Chapitre1

Statistique descriptive

Statistique descriptive univariée (a une variable) :

1)- Définitions:

* statistique : la statistique est ensemble des méthodes qui servent à **organiser** les épreuves fournissant des observations, à **analyser** celles-ci et à **interpréter** les résultats.

L'analyse statistique se subdivise en deux parties

Statistique descriptive : a pour but de décrire c-à-d de résumer ou représenter les données.

Questions typiques
*Représentation graphique

*Paramètres de position, de dispersion, de relation.

Statistique infèrentielle : l'ensemble des méthodes permettant de formuler un jugement. Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).

Questions typiques
* Estimation des paramètres
* Intervalle de confiance
* Tests d'hypothèses
* Modélisation (exemple régression linéaire)

2)- Notions de bases :

*POPULATION La collection d'objets ou de personnes étudiées (élèves, habitants, voitures…).

*INDIVIDU élément de la population étudiée. (un élève, un habitant, une voiture,…).

*ECHANTILLON partie de la population étudiée. Nombre d'individus dans un échantillon noté n est appelé taille de l'échantillon

*VARIABLE (CARACTERE) propriété commune aux individus de la population, que l'on veut étudier.

Un caractère peut être :

a)-qualitatif : on ne peut associer ni valeur numérique ni un ordre naturel (type de voiture, couleur des cheveux,…).

b)- quantitatif : peut prendre des valeurs numérique (poids, longueur) un caractère quantitatif peut être :

*Continue : peut prendre toutes les valeurs numériques d'un intervalle déterminer (taille, poids…).

*Discontinue (discrète) : ne peut prendre que des valeurs numériques isolées (nombre de pièces d'habitations, nombre de fruits endommagés…).

MODALITE l'une des formes particulières d'un caractère. La couleur des yeux est un caractère, ses modalités sont : bleu, vert, marron,…

EFFECTIF OU FREQUENCE ABSOLUE (noté n_i) nombre d'apparitions de la valeur associé à un caractère dans un échantillon.

FREQUENCE RELATIVE (noté f_i) $f_i = \frac{n_i}{n}$.

SERIE STATISTIQUE l'ensemble des valeurs du caractère avec en regard, les fréquences absolue ou relative correspondantes.

*On appelle LES STATISTIQUES (au pluriel) des collections de nombres présentées sous forme de tableaux ou de graphique groupant des observations relatives à un phénomène considéré.

Exemple 1 :

Nombre d'enfant	0	1	2	3	4	5	total
Nombre de famille ou effectif : n_i	16	18	14	11	3	2	64
Fréquence relative : f_i	0,250	0,281	0,218	0,172	0,047	0,031	1

Population étudié : les familles

L'échantillon sur lequel porte l'étude : familles d'un immeuble ; n=64.

Le caractère étudié est le nombre d'enfants par famille. C'est un caractère quantitatif discret.

3) Traitement d'une série statistique :

*Série ordonnée : les valeurs obtenues peuvent être rangées par ordre de grandeur par exemple croissante. On obtient une série statistique ordonnée.

*Etendue de la série : la différence entre les deux valeurs extrêmes est appelée étendue de la série.

*Classe : quand le caractère étudié est quantitatif continu, la série statistique est répartie en classes ou intervalles semi ouverts. Le nombre de classes, k est calculé par l'une des deux formules :

LA règle de Sturge k=1+3.3log(n)

La règle de Yule $k=2.5(n)^{1/4}$

Centre de classe : on appelle centre de classe, la demi-somme des valeurs extrêmes de la classe. On note c_i le centre de la classe numéro i.

*Effectif cumulé : la somme des effectifs des i première classe est appelé effectif cumulé de la $i^{ème}$ classe on le note n_i^{cum} on n_i^c.

*Fréquence cumulée : le rapport $\frac{n_i^c}{n}$ est appellé fréquence cumulé de la $i^{ème}$ classe (n est la taille de l'échantillon).

