Formation Sociale

<u>Parcours</u> : BREVET DE TECHNICIEN EN TRAVAIL SOCIAL

DIPLOME D'ETAT D'AGENT DE PROMOTION SOCIALE

Diplôme de niveau IV
selon la classification internationale des diplômes
établie par l'UNESCO

2ème Année 2020 - 2021

Cours

STATISTIQUES

LA STATISTIQUE DESCRIPTIVE

Chapitre 1 : Notions de base

I- Introduction

D'une manière générale, les statistiques sont définies comme un ensemble de données quantitatives et qualitatives d'observations collectées, élaborées grâce à des méthodes et outils scientifiques aidant à la prise des décisions dans des domaines économique, social et politique.
Il est universellement admis que seules de bonnes statistiques peuvent permettre d'élaborer de bonnes politiques et obtenir de meilleurs résultats dans le domaine du développement économique et social. Il faut aussi souligner l'importante place qu'occupent actuellement les statistiques au cœur de la problématique du développement, de la lutte contre la pauvreté et des objectifs du millénaire pour le développement (OMT).

II- Notion de statistiques

Cette notion peut s'énoncer de trois (3) façons différentes (les, une, la) correspondant à trois (3) réalités distinctes quoi que cependant liées :

2-1- Les statistiques

C'est une collection de données numériques et systématiques sur un sujet donné. Elles sont élaborées périodiquement depuis la haute antiquité. Le mot “STATUS” signifiant à la fois Etat puissance publique et Etat situation. Elles correspondent le plus souvent avec le recensement des personnes et des biens à une attitude empirique basée sur la connaissance préalable en vue de l'action directe et immédiate (lever l'impôt ou une armée).
A titre d'exemple, les statistiques togolaises du commerce extérieur appréhendent les flux d'importation et d'exportation (en volume et en valeur) intéressant les relations commerciales entre le Togo et le reste du monde.

2-2- Une statistique

Une statistique ou une série statistique peut être “fabriquée” à partir des matériaux de base que constituent les statistiques présentée sous forme d'un tableau statistique. Chaque série statistique ainsi individualisée doit être considérée comme un “être nouveau” qu'il s'agira ensuite d'analyser et d'interpréter à travers les résultats obtenus. Ainsi, grâce aux statistiques de l'exemple précédent, il sera possible de dresser un tableau statistique qui rende compte de l'évolution des importations du nombre des voitures japonaises au cours de la dernière décennie.

2-3- La statistique

La statistique est précisément la méthode d'analyse qui s'applique à chaque série statistique. Le champ d'application de la statistique est universel puisque intéressent toutes les sciences (agronomie, biologie, météorologie, médecine...) et toutes les activités : production, gestion, contrôle de fabrication marketing...

Définition

La statistique est une science qui a pour objet l'étude, l'analyse et l'interprétation des observations relatives à des phénomènes collectifs.

III- Vocabulaire succinct de la statistique

Chaque série statistique est identifiée par un titre qui renferme obligatoirement trois (3) termes fondamentaux : population constituée d'unités, caractère et effectif).

3-1- Population et unité statistique

L'enquête statistique qu'elle soit exhaustive ou partielle, porte sur une population ou un ensemble constitué d'un nombre plus ou moins élevé d'unité statistique (objet de l'observation). Ainsi, en se référant à une enquête menée sur une population de 80 ménages (cadre moyen) sous la forme d'un questionnaire simplifié (nombre d'enfants, revenu mensuel), chacun des ménages ou unités statistiques répondra à ces deux (2) questions. Il sera alors possible par le dépouillement des questionnaires remplis (constituant les statistiques) de dresser deux (2) séries statistiques distinctes ; (avec 10 questions on aurait obtenu 10 séries différentes).

3-2- Le caractère statistique

Il correspond précisément à chaque question. Plus généralement, il est présenté par chaque unité statistique avec une intensité variable mesurée par un nombre qui est la valeur du caractère si par exemple, le 8^{e} ménage interrogé pour la question nombre d'enfants répond 3, nous retiendrons que pour ce ménage, le caractère statistique est appelé variable et se note ($\boldsymbol{x}$).

