Cours 1 : Statistique descriptive

1-Terminologie

La statistique est l'ensemble des méthodes scientifiques qui consistent à réunir, à organiser, à résumer des données puis à analyser, à commenter et à critiquer ces données.

Exp : On observe le poids de 1000 enfants d'âge de 6 ans. Les données sont $x_1, x_2, \dots, x_{1000}$.

L'étude statistique porte sur un ensemble d'éléments qui peuvent être des personnes, des objets, des animaux,…D'où la définition suivante :

Définition :

-On note Ω une population : qui est l'ensemble sur lequel porte l'étude statistique.

-Les éléments de Ω sont notés ω , ils sont appelés individus (ω ϵ Ω).

Définition : On appelle variable statistique (ou caractère) toute application de Ω dans un ensemble C.

X : Ω → C

ω → X(ω) = x : donnée ou observation. X est notée V.S.

Remarque :

-Si C⊂ $\mathbb{R}$, alors X est dites V.S. quantitative.

-Si C⊄ $\mathbb{R}$, alors X est dites V.S. qualitative.

2-Etude d'une V.S. quantitative

1-V.S. discrète :

Définition : Soit X : Ω → C = $\{x_1, x_2, \dots, x_m\}$ où $x_1 \leq x_2 \leq \cdots \leq x_m$ (les x_i sont des valeurs isolées et ordonnées). X est dite V.S. discrète.

Exp :

1)Ω : les étudiants de première année STU,

X : la note en EMD1,

C = $\{0, 1, \dots, 20\}$.

2)Ω : les familles d'un quartier,

X : le nombre d'enfant par famille,

C = $\{0, 1, \dots, 7\}$.

3)Exemple de qualité : Une usine fabrique des pièces mécaniques, après fabrication on procède au contrôle d'un lot (un échantillon) de 500 pièces. On s'Intéresse au nombre de défauts par pièce. Les résultats sont résumés dans le tableau suivant :

X= x_i	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$	$x_5 = 4$
n_i	112	170	128	60	30

a-Effectif n_i et fréquence f_i

Définition de l'effectif n_i : Le nombre n_i s'appelle effectif de la valeur x_i

Exp :

D'après le tableau n_1= 112 est l'effectif de la valeur $x_1 = 0$

Proposition : N = card Ω =$\sum_{i=1}^{m} n_i = n_1 + n_2 + \cdots + n_m$

Définition de la fréquence f_i : Le nombre f_i = $\frac{n_i}{N}$ s'appelle fréquence de la valeur x_i

Exp :

X= x_i	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$	$x_5 = 4$
f_i	0,224	0,340	0,256	0,120	0,060

Remarque : f_i représente le pourcentage des individus qui ont un caractère égale à x_i

