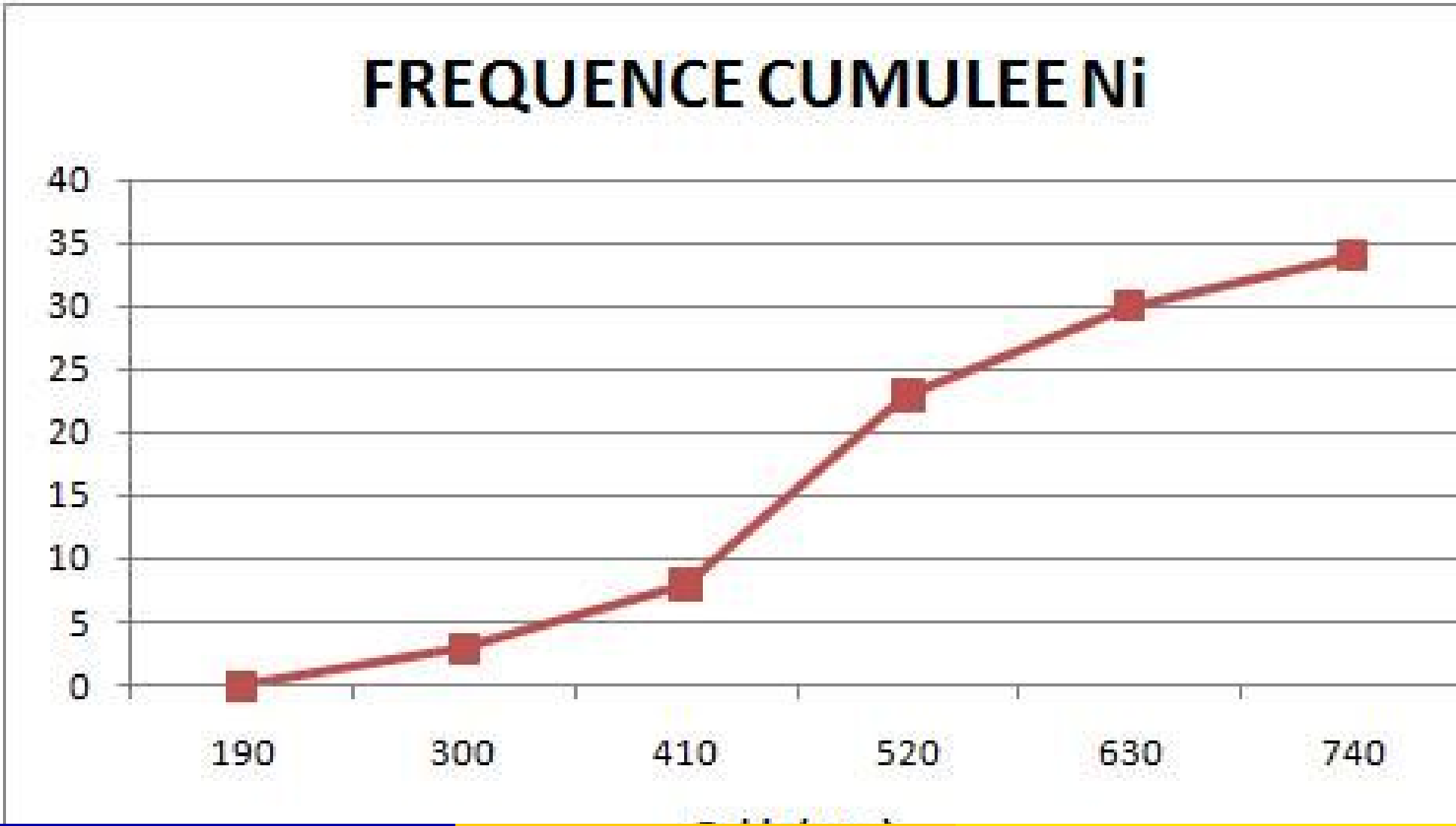


Polygône des fréquences cumulées croissantes (ou ogive)

Pour construire ce graphique, on place sur l'axe des abscisses les différents limites supérieures des classes, et sur l'axe des ordonnées les différentes fréquences cumulées croissantes $N_i = \sum_{j \leq i} n_j$ ou $F_i = \sum_{j \leq i} f_j$. Puis on relie les points obtenus par des segments de droite.