Exemple2 : Le taux de glucose sanguin (glycémie) déterminé chez 32 sujets est donné ci-dessous en g/l

Série ordonnée :

0,85	0,95	1,00	1,06	1,11	1,19
0,87	0,97	1,01	1,07	1,13	1,20
0,90	0,97	1,03	1,08	1,14	
0,93	0,98	1,03	1,08	1,14	
0,94	0,98	1,03	1,10	1,15	
0,94	0,99	1,04	1,10	1,17	

Etendue de la série : 1,20 – 0,85 en g/l = 0,35 g/l.

Classe en g/l	$c_i g/l$	n_i	f_i	n_i cumulés
[0,85 ; 0,91[	0,88	3	3/32	3
[0,91 ; 0,97[	0,94	4	4/32	7
[0,97 ; 1,03[	1,00	7	7/32	14
[1,03 ; 1,09[	1,06	8	8/32	22
[1,09 ; 1,15[	1,12	6	6/32	28
[1,15 ; 1,21]	1,18	4	4/32	32
		$n = \sum n_i = 32$	$\sum f_i = 1$	

On a $n = 32$ et la formule de Yule donne

$$k = 2.5(32)1/4 = 5.94 \approx 6.$$

Au bas de la colonne n_i , on indique la somme de tous les n_i , $\sum n_i$ qui n'est autre que l'effectif total n de l'échantillon.

De la même façon au bas de la colonne des f_i on indique leur somme $\sum f_i$ qui doit être égal à 1.

La dernière colonne, dite des effectifs cumulés croissants a la signification suivante :

-Pour la classe [0,85 ; 0,91[: $n_i^c = 3$, on dit qu'il ya 3 valeurs inferieur à 0,91 g/l.

-Pour la classe [0,91 ; 0,97[: $n_i^c = 3 + 4 = 7$

il ya 7 valeurs inferieures à 0,97 (3 inferieures à 0,91 et 4 comprises entre 0,91 et 0,97).

-Pour la dernière classe on a donc $n_i^c = n$.

On appelle fréquence cumulée croissante pour la $i^{ème}$ classe le rapport $\frac{n_i^c}{n} = f_i^c$

On a donc :

-Pour la $1^{ère}$ classe $f_1^c = \frac{3}{32}$

-Pour la $2^{ème}$ classe $f_2^c = \frac{7}{32}$

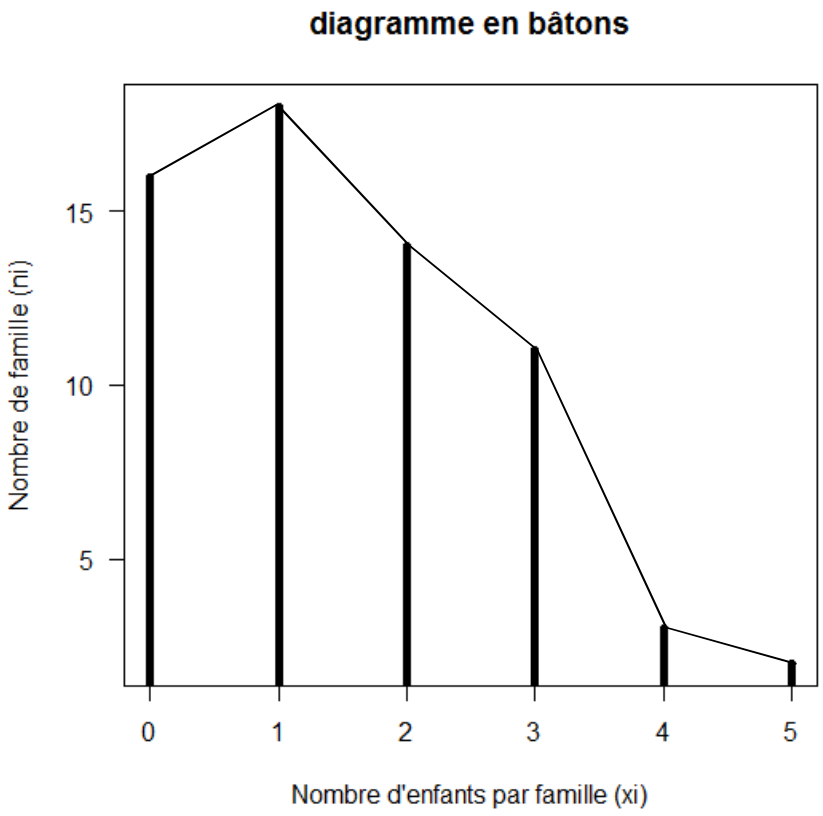
-Pour la dernière $f_6^c = \frac{32}{32} = 1$