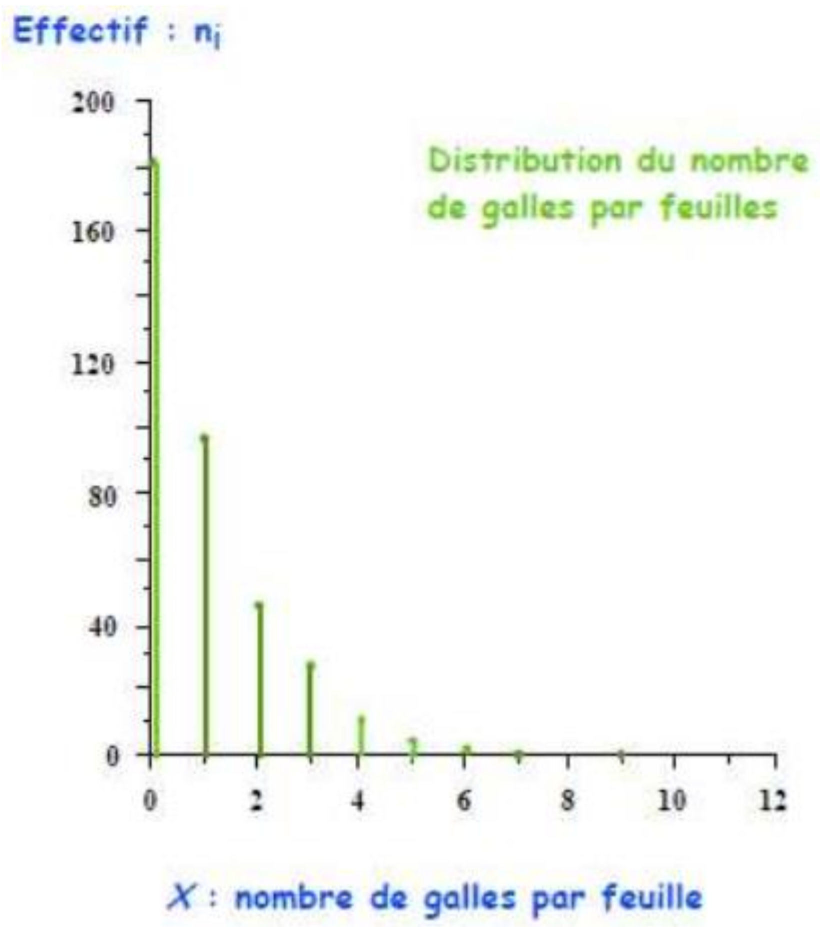
Exp : Il y a 34% de pièces qui ont un défaut.

Proposition : $\sum_{i=1}^{m} f_i = f_1 + f_2 + \cdots + f_m = 1$.

Diagramme en bâtons des effectifs :

Exp : La cécidomyie du h^^être provoque sur les feuilles de cet arbre des galles dont la distribution des effectifs est la suivante :

Caractère X : x_i: nombre de galles par feuille	0	1	2	3	4	5	6	7	8	9	10
n_i : nombre de feuilles portant x_i galles	182	98	46	28	12	5	2	1	0	1	0



Remarque : Le diagramme en bâtons des fréquences est analogue au diagramme en bâtons des effectifs.

b-Effectif cumulé N_i et fréquence cumulé F_i

Définition de l'effectif cumulé N_i : Le nombre $N_i = n_1 + n_2 + \cdots + n_i$ s'appelle effectif cumulé de la valeur x_i

Remarque : $N_m = \sum_{i=1}^{m} n_i = n_1 + n_2 + \cdots + n_m = N$= card Ω

Définition de la fréquence cumulée F_i : Le nombre $F_i = \frac{N_i}{N} = f_1 + f_2 + \cdots + f_i$ s'appelle fréquence cumulée de la valeur x_i

Exp :

$X = x_i$	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	$x_4 = 3$	$x_5 = 4$
N_i	112	282	410	470	500
F_i	0 ,224	0,564	0,820	0,940	1

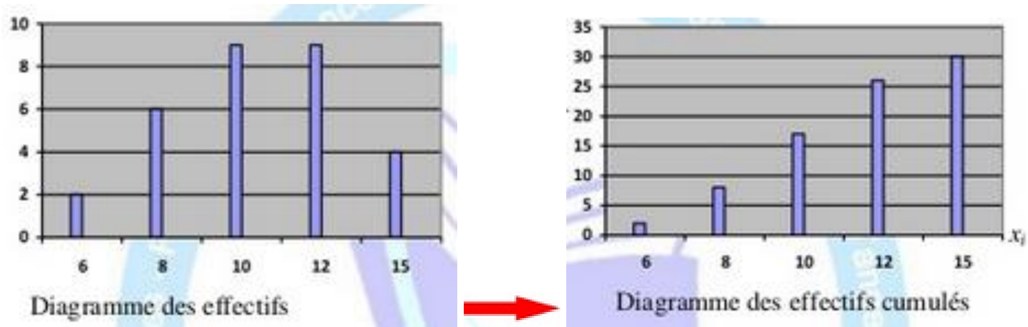
Remarque : F_i représente le pourcentage des individus qui ont un caractère inférieur ou égale à x_i

Exp : Il y a 94% de pièces qui ont au plus trois défauts.