Exemple

Cas du poids des dorades d'élevage :



Mesures de tendance centrale

Introduction

Il est souvent nécessaire de résumer l'ensemble des informations qu'on possède sur une série statistique. Pour cela, on a recours aux mesures de tendance centrale, de dispersion et de position.
Les mesures de tendance centrale sont des valeurs autour desquelles se regroupent les différents résulats de la série. Elles donnent une idée sur l'ordre de grandeur des données ; il s'agit de mesures qui indiquent le score typique d'une distribution.
Les mesures de dispersion informent sur le degré de variabilité des valeurs de la série.
Les mesures de position renseignent sur la situation d'un résultat par rapport à l'ensemble de la distribution et sur la forme générale de celle-ci.

La moyenne (arithmétique)

Il s'agit de la mesure la plus utilisée. Est est spécifique des caractères quantitatifs. La moyenne peut cependant être très affectée par les mesures extrêmes. On note souvent par m ou μ dans le cas d'une population et par $\overline{x}$ dans le cas d'un échantillon.

Cas de données non groupées en classes

$$m = \frac{\sum_{i=1}^{N} x_i}{N}$$

où x_i représente la $i^{ème}$ donnée de la variable X mesurée sur une population de taille N.

Exemple

Moyenne de la classe

Cas de données groupées en classes

Dans le cas d'un caractère discret, on a :

$$m = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$$

où k est le nombre de modalités présentées par le caractère étudiée, n_i (resp. f_i) est l'effectif (resp. la fréquence relative) correspondant à la $i^{ème}$ classe.
Exemple : Cas du nombre d'enfants par foyer

Dans le cas d'un caractère continu, on ne peut calculer qu'une approximation de la moyenne donnée par :

$$\overline{x} \cong \frac{\sum_{i=1}^{k} n_i c_i}{n} = \sum_{i=1}^{k} f_i c_i$$

où c_i est le point milieu ou le centre de la $i^{ème}$ classe.

La médiane

La médiane est la valeur qui sépare, aussi exactement que possible, une série en deux parties égales par rapport au nombre de données, une fois celles-ci classées par ordre croissant. On note Me.

Cas de données non groupées en classes

Si le nombre des données est impair, la médiane est la valeur de la série de rang $\frac{n+1}{2}$.

Exemple

Soit la série : $2 \quad 7 \quad 5 \quad 3 \quad 4 \quad 8 \quad 5$

Le calssement ascendant est : $2 \quad 3 \quad 4 \quad 5 \quad 5 \quad 7 \quad 8$

La médiane est la donnée de rang 4, on a $Me = 5$

Si le nombre de données est pair, la médiane sera la moyenne des valeurs de rang $\frac{n}{2}$ et $\frac{n}{2} + 1$.

Exemple

Soit la série : $1 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 8$

On a : $Me = \frac{3+4}{2} = 3.5$

Cas de données groupées en classes

Graphiquement, on peut utiliser le polygone des fréquences cumulées et la médiane est l'ordonnée corresopndant à une fréquence cumulée de 50% ou à la moitié de l'effectif total.
Ou bien : après avoir localisé la modalité où se situe la médiane, on détermine une valeur approchée de celle-ci par interpolation linéaire.

On obtient :

$$Me \cong L_i + (\frac{50\% - F_{i-1}}{f_i}).a_i$$

Où : L_i est la borne inférieure de la classe contenant la médiane.

f_i est la fréquence associée à cette classe.

F_{i-1} est la fréquence cumulée correspondant à la modalité précedente.

a_i est l'amplitude de la classe contenant la médiane.

N.B. On peut remplacer les fréquences en pourcentage pa les effectifs.

Le mode

Le mode d'une série est la valeur la plus fréquente.
Lorsque les données sont groupées, on utilise le centre de la classe ayant une fréquence maximale comme approximation de la valeur modale ou bien on parle tout simplement de classe modale. On note Mo.

- $2 \quad 3 \quad 4 \quad 3 \quad 2 \quad 2 \quad \Longrightarrow Mo = 2$
- $1 \quad 5 \quad 6 \quad 3 \quad 9 \quad 11 \quad 7 \quad 4 \quad \Longrightarrow$ pas de mode.
- $2 \quad 2 \quad 3 \quad 4 \quad 3 \quad 2 \quad 3 \Longrightarrow Mo = 2$ ou 3 (deux modes ! !)