On peut de la même façon concevoir des effectifs cumulés décroissants et des fréquences cumulés décroissantes.

4)- Graphes :

a)- **Diagramme en bâtons** : si on porte en abscisse les valeurs des n_i et si on trace à partir de chacun de ces points, un segment // à l'axe des ordonnées et de longueur l'effectif n_i on obtient un diagramme en bâton (si on joint les sommets des bâtons on obtient le polygone des fréquences).

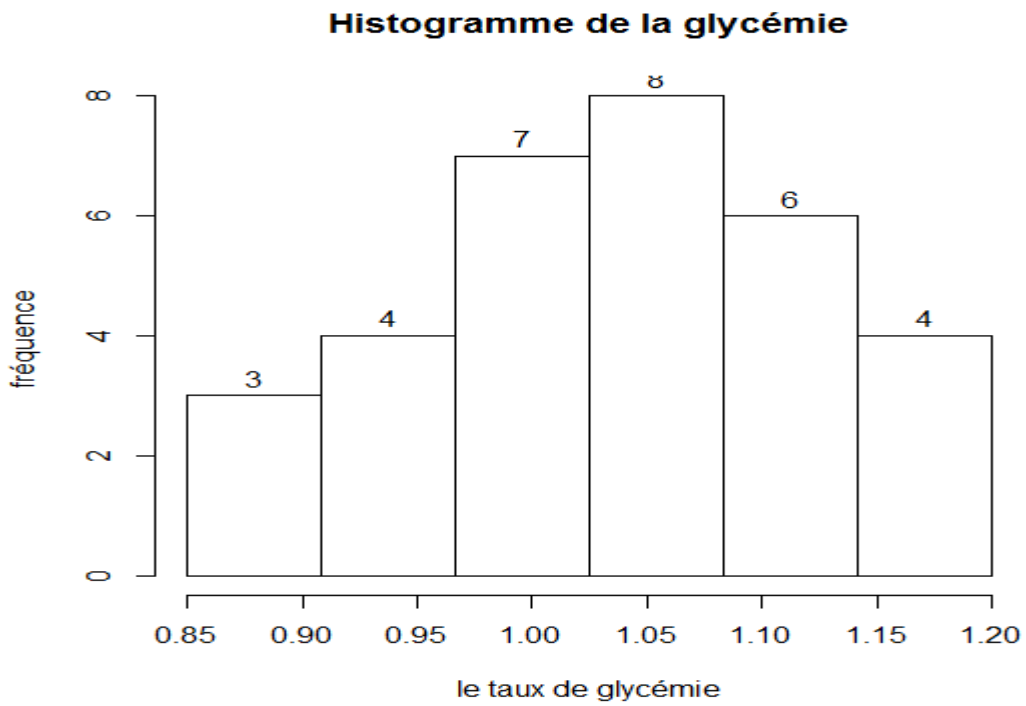
*Diagramme en bâtons de l'exemple1 :



b)- **L'histogramme** : lorsque le caractère étudié est continue on utilise un histogramme.

Chaque classe est représentée par un rectangle dont la base est égale à l'intervalle de la classe et dont la hauteur est égale à l'effectif correspondant. Le polygone des fréquences s'obtient en joignent les points d'abscisses les centres de classes et d'ordonnées les effectifs correspondants.