3-3- Effectif statistique

Il est bien évident que parmi les 80 ménages interrogés, chaque valeur du caractère (0, 1, 2, 3, 4... enfants) pourra se rencontrer un certain nombre de fois. Cette répétition de la même valeur du caractère statistique s'appelle : effectif et se note $\boldsymbol{n}$ (nombre de fois). On vérifiera que la somme des effectifs constatés est égal à l'effectif total qui se note **N** et dans notre cas, **N = 80**.

Pour conclure, nous observons que les valeurs successives de la variable ***x*** correspondantes aux effectifs constatés ***n*** figurent dans deux (2) colonnes distinctes constituant un tableau statistique ayant pour titre “répartition d’une population de 80 ménages selon le nombre d’enfants” et dans lequel nous retrouvons l’unité statistique (ménage), le caractère statistique (nombre d’enfants) ainsi que l’effectif total (80).

IV- Elaboration des statistiques

Elle se déroule en trois (3) étapes :

- Première phase ou phase préliminaire
  Dans cette phase, on procède à :
  - ∗ La définition des faits élémentaires à observer (unité statistique)
  - ∗ La limitation du champ d’investigation (ensemble statistique)

- Deuxième phase : collecte des renseignements (enquête orale ou écrite)
  Ici on définit les instruments de l’enquête et la modalité de l’enquête.

1- Instrument de l’enquête (enquêteur et questionnaire)

∗ Les enquêteurs doivent posséder des qualités requises (connaissances techniques, connaissances professionnelles, bonne psychologie). Par ailleurs, ils doivent être recrutés, encadrés et formés spécialement.

∗ Le questionnaire se caractérise par
  - Le choix des questions (nombre limité, adaptation à l’enquête)
  - La formulation des questions (clarté, précision, facilité de compréhension, interprétation unique et immédiat, objectivité)
  - La présentation (note préliminaire, ordre logique, aspect agréable)

2- Modalité de l’enquête (recensement, sondage)

a- Enquête exhaustive : le recensement

C’est une collecte exhaustive de l’information induisant un coût élevé.

b- Enquête partielle : sondage

C’est une collecte partielle de l’information posant le problème de la représentativité de l’échantillon.

Echantillon : lorsque la population statistique est trop nombreuse, on étudie un sous-ensemble de la population appelé échantillon sur lequel on détermine l’observation des faits.

- Le dépouillement des questionnaires

Dans cette phase, on procède à la :

- Vérification des documents de base (vraisemblance des réponses)
- Définition d'un certain nombre de groupe et de classe en vue de l'établissement du tableau statistique correspondant. La phase technique du dépouillement peut être réalisée manuellement ou par des machines électroniques plus ou moins sophistiquées.

a- Notation utilisée

Une série statistique à une seule variable ***x*** se présente sous la forme d'un tableau à deux (2) colonnes (à simple entrée). Pour chaque ligne du tableau, à chaque valeur du caractère ***x*** correspond un effectif ou répétition en ***n***. Il est donc indispensable de repérer les différentes lignes par la lettre « ***i*** » comme indice.

La notation finalement retenue sera donc pour la variable : x_i (***x*** indice ***i***) et pour l'effectif n_i (***n*** indice ***i***) ainsi, pour la troisième ligne du tableau ci-après : x_3 (***x*** indice **3**) correspondant n_3 (***n*** indice ***3***) c'est-à-dire à l'intensité du caractère 2 enfants correspond 20 manages. En d'autres termes, on dira qu'on rencontre 20 fois deux enfants dans la population des 80 ménages.

b- Présentation d'une série statistique avec effectif

Après dépouillement des 80 réponses du questionnaire relatif au nombre d'enfants (les valeurs du caractère étant placées selon un ordre croissant), on obtient une série statistique faisant l'objet d'un tableau à deux (2) colonnes identifiées par un titre.

3	1	0	2	3	3	2	2	3	1
0	1	3	2	0	4	0	4	2	1
0	1	5	2	3	1	0	4	5	2
3	2	2	4	0	2	3	0	4	1
2	2	1	3	1	2	3	0	1	2
1	0	1	5	5	1	1	1	1	3
3	1	2	1	2	4	1	2	0	4
0	1	2	1	1	1	4	2	2	1

Nombre d'enfants x_i	Nombre de ménages	Interprétation	Lecture du tableau statistique
$x_1 = 0$		n1= 12	12 ménages n'a pas d'enfants
$x_2 = 1$		n2= 24	24 ménages ont 1 enfant
$x_3 = 2$		n3= 20	20 ménages ont 2 enfants
$x_4 = 3$		n4= 12	12 ménages ont 3 enfants
$x_5 = 4$		n5 = 8	8 ménages ont 4 enfants
$x_6 = 5$		n6 = 4	4 ménages ont 5 enfants

Remarque 1

Le vocabulaire utilisé en statistique est riche de notions équivalentes. Ainsi, peut-on utiliser pour désigner la population : ensemble, collectivité, masse ou univers statistique.