Remarque : $F_m = \sum_{i=1}^{m} f_i = f_1 + f_2 + \cdots + f_m = 1$.

Diagramme en bâtons des effectifs cumulés :

Exp :



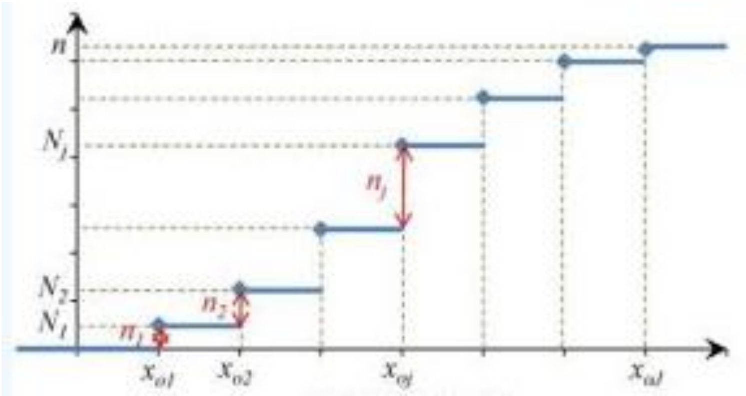
Remarque : Le diagramme en bâtons des fréquences cumulées est analogue au diagramme en bâtons des effectifs cumulés.

La courbe cumulative des effectifs cumulés :

Définition : La courbe cumulative des effectifs cumulés est donnée par :
$N(x) = card\{\omega\epsilon\Omega\ t.q. X(\omega) \leq x\}$ pour tout $x\epsilon\mathbb{R}$.

Remarque : Si $x = x_i$ alors $N(x) = N_i$

$$N(x) = \begin{cases} 0 & si \quad x < x_1 \\ N_1 & si \quad x\epsilon[x_1 , x_2[\\ & \vdots \\ N_i & si \quad x\epsilon[x_i , x_{i+1}[\\ & \vdots \\ N & si \quad x \geq x_m \end{cases}$$



La courbe cumulative des effectifs cumulés

La courbe cumulative des fréquences cumulés :

Définition : La courbe cumulative des fréquences cumulées est donnée par :

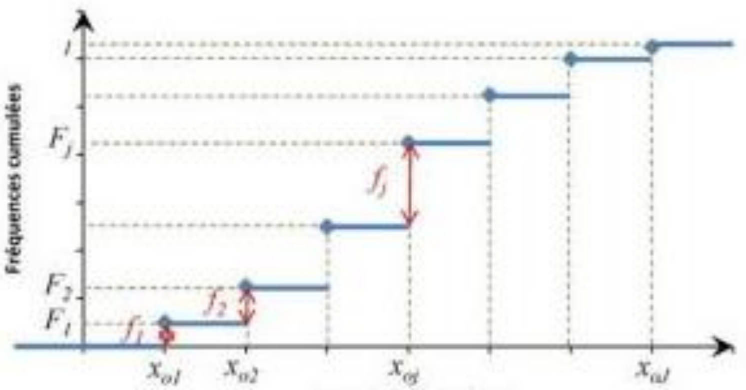
F : ℝ → ℝ

x → F(x)

F(x) représente le pourcentage des individus qui ont un caractère inférieur ou égale à x.

Remarque : Si $x = x_i$ alors $F(x) = F_i$

$$F(x) = \begin{cases} 0 & si \quad x < x_1 \\ F_1 & si \quad x \epsilon [x_1 , x_2[\\ & \vdots \\ F_i & si \quad x \epsilon [x_i , x_{i+1}[\\ & \vdots \\ 1 & si \quad x \geq x_m \end{cases}$$



La courbe cumulative des fréquences cumulés

2-V.S. continue :

Définition : Soit X :Ω →C = $[\alpha, \beta] \subset \mathbb{R}$. On appelle $e = \beta - \alpha$ l'étendue de C.

Exp : Ω : les étudiants de la première année STU,

X : la taille de chaque étudiant,

C : l'ensemble des valeurs prises par X entre 1,5m et 1,98m.