Remarques sur la représentativité des mesures de tendance centrale

La mesure de tendance centrale la plus utile est la moyenne. cependant, il peut arriver dans certains cas, que celle-ci s'avère affectée pa les valeurs extrêmes et donc moins représentative. ça sera le risque chaque fois qu'on utilisera une seule valeur pour représenter une large quantité de données statistiques.

Exemple

Considérons les salaires mensuels de groupes de salariés

- 10000 12000 12200 13000
- 10000 15000 15000 85000

Les moyennes des deux séries sont : $m_1 = 11800$ et $m_2 = 31250$

Les médianes des deux séries sont : $Me_1 = 12100$ et $Me_2 = 15000$

m_1 représente raisonnablement les données de la première série alors que m_2 parait loin des salaires de la majorité des salariés du deuxième groupe ; ceci est dû au fait que le salaire d'une seule personne s'éloigne beaucoup de ceux des autres. On conclut que la moyenne est sonvent influencée par les valeurs atypiques.
La médiane $Me_2 = 15000$ est plus représentative dans le cas de la deuxième série.

Dans le choix de la mesure centrale à utiliser, on ne sera jamais en mesure de satisfaire tous les critères de sélection et on sera amené à faire des compromis.

On ne peut pas évaluer la moyenne d'une distribution de fréquences dont la première ou la dernière modalité n'est pas bornée car, dans ce cas le centre de la classe n'est pas connu. Il faut alors utiliser la médiane ou le mode comme mesure de tendance centrale

Mesures de dispersion

Les mesures de dispersion renseignent sur le degré de variabilité des résultats.
Supposons que 10 étudiants aient obtenu les résultats suivant lors d'un premier examen (en ordre croissant) :
4 5 5 6 6 6 6 7 7 8
On a : $m = Me = Mo = 6$
Au deuxième examen, les résultats sont les suivants :
2 3 4 5 6 6 7 8 9 10
On a encore : $m = Me = Mo = 6$

On ne peut pas avancer qu'on est devant deux séries statistiques équivalentes. On remarque qu'il y a une variabilité plus accentuée des résultats du deuxième examen : dans le premier test les notes varient entre 4 et 8 alors que dans le second, les notes s'étalent entre 2 et 10. On peut conclure qu'il était plus facile pour qu'un étudiant, lors du deuxième test, de se démarquer (en mieux ou en pire) par rapport à la moyenne.
De ce fait, il convient de se donner des mesures pour décrire cet aspect important que constitue la dispersion ou la variabilité dans une série statistique.

L'étendue

Elle mesure l'écart entre la plus grande et la plus petite valeur de la série.

Exemple

Premier examen : Etendue $e = 4$

Second examen : Etendue $e = 8$

Remarque

Supposons que lors d'un troisième test, on a obtenu la distribution suivante :

0 6 6 6 6 6 6 6 8 10

On a toujours : $m = Me = Mo = 6$ et l'étendue est 10 ; ce qui laisse supposer qu'il y a une grande dispersion alors que presque toutes les les données sont égales à 6.

Il devient donc indispensable d'avoir recours à une mesure qui tient de toutes les données d'une série et non seulement des valeurs extrêmes.

La variance

Cette mesure évalue l'étalement des données d'une série par rapport à la moyenne ; elle donne l'écart quadratique moyen des données par rapport à leur moyenne arithmétique.

Cas de données non groupées

Si les données sont celles d'une population ou de tout un groupe d'individus, la variance, notée $V(X)$ ou σ^2 est définie par :

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - m)^2}{N}$$

On vérifie que :

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i)^2}{N} - m^2$$

Si les données recueillies sont celles d'un échantillon alors la variance, notée s^2, est donnée par :

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n-1}$$

On vérifie que :

$$s^2 = \frac{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}$$

Exemple

La série suivante présente le nombre d'bsences d'un échantillon de 6 étudiants à un cours de mathématiques : 2 7 9 4 6 7

On a : $\sum_{i=1}^{6} x_i = 35$; $m = 5.83$; $\sum_{i=1}^{6} x_i^2 = 235 \Longrightarrow s^2 = \frac{6(235)-35^2}{6\times 5} = 6.16$

k

Dans le cas discret,

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i(x_i - m)^2}{N} = \sum_{i=1}^{k} f_i(x_i - m)^2 = \sum_{i=1}^{k} f_i(x_i)^2 - m^2$$

$$s^2 = \frac{n(\sum_{i=1}^{k} n_i x_i^2) - (\sum_{i=1}^{k} n_i x_i)^2}{n(n-1)}$$

Dans le cas continu, on calcule une valeur approchée de la variance en remplaçant x_i par c_i, milieu de la $i^{ième}$ classe.

