STATISTIQUE DESCRIPTIVE

CHAPITRE 1
ETUDE UNIVARIEE

I. GENERALITES ET DEFINITIONS

- On appelle statistiques (au pluriel) ou série statistique des collections de nombres présentées sous formes de tableaux ou de graphiques.
- La statistique (au singulier) est l'ensemble des méthodes scientifiques à partir desquelles on organise, présente et analyse les données numériques et qui permettent de tirer des conclusions et de prendre des décisions judicieuses.
- On appelle population l'ensemble d'individus que ce soient des personnes, des animaux, des plantes ou des objets.

Une étude statistique porte généralement sur un caractère (c'est une « facette » que présente un individu) déterminé présenté par chacun des individus d'une population donnée. Exemple : étude statistique portant sur le poids de nouveau-nés, la taille, le taux de cholestérol, le taux d'urée sanguine de personnes d'une population.

Remarque : Parfois on emploie le terme de variable statistique au lieu de caractère.

L'étude statistique d'un phénomène doit comporter les quatre étapes suivantes :

1. Le recueil des données

On considère une population de laquelle on veut faire une étude statistique portant sur un caractère présenté par chacun des individus. Il est généralement impossible de faire des observations sur chaque individu de la population soit à cause de l'effectif qui est trop grand, soit parce qu'elle est destructive (contrôle de qualité d'un produit) ; on devra choisir une partie composée de n individus appelée échantillon de taille n. Le problème important en statistique consiste avant tout en le choix de l'échantillon. La méthode de choix de l'échantillon la plus fréquente est appelée méthode des sondages ; elle consiste à choisir au hasard un échantillon de la population. L'expression « Hasard » signifie qu'il n'y a aucune raison pour qu'un individu soit choisi de préférence à un autre c'est à dire que chaque individu de la population a la même probabilité d'être choisi. Cette méthode vise à réaliser un échantillon représentatif de la population : les informations obtenues à partir des observations faites sur l'échantillon doivent pouvoir être étendues, sans erreur grave, à l'ensemble de la population.

En mathématiques, un caractère est une application définie sur l'ensemble de la population à valeurs dans un ensemble appelé ensemble des modalités du caractère.

Un caractère peut donc présenter plusieurs modalités :

Exemple

- Le caractère groupe sanguin a des modalités : A, B, AB et O.
- Le caractère sexe a deux modalités : masculin et féminin.

- Le caractère taille des personnes a plusieurs modalités : 165 cm, 157 cm, 173 cm etc.

Parmi les caractères étudiés, on distingue :

- **Caractère quantitatif** : Un caractère est quantitatif si ses modalités sont mesurables.
  Exemple : Taille, Poids, Taux d'urée, Taux de cholestérol, Nombre d'enfants.
  Un caractère quantitatif peut être :
  - **Discret**, si les mesures du caractère sont discrètes, c'est-à-dire $C = \{x_1, x_2, \cdots, x_n\}$.
    Exemple : Nombre d'enfants d'une famille.
  - **Continu**, si les mesures du caractère peuvent prendre n'importe quelle valeur entre des limites données, c'est-à-dire $C = [a, b]$.
    Exemple : Taille, Poids, Taux de cholestérol.
- **Caractère qualitatif** : Un caractère est qualitatif si ses modalités ne sont pas mesurables.
  Exemple : Groupe sanguin, Sexe.

2. La présentation des données

Les données recueillies doivent être présentées sous forme de tableaux ou de graphiques et quelques fois cette présentation donne une idée suffisante de l'information contenue dans ces données.

3. L'analyse des données

Cette étape fondamentale consiste à obtenir des informations concernant le caractère étudié dans la population à partir de celles obtenues sur l'échantillon en utilisant les méthodes du calcul de probabilités appelées méthodes statistiques.

4. La fiabilité des résultats

Il s'agit de préciser le degré de confiance qu'il faut accorder aux résultats obtenus par l'analyse des données en fonction des données observées.