Remarque 2

Lorsqu'on se trouve en présence d'un tableau statistique qui ne précise pas en tête des colonnes, la variable et l'effectif, le titre permet de lever toute ambiguïté.

Exemple : 1^{ere} colonne (nombre d'entreprises)

2^{e} colonne (nombre de salaire)

Titre : ventilation d'une collectivité de cent (100) entreprises selon le nombre de salarié.

V- Effectif et fréquence

Il est souvent intéressant de transformer les effectifs $\boldsymbol{n_i}$ en fréquence correspondante notée ($\boldsymbol{f_i}$) car le calcul statistique s'en trouve simplifié (la somme des fréquences étant toujours égale à l'unité ($\boldsymbol{f_i}$) = 1). La transformation très simple repose sur l'application de la formule : $\boldsymbol{fi = \frac{ni}{N}}$ avec N = $\sum_{ni}$ = n_1 + n_2 + ... n_n.

Nombre d'enfants x_i	Effectif (nombre de ménages)	Fréquence $fi = \frac{ni}{N}$	%
$x_1 = 0$	12	$fi = \frac{12}{80} = 0{,}15$	15%
$x_2 = 1$	24	$fi = \frac{24}{80} = 0{,}30$	30%
$x_3 = 2$	20	$fi = \frac{20}{80} = 0{,}25$	25%
$x_4 = 3$	12	$fi = \frac{12}{80} = 0{,}15$	15%
$x_5 = 4$	8	$fi = \frac{8}{80} = 0{,}10$	10%
$x_6 = 5$	4	$fi = \frac{4}{80} = 0{,}05$	5%

NB : le calcul statistique pourra donc porter indifféremment sur le tableau initial ou sur le tableau transformé. Toutefois, il est préférable d'utiliser les pourcentages. Les effectifs n_i sont également appelés fréquences absolues alors que les f_i sont les fréquences relatives. C'est pourquoi les séries avec effectif sont également désignées comme des distributions de fréquence.

Remarque 1 : lorsque les effectifs sont faibles ou nuls pour les valeurs élevées de la variable, on simplifie le tableau par regroupement. Exemple : si deux (2) ménages ont six (6) enfants, 0 ménages a sept (7) enfants et un (1) ménage a huit (8) enfants, la dernière ligne du tableau s'écrivait : plus de six (6) enfants trois (3) ménages.

VI- Description d'une série statistique

A- Introduction

Une série statistique est constituée par des valeurs numériques résultant de l'observation d'une population. Une population est l'ensemble sur lequel on effectue l'étude statistique.
Exemple de population : les élèves de l'ENFS.

Les éléments d'une population sont appelés unité statistique. Une population donnée peut présenter un trait commun : est le caractère.
Exemple de caractère : la taille, le poids des élèves.

Lorsque les paramètres d'appréciation ne sont pas mesurables, on parle de caractère qualitatif.
Exemple de caractère qualitatif : le sexe, la nationalité, la religion, le teint.

Le caractère est quantitatif si les paramètres d'appréciation sont mesurables.
Exemple de caractère quantitatif : la taille, le poids, l'âge.

1- Le dépouillement

Le tableau statistique suivant donne les notes de cent (100) élèves à la suite d'une évaluation écrite :

* Les notes attribuées aux élèves constituent une série statistique, la note est variable. On la note x_i.
* Le nombre de fois que la note x_i apparait dans la série statistique devient l'effectif. On le note n_i.

Remarque 1 : lorsque les variables sont isolées c'est-à-dire représentées par des entiers naturels, on parle de variable discrète.