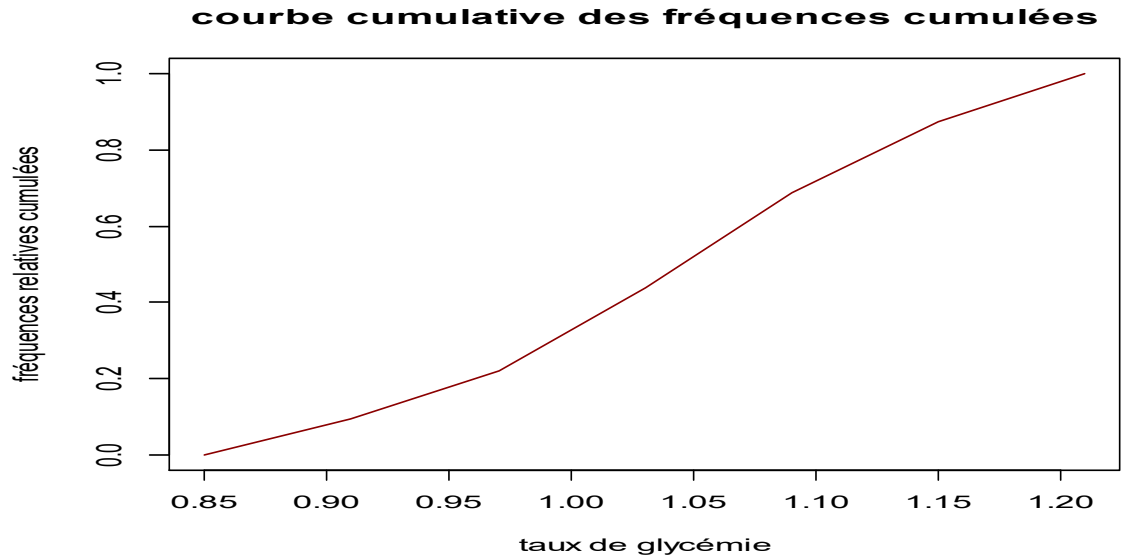
*histogramme de l'exemple 2



*<u>**La courbe cumulative (sigmoïde)**</u>

On considère les points dont les abscisses sont les limites supérieures des classes et d'ordonnées les n_i^c correspondants. La limite inferieure de la première classe à pour ordonnée le zéro. En reliant entre ces points par des segments, on obtient la courbe cumulative.

Le graphique suivant est la courbe cumulative de l'exemple2 :



