Chapitre 1 : Statistique descriptive

Définition la statistique est l'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).

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 est appelé taille de l'échantillon, noté n.

*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 :

*Continu : peut prendre toutes les valeurs numériques d'un intervalle déterminer (taille, poids…), il relève d'une mesure

*Discontinu (discrèt) : ne peut prendre que des valeurs numérique isolées (nombre de pièces d'habitations, nombre de fruits endommagés…), il relève d'un comptage.

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

Notons par Ω l'échantillon, C ensemble des modalités du caractère ou la variable X, alors :

Si X une variable statistique discrète ; $X : \Omega \to \{x_1, x_2, \dots x_k\}$, $X(\omega) = x_i$.

Si X une variable statistique continue ; $X : \Omega \to [\alpha, \beta]$.

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.

Exemple un opérateur téléphonique veut savoir si ses abonnés au téléphone portable sont satisfaits de leur forfait. Il fait une enquête auprès de 300 abonnés. Le tableau suivant résume les réponses obtenues ;

	Forfait Ultra - Prime	**Forfait Super -Plus**	**Total**
Satisfait	150	50	200
Non satisfait	50	50	100
Total	200	100	300

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

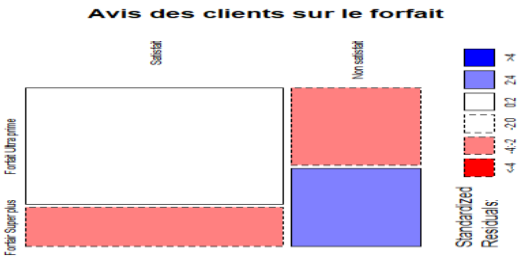
*La fréquence correspondante 200/300

*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.

Une des représentation graphique parmi plusieurs est la représentation mosaique



Etude d'une variable discrète :

Notons par Ω l'échantillon, C ensemble des modalités du caractère ou de la variable X, alors: Si X une variable statistique discrète ; $X : \Omega \to \{x_1, x_2, \dots x_k\}$, $x_1 < x_2 < \dots < x_k$, $X(\omega) = x_i$

Effectif total-Effectif partiel-Fréquence relative

Effectif total : noté n, la taille de l'échantillon, n=card(Ω).

Effectif partiel : noté n_i, le nombre d'individus qui ont la modalité x_i ;

n_i= card{ ω, $X(\omega) = x_i$}=$X^{-1}(\{ x_i \})$. $\sum_{i=1}^{k} n_i = n$.

Fréquence relative : noté f_i, la proportion des individus qui ont la modalité x_i ; $f_i = n_i/n$. $\sum_{i=1}^{k} f_i = 1$.

Effectif cumulé- Fréquence cumulée

Effectif cumulé : noté n_i^{cum} le nombre d'individus qui ont la modalité inférieure ou égale à x_i.

n_i^{cum}=card{ ω, $X(\omega) \leq x_i$}. $n_1^{cum} = n_1$ et $n_k^{cum} = n$

Fréquence cumulée : noté f_i^{cum} le pourcentage des individus qui ont la modalité inférieure ou égale à x_i. $f_i^{cum} = n_i^{cum}/n$. $f_1^{cum} = f_1$ et $f_k^{cum} = 1$.

Fonction de répartition

F_X est dite fonction de répartition de X si $F_X : \mathbb{R} \to [0,1]$,

$$F_X(x) = \begin{cases} 0 & si \ x < x_1 \\ f_i^{cum} \ si & x_i \leq x < x_{i+1} \quad 1 \leq i \leq k-1 \\ 1 & si \ x \geq x_k \end{cases}$$

F_X représente le pourcentage des individus ω tels que $X(\omega) \leq x$.

Représentation graphique de F_X appelée la courbe cumulative des fréquences, c'est le graphe d'une fonction en escalier.