Remarque 2 : si les variables sont regroupées dans des classes, on parle alors de variable continue.
Regroupons la série statistique en variable avec effectif correspondant.

x_i	n_i	x_i	n_i
1	2	11	10
2	1	12	8
3	1	13	8
4	2	14	4
5	4	15	6
6	6	16	3
7	7	17	3
8	9	18	4
9	9	19	1
10	12	20	0

Classes	n_i
[0 ; 4 [	2+1+1= 4
[4 ; 8 [	2+4+6+7 = 19
[8 ; 12 [	9+9+12+10 = 40
[12-16 [	8+8+4+6 = 26
[16-20 [	3+3+4+1 = 11

Remarque 3 : les deux (2) variables précédentes continues et discontinues sont quantitatives (puisque l'intensité du caractère peut être mesurée). Il existe également des variables qualitatives : sexe, nationalité, religion.
Remarque 4 : deux (2) variables peuvent intervenir simultanément pour une même série : le tableau statistique est alors à double entrée. Exemple : cas d'une population de conscrits (les recrus) repartis selon l'âge et le poids.

B- Type de variable

Le dépouillement de l'enquête sur les 80 ménages permet d'isoler deux (2) séries statistiques puisque deux (2) questions seulement sont posées :

- La première à partir du nombre d'enfants
- La seconde en fonction du revenu perçu par ménage.

La variable "nombre d'enfants" prend des valeurs isolées bien déterminées et peu nombreuses (0, 1, 2, 3, 4, 5) : nous dirons que dans ce cas x est une valeur discontinue ; il y a bien rupture dans un ménage lorsqu'on passe de 0 à 1 enfant, puis de 1 à 2. Cependant, il faut bien noter que la "discontinuité" s'atténue lorsque l'intensité du caractère statistique croit.
Exemple : dans les entreprises artisanales, il y a moins de 10 salariés, ici le nombre d'ouvriers correspond à une variable discrète. Par contre, dans les groupes industriels employant des milieux de travailleurs, la variable devient continue car l'embauche d'un ouvrier supplémentaire n'a pas d'incidence

significative sur la masse salariale ou le niveau de la production. Par contre, la "variable niveau de revenu" peut prendre 80 valeurs différentes (une par ménage) car les numérotations sont déterminées au franc près (différents éléments constitutifs entrant en jeu : salarié de base, compléments familiaux, prime et indemnité diverses). Or la différence de 1 ou quelques francs entre deux (2) revenus voisins n'est pas significative : x dans ce cas est considéré alors comme une variable continue (pouvant prendre théoriquement une infinité de valeur). Si on prenait en compte systématiquement les 80 revenus distribués, le tableau statistique qui en résulterait comprendrait 80 lignes, il est évident que le calcul statistique deviendrait rapidement fastidieux, lourd et presque impossible pour les populations nombreuses.

D'où la nécessité de transformer la série initiale des 80 valeurs ponctuelles de x_i par leur regroupement dans des tranches de revenu plus ou moins large (appelé places).

1- L'utilisation des classes

a- La classe

La classe est un intervalle (ou tranche de variable) qui regroupe plusieurs valeurs ponctuelles initiales de la variable. La simplification du tableau statistique entraine cependant deux (2) inconvénients (qui s'aggravent avec l'importance croissante de l'intervalle) :

- Perte d'une partie notable de l'information
- Possibilité de risque d'erreurs lorsque les valeurs de x_i se distribuent inégalement. En effet, plus le tableau est simplifié (nombre réduit de classe de grande amplitude) et plus on s'éloigne des données initiales.

b- Les trois éléments de la classe

A partir d'une tranche de revenus (30 000 à 60 000), nous noterons que cette classe comprend :

- Une borne inférieure (30 000)
- Une borne supérieure (60 000)

Ces bornes sont les limites de classe.

- Un centre de classe qui est égal à la moyenne arithmétique des deux (2) bornes.
- Une amplitude de classe notée a qui est la différence entre les valeurs des deux (2) limites.

2- Les séries statistiques

Nous choisirons pour appliquer la méthode statistique, trois (3) séries volontairement simplifiées pour alléger le calcul. Par ailleurs, nous nous attacherons à leurs compréhensions plutôt qu'à leur complexité.

a- Série N°1 ou série sans effectif

Exemple : notes obtenues par un échantillon de dix (10) élèves (notes attribuées sur 20 points et classées selon un ordre croissant).

$x_i \longrightarrow$ 1, 5, 7, 8, 9, 12, 13, 14, 15, 16

En réalité, toute série comporte un effectif qui est ici implicite et égal à 1 (uniquement pour chaque valeur de la variable) car il existe un titulaire pour chaque note obtenue.