L'écart-type

L'écart type est la racine carrée de la variance ; il permet de mesurer l'écart moyen des données par rapport à leur moyenne ; il est donc exprimé dans l'unité de la variable. L'écart type σ ou s joue un rôle aussi important que la moyenne par rapport aux mesures de tendance centrale ; c'est l'indicateur de variabilité par excellence.

Le coefficient de variation

Considérons les deux séries :

$1 \quad 3 \quad 5 \quad 7 \quad 9 \quad \Longrightarrow \quad m_1 = 5\,; \quad \sigma_1 = 2.83$

$101 \quad 103 \quad 105 \quad 107 \quad 109 \quad \Longrightarrow \quad m_2 = 105\,; \sigma_2 = 2.83$

Les deux séries présentent un même écart-type. Une varibilité de 2.83 est très importante par rapport aux données de la première série et parait faible dans le cas de la seconde ; ceci est mis en évidence par le coefficient de variation qui mesure le % de variabilité par rapport à la moyenne.

$$c.v. = \left|\frac{\sigma}{m}\right| \times 100$$

Pour la première série $c.v. = 56.6\%$
Pour la sconde série $c.v. = 2.7\%$

Mesures de position : Les quantiles

Ces mesures renseignent sur la position d'un résultat par rapport à l'ensemble de la distribution et sur la forme générale de celle-ci.
Les *quantiles* correspondent aux valeurs numériques qui occupent des positions bien précises.
Donnons par exemple les *quartiles* qui sont les quantiles les plus utilisés.
Les quartiles divisent la série en quatre parties contenant chacune autant que possible 25% des données ordonnées par ordre croissant. Ces valeurs sont notées Q_1, Q_2 et Q_3. Notons que $Q_2 = Me$.

Exemple

Soit la série des salaires (en \$)

1650 1750 1680 1555 1510 1850 1690 1930 1740 2125 1720 1680

Classons les salaires

1510 1555 [1650 1680] 1680 1690 1720 1740 [1750 1850] 1930 2125

$\frac{n}{4} = 3$ et $\frac{3n}{4} = 9 \Longrightarrow \begin{cases} Q_1 = \frac{1650+1680}{2} = 1665 \\ Q_3 = \frac{1750+1850}{2} = 1800 \end{cases}$

Le salaire de 25% des employés est $\leqslant 1665$ \$
Le salaire de 25% des employés est $\geqslant 1800$ \$
50% des salaires gagnent entre 1665 et 1800 \$

L'*intervalle inter-quartile :* $[Q_1, Q_3]$ contient 50% de la population et laisse 25% de chaque côté.
L'*étendue inter-quartile* est $Qs = Q_3 - Q_1 = 135\ \$$ est l'amplitude de l'*intervalle inter-quartile* ; il mesure la dispersion de la population.
Le coefficient inter-quartile relatif est donné par :

$$c.i. = \frac{Qs}{Q_3 + Q_1}$$

Interessons nous maintenant au cas où les données sont groupées :
On utilise souvent le polygône des fréquences cumulées et on a, graphiquement, Q_1 (resp. Q_3) est l'abscisse correspondant à $\frac{n}{4}$ ou 25% (resp. $\frac{3n}{4}$ ou 75%) selon que les fréquences sont en nombre ou relatives.

En utisant une interpolation linéaire, on a, après avoir localisé la classe contenant la quartile en question :

$$Q_1 \cong L_i + (\frac{25\% - F_{i-1}}{f_i}).a_i$$

$$Q_3 \cong L_i + (\frac{75\% - F_{i-1}}{f_i}).a_i$$

En gardant les mêmes notations que celles concernant la médiane.

Remarque

On peut déterminer les *déciles* : $q_{\frac{1}{10}}$, $q_{\frac{2}{10}}$, ...,$q_{\frac{9}{10}}$

ou les *centiles* : $q_{\frac{1}{100}}$, $q_{\frac{2}{100}}$, ...,$q_{\frac{99}{10}}$