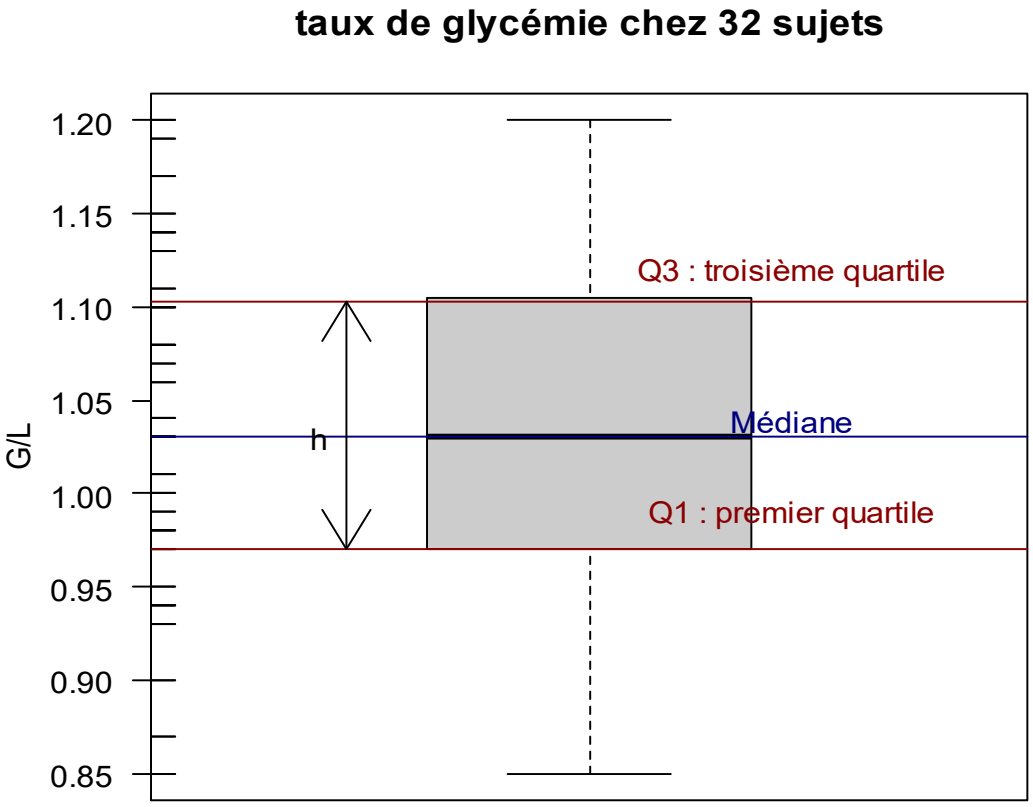
c)- <u>**Diagramme en boite**</u> **(la boite à moustaches)** : c'est un résumé visuel du sommaire d'une série de données ; la médiane, les quartiles, la plus petite et la plus grande valeur

de la série des valeurs aberrantes (les valeurs qui s'écarte de façon marquée de l'ensemble de l'ensemble des données).

Ce diagramme est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes.

REMARQUE une donnée qui s'écarte de façon marquée, d'après la règle de Tuckey, si elle s'écarte d'une distance de $1.5(Q_3 - Q_1)$ au dessus de Q_3 ou en dessous de Q_1.

Pour l'exemple 2



5)- Paramètres de position ou de tendance centrale :

a- **Le mode** : le mode d'un ensemble de nombres est la valeur qui y apparait le plus, c'est-à-dire la valeur dominante. Le mode peut ne pas exister et, même s'il existe, peut ne pas être unique (dans le cas continue on parle de classe modale).

Exemple : l'ensemble 2,2,5,7,9,9,9,10,10,11,12 et 18 a comme mode 9.

Exemple : l'ensemble 3, 5, 8, 10, 12,15 et 16 n'a pas de mode.

Exemple : l'ensemble 2, 3, 4, 4, 4, 5, 5, 7, 7,7 et 9 a deux mode 4 et 7. La série est appelée bimodale.

*Une série ayant un seul mode est appelée uni modale.

Exemple : dans le cas d'une variable continue, on applique la formule suivante ;

$$M_O = l_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} A_i$$

l_i : la limite inférieur de la classe modale

Δ_1 : la différence entre la fréquence de la classe modale et celle d'avant.

Δ_2 : la différence entre la fréquence de la classe modale et celle d'après.

A_i : la longueur de la classe modale.

Pour l'exemple 2, $M_O = 1.03 + \frac{(8-7)}{(8-7)+(8-6)} (1.09 - 1.03)$

$M_O = 1.05$.

b- **La médiane** : la médiane d'un ensemble de nombre rangés par ordre croissant est :

* la valeur du milieu si le nombre des données est impaire

* la moyenne arithmétique des deux valeurs du milieu si le nombre des données est pair.

Exemple : l'ensemble des nombre 3, 4, 4, 5, 6, 8, 8,8 et 10 a comme médiane 6.

Exemple : l'ensemble des données 5, 5, 7, 9, 11, 12,15 et 18 a comme médiane (9+11)/12= 10.