Remarque : Vu le nombre très grand des valeurs entre 1,5 et 1,98 on assimile C à l'intervalle $[1{,}5\,;1{,}98]$.

Pour étudier une V.S. continue, on se ramène à une V.S. discrète en considérant la notion de classe. Nous utilisons le procédé suivant :

on partage C $= [\alpha, \beta]$ en m classes, avec $m = \sqrt{N}\ \big(N = card(\Omega)\big)$.

On choisit les classes de même longueur h.

On note par $\mathcal{E}_i = [a_{i-1}, a_i]$ la i[ème] classe $i = 1, 2, \cdots, m$ et par c_i le centre de cette classe t.q. $c_{i=\frac{a_{i-1}+a_i}{2}}$, d'autre part $h = a_i - a_{i-1}$

h est choisit t.q. $h > \frac{e}{m}$

à chaque classe $\mathcal{E}_i$ on associe :

*l'effectif n_i

*la fréquence $f_i = \frac{n_i}{N}$

*l'effectif cumulé $N_i = \sum_{k=1}^{i} n_k = n_1 + n_2 + \cdots + n_i$

*la fréquence cumulée $F_i = \frac{N_i}{N} = f_1 + f_2 + \cdots + f_i$

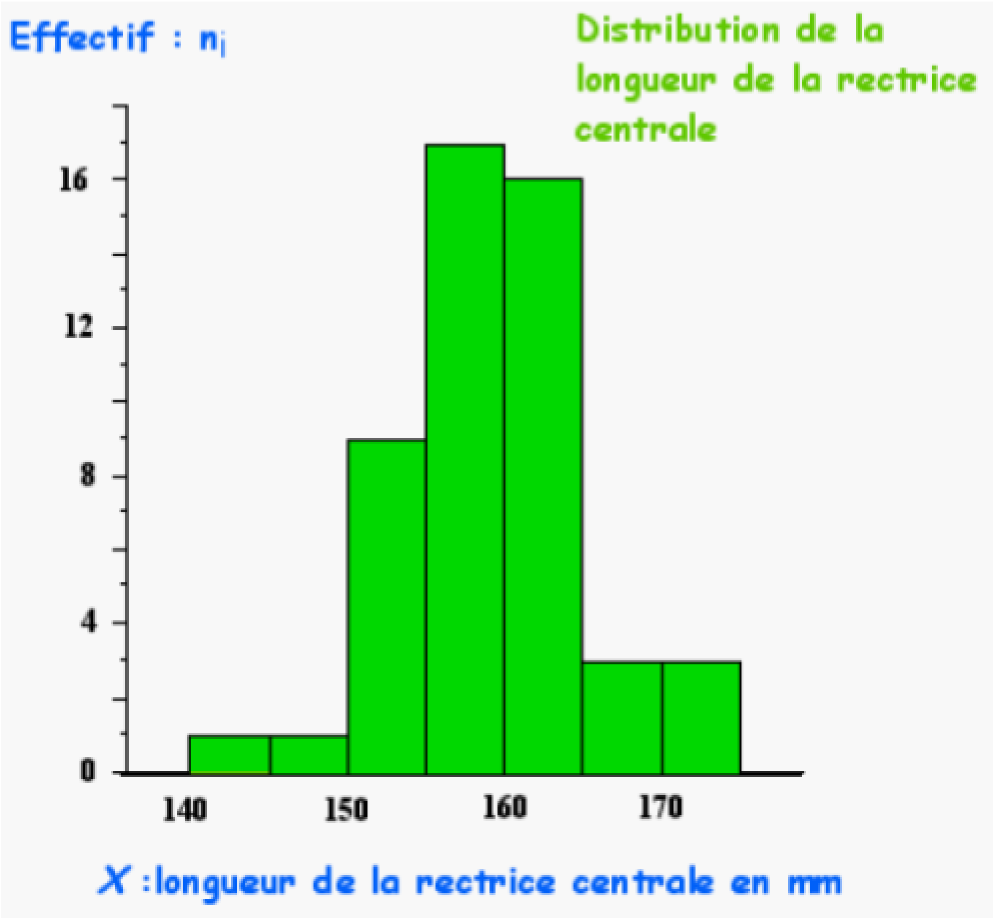
Ainsi la V.S. continue est discrétisée et son étude sera identique à celle d'une V.S. discrète.