Présentation des données sous forme de tableaux

Les données ou la série statistique sont résumé sous forme de tableaux qui comprend les différentes modalités x_i, les effectifs et les fréquences relatifs ainsi que les effectifs et fréquences cumulés.

x_i	n_i	f_i	n_i^{cum}	f_i^{cum}

Exemple

Nombre d'enfants : x_i	Nombre de famille : n_i	f_i	n_i^{cum}	f_i^{cum}
0	16	0.25	16	0.25
1	18	0.281	34	0.531
2	14	0.218	48	0.749
3	11	0.172	59	0.921
4	3	0.047	62	0.968
5	2	0.031	64	$0.999 \cong 1$
Total	64	1		

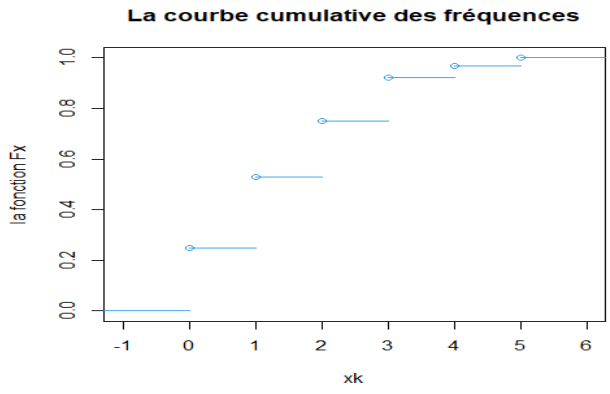
Ω=les familles d'un immeuble

X : le nombre d'enfants par famille, une variable statistique discrète.

L'ensemble des modalités {0,1,2,3,4,5}

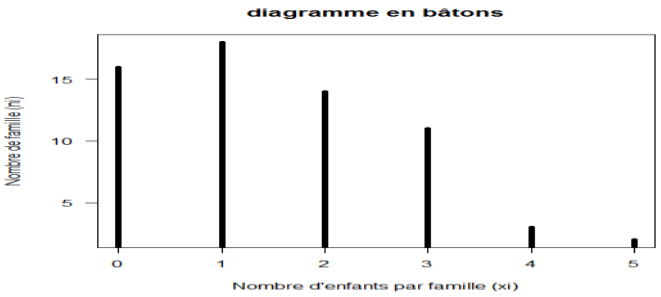
Déterminons sa fonction de répartition $F_X(x) = \begin{cases} 0 & si \quad x < 0 \\ 0.25 & \text{si } 0 \leq x < 1 \\ 0.531 & si \quad 1 \leq x < 2 \\ 0.749 & si \quad 2 \leq x < 3 \\ 0.921 & si \quad 3 \leq x < 4 \\ 0.968 & si \quad 4 \leq x < 5 \\ 1 & si \quad x \geq 5 \end{cases}$

Sa représentation graphique sera alors sous la forme suivante



Représentation graphique des données

Le diagramme en bâton est une des représentations les plus utilisée. En voici celle de l'exemple précédent



Paramètres de position ; médiane, mode, moyenne :

Le mode : noté M_o, est la valeur qui apparaît dans une série statistique plus que les autres, c'est le x_i qui a le plus grand n_i.

La médiane : noté M_e, est la valeur qui sépare une série statistique ordonnée en deux parties égales.

La moyenne : noté $\bar{x}=\frac{1}{n}\sum_{i=1}^{k} n_i\, x_i = \sum_{i=1}^{k} f_i\, x_i$.

Exemple d'après l'exemple précédent ; $M_o=1$, $M_e=1$,

$\bar{x} = \frac{16*0+18*1+14*2+11*3+3*4+2*5}{64} = 1.58$.

Paramètres de dispersion ; variance, écart type.

Variance : $V=\frac{1}{n}\sum_{i=1}^{k} n_i\,(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{k} n_i\, x_i{}^2-\bar{x}^2=\overline{x^2}-\bar{x}^2$.

écart type $\sigma=\sqrt{V}$

Etude d'une variable continue :

Quand la variable statistique est continue, les données sont regroupées en classes ou en intervalles semi ouverts, de nombre k. Ce nombre est calculé par une des deux formules

La règle de Sturge $k=1+3.3\log(n)$

La règle de Yule $k=2.5\,(n)^{1/4}$

n est le nombre de valeurs.

*L'étendue d'une série statistique $e=x_{max}-x_{min}$.

*l la longueur de la classe $l>e/k$.

*c_i centre d'une classe [a ,b[, $c_i=(a+b)/2$.

Effectif total-Effectif partiel-Fréquence relative

Effectif total :$n=card(\Omega)$.

Effectif partiel : n_i le nombre de valeurs de la série qui appartiennent à l'intervalle (ou la classe).