L'effectif uniforme unitaire n'a aucune incidence sur le calcul, c'est pourquoi il n'est pas indiqué dans la série.

b- Série avec effectif : variable discontinue

Exemple : répartition d'une population de 80 ménages selon le nombre d'enfants.

c- Série avec effectif (variable continue avec classe)

On suppose que 50 ménages seulement sur 80 ont répondu à la question "montant du revenu mensuel perçu".

Tranche de revenu (*i*) réelle (FCFA)	simplifiée (10^2F)	Nombre de ménages (n_i)
17 500 à moins de 20 000	175 à 200	2
20 000 à moins de 22 500	200 à 225	8
22 500 à moins de 25 000	225 à 250	18
25 000 à moins de 27 500	250 à 275	11
27 500 à moins de 30 000	275 à 300	7
30 000 à moins de 32 500	300 à 325	4
		50

Remarque 1 : la simplification du tableau initial (relevé des 50 revenus) dépend du choix de l'intervalle de classe : si ce dernier croit, le nombre de classe décroit.

Remarque 2 : certaines séries comportent des intervalles de classe d'amplitudes inégalement croissantes (lorsque les effectifs décroissent rapidement pour les valeurs les plus fortes du caractère : cas du faible nombre de salariés titulaires de hauts revenus).

Chapitre 2 : Utilisation de l'opérateur statistique somme (noté S ou $\sum$)

Le calcul statistique repose essentiellement sur des sommations. Pour la série N°2, (les 80 ménages), la somme des effectifs s'écrira en utilisant la notation n_i. Pour éviter cette écriture trop fastidieuse, on utilise l'opérateur statistique "somme" qui se note S ou encore $\sum$.

$$\sum_{i=4}^{6} = 24$$

Ce qui signifie que l'on fait la somme de tous les effectifs n_i pour les valeurs de i comprises entre 1 et 6 puisqu'il n'existe que six (6) lignes.

Remarque 1 : les sommes figurent implicitement au bas des colonnes des tableaux statistiques. L'indice i indique le rang occupé par le terme dans la suite.
Exemple : U(n + 1) est le (n + 1)eme terme de la suite.
Remarque 2 : la borne < indique le premier terme et la borne > le dernier terme de la suite.
Exemple :

$$\sum_{i=3}^{i=8} = xi$$

Le premier terme c'est x_3 et le dernier terme est x_8. Les deux bornes se trouvent comprises dans la somme.

Conséquence :

$$\sum_{i=3}^{i=8} = xi$$

Compte (8-3+1) = 6 termes.

1- Sommation des termes constants

Soit une suite de termes Ui dont tous les termes sont égaux à une constante a.
On a : $Ui=U_1=U_2$.........Un=a

Si tous les termes sont égaux :

$$\sum_{i=1}^{i=n} a = a + a + a + \cdots \dots \dots + a = na$$

$$\sum_{i=1}^{i=n} a = na$$

Remarque :

Pour les séries avec effectif, le calcul statistique ne fait jamais intervenir isolement la somme des valeurs de x_i. En réalité, chaque valeur x_i se trouve être "pondérée" par l'effectif n_i correspondant. La sommation ne peut alors intéresser que le produit ($x_i n_i$).

3- S ou $\sum$ est un opérateur linéaire

$$\sum_{i=1}^{n=5} 4 = (5 - 1 + 1) = 20$$

Chapitre 3 : La représentation graphique

I- Pourquoi une représentation graphique

La représentation graphique sous forme de graphe ou de diagramme construit géométriquement, donne une image plus ou moins fidèle de l'ensemble des données consignées dans les tableaux statistiques. Elle constitue d'autre part une étape préliminaire, obligatoire et riche d'enseignements de l'analyse statistique.

A- Avantage de la représentation graphique

En effet, un diagramme bien construit présente quatre (4) avantages essentiels :

1- Traduction concrète d'éléments abstraits constituant la série statistique (l'allure générale du diagramme en donne une vue synthétique) ;
2- Contrôle des données initiales (les erreurs, omissions ou anomalies sont très facilement détectés sur le graphe) ;
3- Comparaison entre différentes distributions similaires (le type de liaison entre les phénomènes observés simultanément sur un même graphique est mis en évidence) ;
4- Orientation du staticien quant au choix de la méthode (d'analyse à appliquer et des caractéristiques à déterminer) ;

Encore faut-il que le graphe soit simple (facile à construire), clair (nombre du renseignement limité), équilibré et précis (par un choix judicieux des échelles), bien adapté aux phénomènes représentés ou étudiés.
De fait, les graphes sont très nombreux, mais puisque la méthode statistique concerne essentiellement les séries de fréquence, la représentation graphique sera donc limitée à ses distributions particulières.