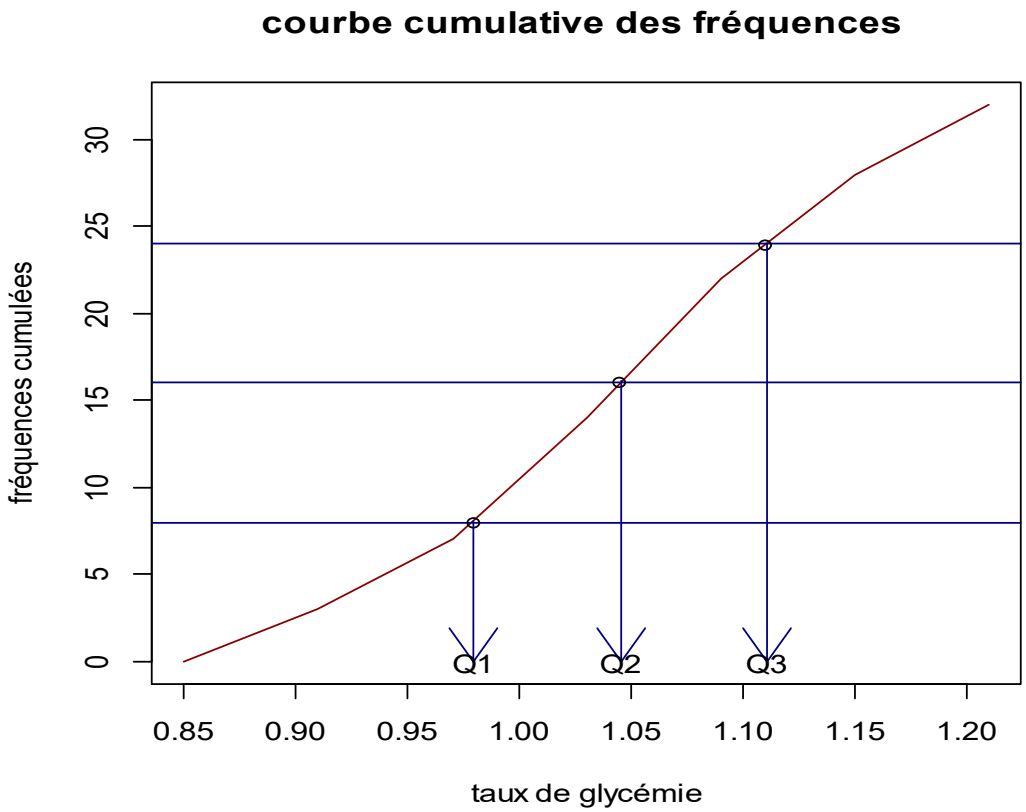
Pour déterminer la médiane dans le cas continue il est nécessaire de considérer les effectifs cumulés croissants ou décroissants et de chercher le cas échéant par interpolation, la valeur du caractère correspondant à 50% de l'effectif total.

Pour l'exemple 2 on a :

M_e est l'abscisse de $32 \times 50\% = 32 \times 0{,}5 = 16$

On a :

$$\frac{M_e - 1{,}03}{16 - 14} = \frac{1{,}09 - 1{,}03}{22 - 14} \Rightarrow \frac{M_e - 1{,}03}{2} = \frac{0{,}06}{8} \Rightarrow M_e = 1{,}045$$



c- **<u>Les percentiles</u>** : le $k^{ème}$ percentile est la valeur du caractère C_k.

-Telle que l'ensemble des individus dont le caractère est au plus égal à C_k représente les k % de l'effectif total.

-Telle que l'ensemble des individus dont le caractère est au moins égal à C_k représente les $(100 - k)$% de l'effectif total.