La représentation graphique :

a-L 'histogramme des effectifs : C'est une suite de rectangles associés à chaque classe $\mathcal{E}_i$ la base du rectangle ayant pour longueur h et pour hauteur n_i.

Exemple : Dans le cadre de l'étude de la population de gélinottes huppées (la gélinotte huppée est un genre d'oiseau), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

Caractère X : x_i : longueur de la rectrice bornes des classes	[140-145[	[145-150[	[150-155[	[155-160[	[160-165[	[165-170[	[170-175[
n_i : nombre d'individu par classe de taille x_i	1	1	9	17	16	3	3



L'histogramme des effectifs

<u>b-Le polygone des effectifs :</u> C'est une ligne brisée reliant les points (c_i, n_i) $i = 1, 2, \cdots, m$ où $c_i = \frac{a_{i-1}+a_i}{2}$

<u>Remarque :</u>

-L'histogramme des fréquences est analogue à l'histogramme des effectifs.

-De même le polygone des fréquences est analogue au polygone des effectifs.

<u>c-La courbe cumulative des fréquences cumulées (ou fonction de répartition) :</u> à l'histogramme des fréquences cumulées correspond une courbe continue $F(x)$

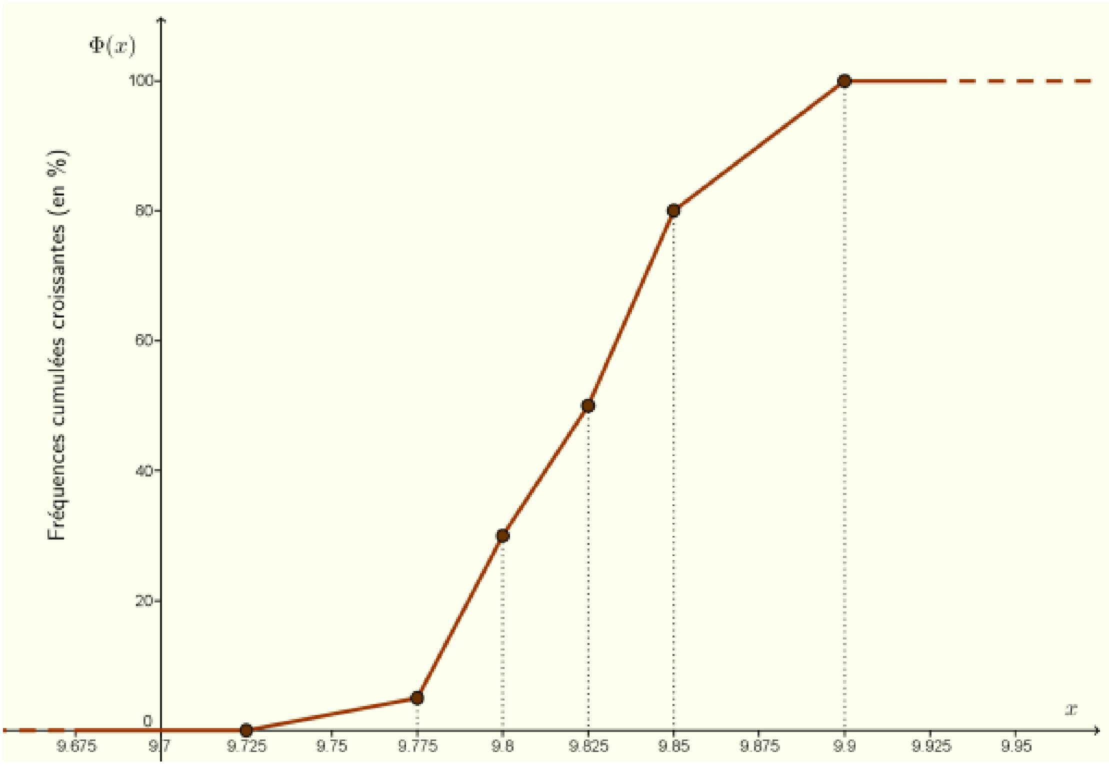
F : $\mathbb{R} \to \mathbb{R}$

$x \to$ F(x)