B- Inventaire succinct des autres graphes

En général, on distingue :

- Des graphes à coordonnées polaires et triangulaires
- Des graphes logarithmiques et semi logarithmiques
- Des graphes circulaires à secteur ou des graphes carrés.

II- Les diagrammes de fréquence

A- Construction générale des graphes

Le graphique est construit dans un repère orthonormé, c'est-à-dire en utilisant des coordonnées cartésiennes (le plus souvent rectangulaire lorsque les échelles sont différentes sur les deux (2) axes) relatives au premier (1^{er}) cadrant (abscisses et ordonnées positives).

NB : pour des séries de fréquence, le caractère statistique x_i figure sur l'axe horizontal (OX) et les fréquences absolues (n_i) ou relatives (f_i) sont portés sur l'axe vertical (OY).

B- Types de diagramme

Selon le type de variable, les graphes représentant les distributions de fréquence sont de deux (2) types :

- Le diagramme en bâtons (lorsque le caractère est discontinu)
- L'histogramme (lorsque la variable est continue)

NB : les diagrammes sont simples lorsque les fréquences (portées en ordonnées sont simples). Les graphes sont cumulatifs c'est-à-dire lorsque (n_i ou f_i sont cumulés).

III- Principe de construction des graphes

A- Le diagramme en bâton

Sur l'axe des abscisses on positionne régulièrement les différentes valeurs isolées de la variable x. La différence entre deux (2) valeurs successives représente l'unité d'abscisse (relative à l'échelle choisie). Les fréquences figurent sur l'axe des ordonnées (avec une échelle différente ou non). Pour chaque valeur x_i, on trace des bâtons perpendiculairement à l'axe des abscisses et dont la longueur est proportionnelle à la fréquence correspondante : le sommet de chaque bâton a donc pour coordonnées (x_i ; n_i) ou (x_i ; f_i).

B- Le graphique de construction de l'histogramme

Lorsque l'amplitude de classe (a) est constante, la différence entre deux (2) bornes successives (également constante) sera prise comme unité d'abscisse (relative à l'échelle choisie). L'histogramme est constituée par une succession de rectangles accolés ou adjacents, dont les bases (B) sont précisément les intervalles de classe unitaire et dont les hauteurs respectives "H" sont mesurées par les fréquences (absolues ou relatives) correspondant au centre de classe xi (ou aux moyennes arithmétiques des bornes).

NB :

- Si l'axe des ordonnées porte fréquence relative n_i, l'aire de l'histogramme sera mesurée par N = Sn_i
- Si l'axe des ordonnées porte fréquence relative f_i, l'aire de l'histogramme sera mesurée par l'équation $\boldsymbol{Si_{fi} = 1}$.
- Comme $fi = \frac{\mathrm{ni}}{\mathrm{N}}$, l'histogramme (avec n_i) devient l'histogramme (fi) par simple changement d'échelle (sur l'axe des ordonnées).

Application de la série N°3

Répartition d'une population de 50 ménages selon le revenu mensuel.

Revenu mensuel ***i***	Effectif simple		Fréquence cumulée Jusqu'à		Fréquence cumulée Plus de	
	Ni	Fi	ni ↗	fi ↗	ni ↘	fi ↘
[175-200[	2	0,04	2	0,04	50	1
[200-225[	8	0,16	10	0,2	48	0,96
[225-250[	18	0,36	28	0,56	40	0,80
[250-275[	11	0,22	39	0,78	22	0,44
[275-300[	7	0,14	46	0,92	11	0,22
[300-325[	4	0,08	50	1	4	0,08

C- L'histogramme à la courbe de fréquence

1- L'histogramme

La surface de chaque rectangle élémentaire ayant la même base (B = à une unité d'abscisse) et une hauteur H correspondant à une fréquence déterminée (n_i ou f_i) est mesurée par la valeur du produit B × H = 1 × n_i.

B × H = 1 × n_i (pour les fréquences absolues), la surface du premier rectangle est égal à 2.

B × H = 1 × f_i (pour les fréquences relatives, la surface du deuxième rectangle est égal à 0,16.