Parmi les percentiles, on distingue :

<u>Les déciles</u> pour lesquels $k = 10, 20, 30, \ldots..$

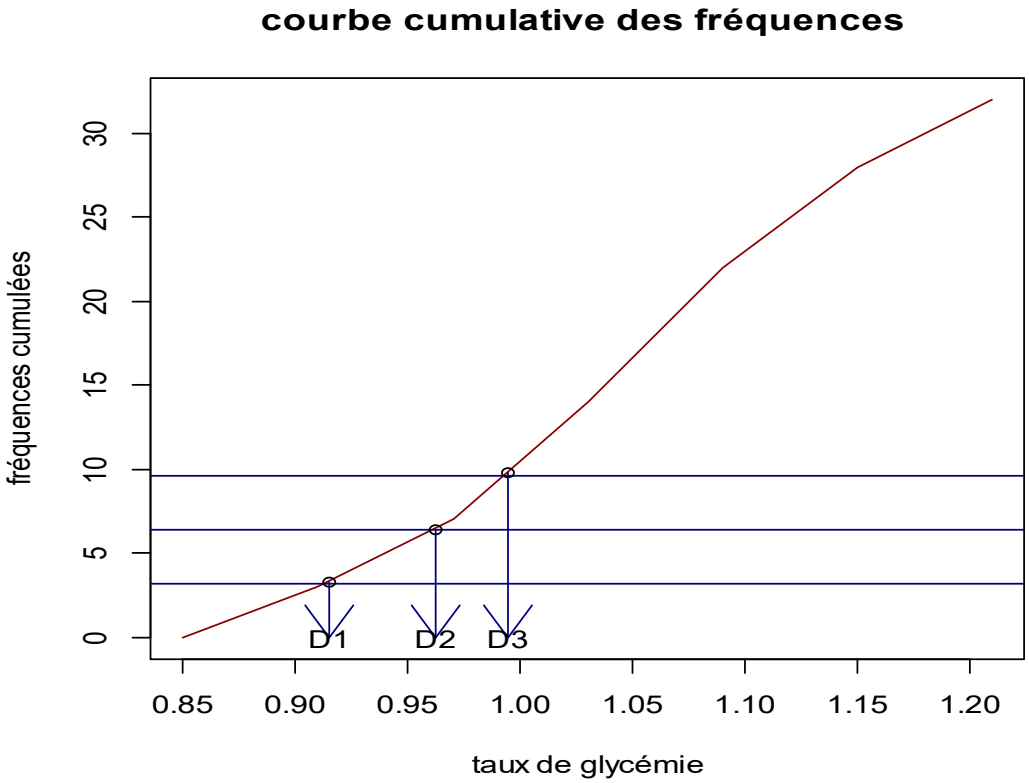
$$C_{10} = D_1 \qquad C_{20} = D_2 \quad \ldots\ldots\ldots$$

Pour l'exemple2 :

D_1 est l'abscisse de $32 \times 10\% = 32 \times 0{,}1 = 3{,}2$

$$\frac{D_1 - 0{,}91}{3{,}2 - 3} = \frac{0{,}97 - 0{,}91}{7 - 3} \Longrightarrow D_1 = 0{,}913$$

$D_2 = 0.961$, $D_3 = 0.992$



<u>Les quartiles</u>, pour lesquels $k = 25, 50$ et 75

$C_{25} = Q_1$ $C_{50} = Q_2$ $C_{75} = Q_3$

Pour l'exemple2 : Q_1 est l'abscisse de $32 \times 25\% = 32 \times 0{,}25 = 8$

$$\frac{Q_1 - 0{,}97}{8 - 7} = \frac{1{,}03 - 0{,}97}{14 - 7} \Longrightarrow Q_1 = 0{,}978$$

On a $Q_3 = 1{,}11$.

d-**<u>Moyenne arithmétique</u>** : Soit $x_1, x_2, x_3, \ldots, x_n$ une suite finie de nombres. La moyenne arithmétique est :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si chaque valeur x_i apparait n_i fois dans la série ; on peut encore écrire :

$$\bar{x} = \frac{1}{n}\sum_{i} n_i\, x_i.$$

On remarquant que $\frac{n_i}{n}$ est la fréquence relative f_i qui correspond à la valeur x_i , on a aussi :

$$\bar{x} = \sum_i f_i \, x_i$$

Exemple : dans l'exemple1 le nombre moyen d'enfants par famille est :

$$\bar{x} = \frac{16\times0+18\times1+14\times2+11\times3+3\times4+2\times5}{64} \cong 1{,}58$$

$$= 0{,}25 \times 0 + 0{,}281 \times 1 + 0{,}218 \times 2 + 0{,}17 \times 3 + 0{,}047 \times 4 + 0{,}31 \times 5 \cong 1{,}58\,.$$