II. PRESENTATION NUMERIQUE ET GRAPHIQUE

Une étude statistique portant sur un caractère quelconque présenté par les individus d'un échantillon de taille n est appelée étude statistique univariée. Dans cette partie, nous allons apprendre comment on représente, pour chaque nature du caractère, numériquement et graphiquement une série statistique simple.

1. Caractère quantitatif discret

Soit la série statistique $x_1, x_2, \ldots, x_n$ où x_i est la valeur du caractère quantitatif discret présentée par l'individu i avec $i = 1,2, \ldots, n$.

L'effectif total : c'est le nombre n de valeurs x_i.

L'étendue de la série : c'est la différence entre la valeur maximale et la valeur minimale.

$$\boldsymbol{E = x_{max} - x_{min}}$$

Une série statistique sur un caractère quantitatif discret est très souvent représentée sous la forme d'un tableau :

Valeur du caractère	x_1	x_2	…	x_k
Effectif	n_1	n_2	…	n_k

L'effectif n_i de x_i : c'est le nombre d'individus présentant la valeur du caractère x_i.

L'effectif total n : c'est la somme des effectifs des classes.

$$\boldsymbol{n = \sum_{i=1}^{k} n_i}$$

La fréquence f_i de x_i : c'est le rapport de l'effectif de x_i à l'effectif total n.

$$\boldsymbol{f_i = \frac{n_i}{n}}$$

Exemple A

Dans une région donnée, on étudie le nombre d'enfants par famille. On choisit au hasard un échantillon de taille 100 et on fait les observations suivantes :

Nombre d'enfants (x_i)	Effectif (n_i)	Fréquence $\left(f_i = \frac{n_i}{n}\right)$
0	5	0,05
1	15	0,15
2	25	0,25
3	20	0,20
4	15	0,15
5	13	0,13
6 ou plus	7	0,07
TOTAL	**100**	**1**

Considérons la série statistique suivante :

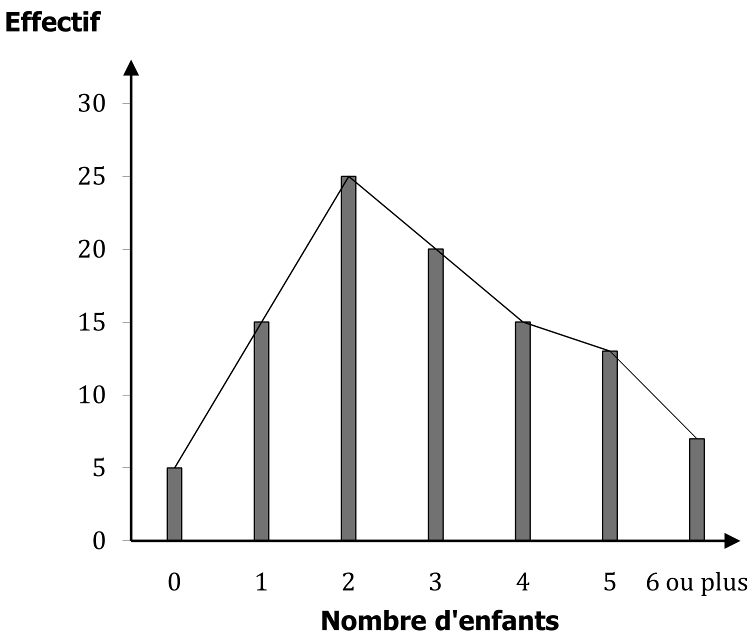
Valeur du caractère	x_1	x_2	…	x_k
Effectif	n_1	n_2	…	n_k

La représentation graphique de ce tableau peut être sous deux formes :

- **Diagramme en bâtons des effectifs :** c'est un ensemble de bâtons ayant pour abscisses les valeurs du caractère x_i et pour hauteurs les effectifs n_i.
- **Polygone des effectifs :** c'est la ligne brisée joignant les extrémités des bâtons.