Fréquence relative : noté f_i, la proportion des individus qui ont la modalité x_i ; $f_i = n_i/n$. $\sum_{i=1}^{k} f_i = 1$.

Effectif cumulé- Fréquence cumulée

Effectif cumulé : noté n_i^{cum} le nombre d'individus qui ont la modalité inférieure ou égale à x_i.

$n_i^{cum} = card\{\omega, X(\omega) \leq x_i\}$. $n_1^{cum} = n_1$ et $n_k^{cum} = n$

Fréquence cumulée : noté f_i^{cum} le pourcentage des individus qui ont la modalité inférieure ou égale à x_i. $f_i^{cum} = n_i^{cum}/n$. $f_1^{cum} = f_1$ et $f_k^{cum} = 1$.

Fonction de répartition

Soit X une V.S définie de Ω vers $[a_0, a_1[\cup [a_1, a_2[\cup \ldots \cup [a_{k-1}, a_k]$

F_X est dite fonction de répartition de X si $F_X : \mathbb{R} \to [0,1]$,

$$F_X(x) = \begin{cases} 0 & si \quad x < a_0 \\ f_1 \frac{x - a_0}{l} & \text{si} \quad a_0 \leq x < a_1 \\ \vdots & \\ f_{i-1}^{cum} + f_i \frac{x - a_i}{l} & si \quad a_i \leq x < a_{i+1} \\ \vdots & \\ 1 & si \quad x \geq a_m \end{cases}$$

Représentation graphique de F_X appelée la courbe cumulative des fréquences, c'est le graphe d'une fonction croissante continue.

Présentation des données sous forme de tableaux

Les données ou la série statistique sont résumé sous forme de tableaux qui comprend les différentes classes C_i, les centres de classes c_i, les effectifs et les fréquences relatifs ainsi que les effectifs et fréquences cumulés.

Ci	C_i	n_i	f_i	n_i^{cum}	f_i^{cum}

Exemple: 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

Population étudié : sujets humains

L'échantillon sur lequel porte l'étude : 32 sujets.

Le caractère étudié est Le taux de glucose sanguin. C'est un caractère quantitatif continu.

* On a $n = 32$ et la formule de Yule donne dans ce cas

$$k = 2.5(32)^{\frac{1}{4}} = 5.94 \approx 6.$$

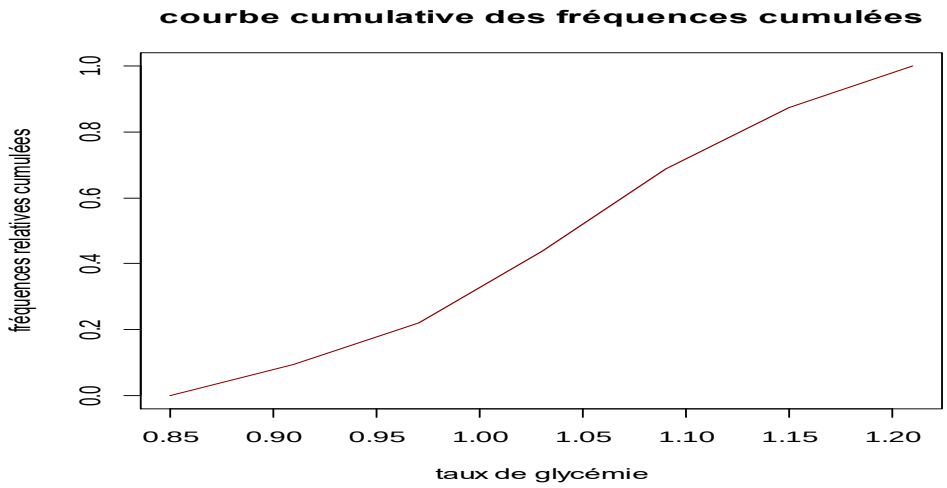
Etendue de la série: e= 1,20 – 0,85 en g/l = 0,35 g/l.