Dans le cas de données groupées en classes, on prend pour valeurs des x_i les centres de classes. Dans l'exemple2 on a :

$$\bar{x} = \frac{3\times0{,}88+4\times0{,}94+7\times1{,}00+8\times1{,}06+6\times1{,}12+4\times1{,}18}{32} = 1{,}04 \;\; g/l$$

e)- **Comparaison des différents paramètres de position** :

* **La moyenne** arithmétique est peu sensible aux fluctuations d'échantillonnage. Elle se prête bien aux comparaisons. Des valeurs aberrantes peuvent toutefois la modifier sensiblement.

* **La médiane** est plus sensible aux fluctuations d'échantillonnage, elle l'est moins à des valeurs aberrantes. Toutefois, elle se prête moins bien à des calculs algébriques.

* **Le mode** est représentatif de la valeur du caractère le plus courant, le plus typique, mais il peut présenter une certaine ambigüité.

La comparaison des trois permet de se faire une idée plus complète de la distribution. (Si les trois sont a peu prés égales alors la série statistique est a peu prés symétrique).

6)- Paramètres de dispersions :

Un paramètre de dispersion se rapporte à la différence de deux valeurs du caractère. Alors qu'un paramètre de position représente une valeur du caractère.

a)- **Ecart moyen arithmétique** : il est donné par

$$\bar{E} = \frac{1}{n}\sum_i n_i \, |x_i - \bar{x}|.$$

b)- **Variance** : La variance du caractère dans l'échantillon, notée $s^2_{échan}$, est donnée par

$$s^2_{échan} = \frac{1}{n}\sum_i n_i(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2.$$

La variance du caractère dans la population, notée σ^2, est en général inconnue.

L'estimateur de la variance de la population, noté s^2, est donné par

$$s^2 = \frac{n}{n-1} s^2_{échan}.$$

c)- **Ecart – type** : est la racine de la variance $s = \sqrt{s^2}$.

Exemple : (pour l'exemple1) $s^2_{échan} = 7{,}75 \times 10^{-3}$, $s = 0{,}089$

b)- **Moment d'une série statistique** : On appelle moment d'ordre q par rapport à x_0 la quantité

$$m_q = \frac{1}{n}\sum_i n_i(x_i - x_0)^q.$$

Si $x_0 = 0 \; et \; q = 1 \; on \; a \;\; m_1 = \bar{x}$

Si $x_0 = \bar{x} \; et \; q = 2 \; on \; a \;\; m_2 = s^2$.

Remarque : les moments d'ordre q supérieur à 2 améliorent la représentation du caractère étudié.

Variable statistique qualitative

Définition les variables qualitatives contiennent des valeurs qui expriment une qualité comme le sexe, la couleur ou le nom

Elles peuvent être:

Nominale comme le nom des journaux, le nom des personnes, la couleur.

Ordinale désigne le rang ou la préférence comme: un peu, moyen, beaucoup

Les variables

Type de forfait: variable qualitative nominale à deux modalités: Ultra- prime et Super-plus

Réponses des abonnés: variable qualitative ordinale à deux modalités: satisfait et non satisfait

Effectifs correspondant à la modalité « satisfait » 200

La fréquence correspondante 200/300

La fréquence correspondante à la modalité « Forfait ultra- prime » 200/300

La proportion de « non satisfait » parmi les abonnés au « forfait ultra-prime »

50/200

Références

http://math.univ-lyon1.fr/~chekroun/Files/chekroun_statistiques.pdf

http://pf-mh.uvt.rnu.tn/32/1/SN1011.pdf