Et sa courbe est une ligne brisée qui relie les points $(a_0, 0), (a_1, F_1), \cdots, (a_m, 1)$

$$F(x) = \begin{cases} 0 & si \quad x < a_0 \\ \vdots & \\ F_i + \frac{f_{i+1}}{a_{i+1}-a_i}(x - a_i) & si \quad x \epsilon [a_i \,, a_{i+1}[\\ \vdots & \\ 1 & si \quad x \geq a_m \end{cases}$$

Remarque : Le chiffre F(x) représente le pourcentage d'individu dont la valeur du caractère X est inferieure ou égale à x.



Courbe cumulative des fréquences cumulées

3-Caractéristiques de position et de dispersion :

a-La médiane :

Définition : La médiane M est la valeur de X qui partage Ω à deux effectifs égaux.

Remarque : $F(M) = \frac{1}{2}$

*Cas discret :

•Si $F_i \neq \frac{1}{2} \forall i$, dans ce cas on a : $F(x_p - 0) < \frac{1}{2} \leq F(x_p + 0) \Rightarrow M = x_p$

•Si $F_i = \frac{1}{2} \Rightarrow M = x_i$

On conclut alors que $F(M-0) < \frac{1}{2} \leq F(M+0)$

Cas continue :

Par définition on a $F(M) = \frac{1}{2}$, si on suppose que $M \epsilon [a_{i-1}, a_i[$ alors le point $\left(M, \frac{1}{2}\right)$ appartient à la droite qui passe par les points (a_{i-1}, F_{i-1}), (a_i, F_i).

b-Le mode :

*Cas discret : C'est la valeur de X pour laquelle on a la plus grande fréquence.

*Cas continue : La classe modale est la classe des valeurs de X pour la quelle on a la plus grande fréquence et le mode est le centre de la classe modale.

c-La moyenne :

*Cas discret : La moyenne de X est notée $\bar{X}$ et elle est donnée par :
$\bar{X} = \frac{1}{N}(n_1x_1 + n_2x_2 + \cdots + n_mx_m) = \frac{1}{N}\sum_{i=1}^{m} n_ix_i$

Proposition : $\bar{X} = f_1x_1 + f_2x_2 + \cdots + f_mx_m = \sum_{i=1}^{m} f_ix_i$

*Cas continue : $\bar{X} = \frac{1}{N}\sum_{i=1}^{m} n_ic_i = \sum_{i=1}^{m} f_ic_i$ où c_i est le centre de la classe $[a_{i-1}, a_i[$

d-La variance :

*Cas discret : La variance de X est notée $V(X)$ et elle est donnée par :
$V(X) = \frac{1}{N}(n_1x_1^2 + n_2x_2^2 + \cdots + n_mx_m^2) - \bar{X}^2 = V(X) = \frac{1}{N}\sum_{i=1}^{m} n_ix_i^2 - \overline{X}^2$

Proposition : $V(X) = (f_1x_1^2 + f_2x_2^2 + \cdots + f_mx_m^2) - \bar{X}^2 = V(X) = \sum_{i=1}^{m} f_ix_i^2 - \overline{X}^2$

*Cas continue : $V(X) = \frac{1}{N}\sum_{i=1}^{m} n_ic_i^2 - \overline{X}^2 = \sum_{i=1}^{m} f_ic_i^2 - \overline{X}^2$ où c_i est le centre de la classe $[a_{i-1}, a_i[$

e-L'écart type :

L'écart type de X est notée σ_X et il est donné par : $\sigma_X = \sqrt{V(X)}$