La surface totale de l'histogramme est mesurée par la somme des fréquences. En d'autres termes, S*ni* = 50

S*fi* = 1

2- Le polygone de fréquence

a- Pour l'histogramme simple

Le polygone de fréquence simple est la ligne brisée qui joint le milieu des paliers supérieurs des rectangles. Aux extrémités de la distribution, on complète le tableau par deux (2) classes fictives (= 150 à 175 et 325 à 350) de fréquence nulle, qui vont permettre de relier le polygone à l'axe des abscisses. Ainsi, la surface comprise entre le polygone et l'axe des abscisses est celle de l'histogramme. En effet, il y a conservation de l'air puisque les triangles "retranchés" compensent exactement les triangles "ajoutés".

b- Pour l'histogramme cumulé

Le polygone de fréquence cumulée est constitué par les diagonales des rectangles "additif" correspondant aux fréquences simples et figurant en pointillés sur le graphe.

3- La courbe de fréquence

Théoriquement, lorsque la taille de la population croit indéfiniment (l'effectif tend vers l'infini) alors que l'intervalle de classe tend vers zéro (le nombre de classe tend également vers l'infini). Les surfaces des rectangles élémentaires se « réduisent à une infinité de bâtons accolés » ; la courbe enveloppante reliant leurs extrémités devient alors la courbe de fréquence. Pratiquement, on peut ajuster le polygone de fréquences empiriquement c'est-à-dire à main levée en respectant la conservation des aires : les surfaces « ajoutées » compensant les surfaces « retranchées ».

Remarque :

1- L'ajustement peut être analytique (recherche de la fonction qui ajuste le mieux le polygone)
2- Pour l'histogramme cumulatif, on assimilera le polygone à la courbe de fréquence.

Les paramètres de position

I- Les paramètres

Dans un premier temps, la représentation graphique est susceptible de rendre compte dans une certaine mesure de la série étudiée. En particulier, l'allure de l'histogramme permet d'apprécier approximativement si la distribution de fréquence est plus ou moins systématique ou plus ou moins étalée.

Encore faut-il que le graphe soit bien construit car "l'impression visuelle" peut être déformée par un choix peu judicieux des échelles. Dans un second temps, il s'agit de décrire plus rigoureusement la série en déterminant certains paramètres ; ces derniers encore appelés caractéristiques, en réduisant les données statistiques, vont résumer la distribution en quelques valeurs types. Cet effort de réduction en quelques valeurs synthétiques doit concilier deux (2) impératifs contradictoires :

- Une présentation simple conduisant à ne retenir qu'un petit nombre de valeur type (dans un souci légitime de simplification).
- Une identification sure de l'ensemble étudié (par le choix des paramètres les plus représentatifs comme la moyenne arithmétique et l'écart type). Pour bien saisir ce double effort de réduction et de représentation, nous ferons appel à l'image suivante : l'élection d'un député a un effet réducteur sur le débat politique et dispense du recours systématique au référendum. Par ailleurs, l'élu est censé représenter au parlement la population de sa circonscription qui l'a désigné.

Avant d'aborder la détermination des paramètres, il convient d'en dresser l'inventaire et de proposer leur classement. Ces paramètres très nombreux peuvent se ranger en trois (3) catégories :

- Les paramètres de position (ou caractéristiques de valeur centrale) : ce sont des valeurs types situées dans la zone centrale de la série et autour desquelles se regroupe la majorité des observations. Ce sont :
  - ∗ Les moyennes de position (médiane et médiale)
  - ∗ La moyenne de fréquence (le mode)
  - ∗ La moyenne de grandeur (la moyenne arithmétique)
- Les paramètres de dispersion (ou de fluctuation) : ils déterminent dans quelle mesure les observations s'écartent des valeurs types précédentes ; ces caractéristiques de dispersion rendent compte globalement de la structure interne de la série. Ce sont :
  - ∗ Les intervalles inter percentiles (ayant pour borne 2 percentiles particulier parmi les 99 recensés)
  - ∗ Les intervalles inter déciles (ayant pour limite 2 déciles choisis parmi les 9)
  - ∗ L'intervalle inter quartile (borné par Q3 et Q1)
  - ∗ L'écart type noté par (sigma) et la variante notée (V)
- Les paramètres de forme complète (l'impression visuelle) : en mesurant précisément les degrés des symétries et d'aplatissement de l'histogramme (représentée par sa courbe de fréquence). Ce sont :
  - Les moments non centrés m_q
  - Les moments centrés u_q