Exemple : Représenter graphiquement la série statistique de l'exemple A.

Diagramme en bâtons et polygone des effectifs



Remarque

On peut tracer le diagramme en bâtons et le polygone des fréquences en portant en ordonné les fréquences f_i.

Effectif cumulé n_i^c jusqu'à la $i^{\text{ème}}$ valeur x_i : c'est la somme de l'effectif de x_i et de tous les effectifs des valeurs qui précèdent x_i.

$$\boldsymbol{n_i^c = n_1 + n_2 + \cdots + n_i}$$

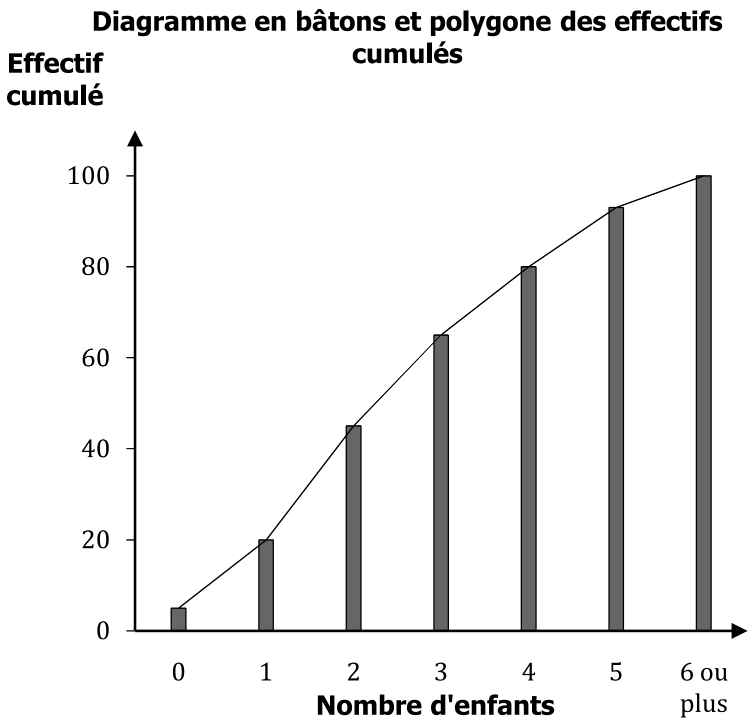
Fréquence cumulée f_i^c jusqu'à la $i^{\text{ème}}$ valeur x_i : c'est la somme de la fréquence de x_i et de toutes les fréquences des valeurs qui précèdent x_i.

$$\boldsymbol{f_i^c = f_1 + f_2 + \cdots + f_i}$$

Exemple

Nombre d'enfants (x_i)	**Effectif** (n_i)	**Effectif cumulé** (n_i^c)	**Fréquence** (f_i)	**Fréquence cumulée** (f_i^c)
0	5	5	0,05	0,05
1	15	20	0,15	0,20
2	25	45	0,25	0,45
3	20	65	0,20	0,65
4	15	80	0,15	0,80
5	13	93	0,13	0,93
6 ou plus	7	100	0,07	1,00

De la même façon, on établit le diagramme en bâtons et le polygone des effectifs cumulés (ou des fréquences cumulées).



2. Caractère quantitatif continu

Lorsque l'étude statistique porte sur un caractère quantitatif continu la représentation précédente n'est plus possible parce que, entre deux valeurs du caractère, on trouve toujours une infinité de valeurs ; par conséquent, si l'échantillon est de taille assez élevée, plusieurs valeurs sont extrêmement proches. La représentation précédente entraînerait alors une grande dispersion des effectifs et ne permettrait pas de suivre les variations du caractère dans l'échantillon. Pour cela, il est d'usage de répartir son étalement en différentes classes disjointes limitées chacune par une borne inférieure et une borne supérieure. La différence entre ces deux limites s'appelle amplitude de la classe, et dans la majorité des cas les amplitudes sont égales. On procède de la façon suivante à l'aide d'un exemple.