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

Déterminons la fonction de répartition

$$F_X(x)=\begin{cases} 0 & si \quad x > 0.85 \\ (3/32)\frac{x-0.85}{0.06} & si \quad 0.85 \le x < 0.91 \\ (3/32)+(4/32)\frac{x-0.91}{0.06} & si \quad 0.91 \le x < 0.97 \\ (7/32)+(7/32)\frac{x-0.97}{0.06} & si \quad 0.97 \le x < 1.03 \\ (14/32)+(8/32)\frac{x-1.03}{0.06} & si \quad 1.03 \le x < 1.09 \\ (22/32)+(6/32)\frac{x-1.09}{0.06} & si \quad 1.09 \le x < 1.15 \\ (28/32)+(4/32)\frac{x-1.15}{0.06} & si \quad 1.15 \le x < 1.21 \\ 1 & si \quad x \ge 1.21. \end{cases}$$

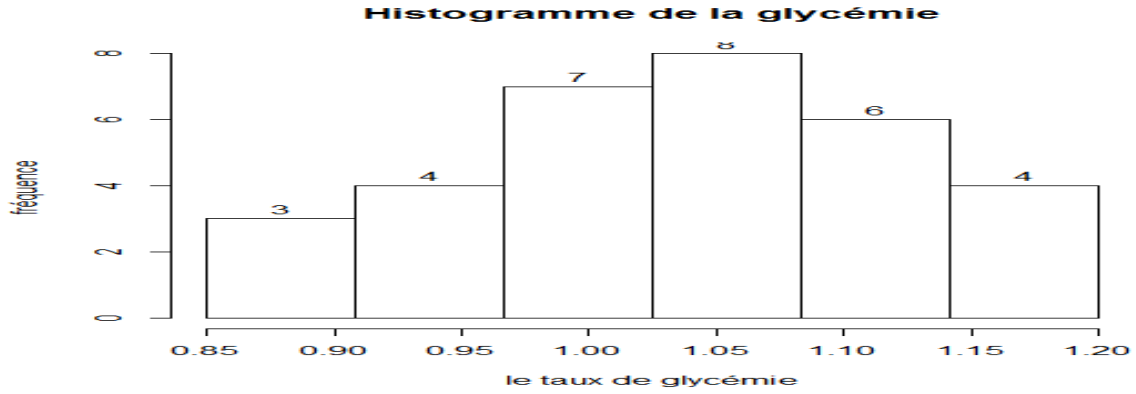
La représentation de F_X



Représentation graphique des données

Lorsque le caractère étudié est continue on utilise un histogramme ; Chaque classe est représentée par un rectangle dont la base est égale à la longueur de la classe et dont la hauteur est égale à l'effectif correspondant.

Exemple : pour l'exemple du taux de glycémie, on a l'histogramme suivant



Paramètres de position ; médiane, mode, moyenne :

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

$M_e \in [a_i, b_i[$, $l = b_i - a_i$, n la taille de l'échantillon,

$$M_e = a_i + l\frac{0.5 * n - n_{i-1}^{cum}}{n_i^{cum} - n_{i-1}^{cum}}$$

mode: Dans le cas continu, on parle de classe modale, c'est la classe qui a le plus grand n_i.

$M_O = a_i + \frac{\Delta_1}{\Delta_1+\Delta_2} l_i$;

a_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.

l_i : la longueur de la classe modale.

Moyenne : Pour la moyenne on applique la même formule du cas discret en remplaçant les x_i par les c_i centre de classes.

Exemple : Considérons le même exemple

***** $M_e \in [1{,}03 ; 1{,}09[$, $L = 1{,}09 - 1{,}03 = 0{,}06$, $n = 32$

$$M_e = 1{,}03 + 0{,}06\frac{16-14}{22-14}$$

$$\Longrightarrow M_e = 1{,}045$$

*M_o=1.03+$\frac{(8-7)}{(8-7)+(8-6)}(1.09-1.03)$

$M_O = 1.05.$

* $\bar{x}$ =(0.88*3+0.94*4+1*7+1.06*8+1.12*6+1.18*4)/32=1.04

Paramètres de dispersion : variance -écart type

Variance V=$\frac{1}{n}\sum_{i=1}^{k} n_i\,(c_i-\bar{x})^2$=$\frac{1}{n}\sum_{i=1}^{k} n_i\,{c_i}^2$-$\bar{x}^2$=$\overline{x^2}$- $\bar{x}^2$.

écart type σ=$\sqrt{V}$.

Exemple : (0.88^2*3+0.94^2*4+1^2*7+1.06^2*8+1.12^2*6+1.18^2*4)/32-1.04^2=0.01035

Ecart type= 0.1017