Exemple : La pesée de 37 nouveau-nés a donné les résultats suivants (exprimés en Kg) :

2,00 2,05 2,07 2,11 … 4,93 4,94 4,94 4,95 4,95

On partage la série statistique en N classes de même amplitude k en procédant de la façon suivante :

1. On calcule l'étendue E :

$$E = x_{max} - x_{min} = 4{,}95 - 2{,}00 = 2{,}95$$

2. On détermine l'amplitude des classes k sachant que :

$$E = N.k$$

Le nombre de classes N est généralement pris égal à la valeur approchée de $\sqrt{n}$ où n est la taille de l'échantillon (dans notre exemple $n = 37$). On prendra par exemple $N = 6$.
D'où :

$$k = \frac{E}{N} = \frac{2{,}95}{6} = 0{,}49 \approx 0{,}50$$

3. On partage la série statistique en 6 classes de même amplitude 0,50 ensuite on attribue les effectifs n_i à chaque classe.

Classe	**Effectif**
[2,00 , 2,50[	n_1
[2,50 , 3,00[	n_2
[3,00 , 3,50[	n_3
[3,50 , 4,00[	n_4
[4,00 , 4,50[	n_5
[4,50 , 5,00[	n_6
TOTAL	n

D'une façon plus générale, on obtient la série statistique présentée sous la forme du tableau suivant :

Classe	**Centre de classe**	**Effectif**
$[a_0, a_1[$	x_1	n_1
$[a_1, a_2[$	x_2	n_2
...	...	...
$[a_{k-1}, a_k[$	x_k	n_k
Total		n

a_{i-1} et a_i sont les extrémités de la classe $[a_{i-1}, a_i[$ (la valeur extrême droite a_i n'appartient pas à la classe $[a_{i-1}, a_i[$).
n_i est l'effectif de la classe $[a_{i-1}, a_i[$.
x_i représente le centre de la classe $[a_{i-1}, a_i[$.

$$x_i = \frac{a_{i-1} + a_i}{2}$$

n est l'effectif total.

$$n = \sum_{i=1}^{k} n_i$$

La représentation graphique de ce tableau peut être sous deux formes :

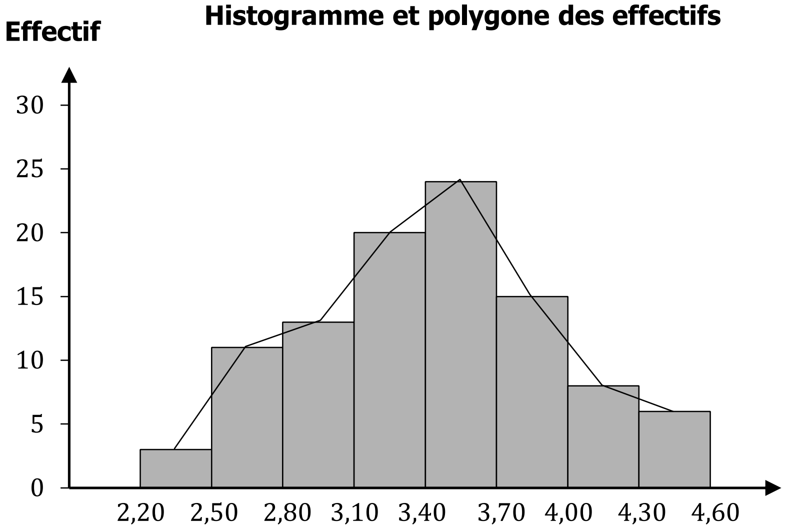
- **L'histogramme des effectifs (ou des fréquences)** est l'ensemble des rectangles ayant pour largeur l'amplitude de la classe et pour hauteur l'effectif (ou la fréquence) de la classe.
- **Le polygone des effectifs (ou des fréquences)** est la ligne brisée joignant les milieux des bases supérieures des différents rectangles adjacents.

Exemple B

On effectue l'opération de pesage sur un échantillon de 100 nouveau-nés. Les valeurs des poids exprimées en Kg et réparties en 8 classes figurent dans le tableau suivant :

Classe des poids	Effectif (n_i^c)	Centre de classe (x_i)	Fréquence (f_i)
[2,20 , 2,50[	3	2,35	0,03
[2,50 , 2,80[	11	2,65	0,11
[2,80 , 3,10[	13	2,95	0,13
[3,10 , 3,40[	20	3,25	0,20
[3,40 , 3,70[	24	3,55	0,24
[3,70 , 4,00[	15	3,85	0,15
[4,00 , 4,30[	8	4,15	0,08
[4,30 , 4,60[	6	4,45	0,06
TOTAL	100		1

Histogramme et polygone des effectifs

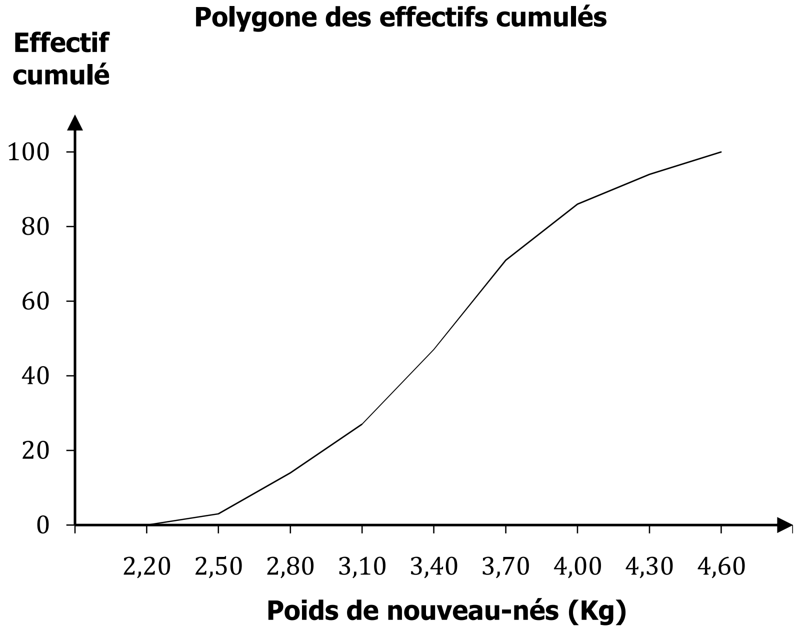


La représentation graphique peut être aussi le polygone des effectifs cumulés : c'est la ligne brisée joignant les points (a_i, n_i^c) où a_i est la valeur extrême droite de la classe $[a_{i-1}, a_i[$ et n_i^c est l'effectif cumulé jusqu'à a_i.

Exemple : Tracer le polygone des effectifs cumulés de l'exemple B.

Classe des poids	Effectif (n_i)	Centre de classe (x_i)	Effectif cumulé (n_i^c)
[2,20 , 2,50[	3	2,35	3
[2,50 , 2,80[	11	2,65	14
[2,80 , 3,10[	13	2,95	27
[3,10 , 3,40[	20	3,25	47
[3,40 , 3,70[	24	3,55	71
[3,70 , 4,00[	15	3,85	86
[4,00 , 4,30[	8	4,15	94
[4,30 , 4,60[	6	4,45	100

Polygone des effectifs cumulés



Remarque : De la même façon, on peut tracer le polygone des fréquences cumulées en portant en ordonnée les fréquences cumulées.

3. Caractère qualitatif

Il n'est plus alors possible d'utiliser un diagramme cartésien puisque les modalités ne sont pas mesurables. Diverses méthodes sont possibles ; nous indiquerons deux d'entre elles à partir de l'exemple suivant :

Exemple C

Pour étudier les réactions d'enfants à un vaccin, on considère un échantillon de 100 de ces enfants et, pour chacun d'eux, sa réaction au vaccin. On a le tableau suivant :