II- Les paramètres de position

Ses caractéristiques de tendance centrale sont en fait des moyennes qui appartiennent à trois (3) types :

- ➢ Les moyennes de position (médiane et médiale)
- ➢ La moyenne de fréquence (mode)
- ➢ La moyenne de grandeur (moyenne arithmétique)

Dans une première approche, nous nous limiterons au calcul du mode et de la médiane pour trois (3) raisons :

- Ses caractéristiques ont une signification concrète (donc facile à appréhender)
- Elles font l'objet d'une détermination graphique (il s'agit d'exploiter les graphes précédents)
- Elles ne sont concernées que par certaines valeurs de la série (le calcul est donc simple et rapide)

La médiale sera abordée avec l'étude de la concentration car cette caractéristique intervient dans la construction et l'interprétation des courbes de concentration. La moyenne arithmétique contrairement aux paramètres précédents n'a pas de

signification concrète et fait intervenir les termes de la série pour sa détermination. Elle sera étudiée avec les moyennes de grandeur dont elle ne constitue qu'un cas particulier.
Parmi ces dernières, nous verrons que la moyenne quadratique est utilisée pour la détermination du paramètre de dispersion le plus intéressant, c'est-à-dire l'écart type.

III- Le mode ou la dominante (Mo)

Le mode est la valeur du caractère X pour laquelle la fréquence (absolue ou relative) est maximum. On dit encore que c'est la valeur la plus fréquente de la série (on peut rapprocher cette notion de celle de la "mode" qui caractérise un comportement majoritaire). Il est évident que le mode ne peut être déterminé que pour les distributions de fréquence.

IV- Médiane ou médiante (Me)

La médiane est la valeur du caractère tel que :

- 50% de l'effectif possède des valeurs x_i < Me et -50% donc l'effectif possède des valeurs x_i > Me.

Ce qui suppose implicitement que les valeurs de la variable x sont ordonnées préalablement d'ordre croissant ou de y. la médiane est donc la valeur de x_i qui partage la série globale en deux (2) séries partielles de même effectif (l'une à caractère "faible", l'autre à caractère "fort"). Cette définition est mise en évidence graphiquement.

V- Détermination du mode

Pour les distributions de fréquences dont le caractère est discontinu, la détermination du mode est instantanée. Le mode est en effet l'abscisse du bâton le plus important.
Pour la série N°2 (nombre d'enfants/ménages), le mode est Mo = 1 enfant puisque la majorité des ménages 24/80 ont précisément un enfant.

VI- Détermination de la médiane

a- Série sans effectif

Soit n le nombre de terme de la série :
Si n est impair (n = 2p + 1), la médiane Me est la valeur (P + 1^{er}) terme. Il existe en effet P termes de part et d'autres de la médiane.
Si n est pair (n = 2p), il existe un intervalle médian dont les limites sont respectivement les valeurs du P^{eme} et P + 1^{eme} terme.

Application

A l'issu d'un devoir sur table, les élèves de la 2e ACI ont obtenus les notes suivantes :

$x_1 \longrightarrow$ 1, 5, 7, 8, 9, 12, 13, 14, 15, 16. Dans ce cas, le nombre de terme étant pair, Me est comprise dans l'intervalle médian, dont les bornes sont les valeurs du 5e terme (9) et la 6e (P + 1eme) (12).

Pour lever l'indétermination, on peut calculer la médiane arithmétique. En supprimant la 1ere note, le nombre de terme devient impair (9 au lieu de 10) et la médiane est parfaitement connue Me = 12. En effet, dans ce cas, quatre (4) notes sont inférieures à 12 (5, 7, 8, 9) et 4 supérieures à 12 (13, 14, 15, 16), conformément à la définition de la médiane.

Remarque 1 :

L'inconvénient majeur de la médiane réside dans le fait que Me dépend de l'ordre des termes et non de leur valeur. Par exemple si le correcteur augmentait de deux (2) points, les quatre (4) notes supérieures, on obtiendrait (15, 16, 17, 18), la médiane ne s'en trouverait nullement affectée.

b- Série avec effectif (variable discontinue)

Dans ce cas, la détermination de la médiane se réalise en trois (3) temps :

1- Cumul des effectifs ou des fréquences
2- Repérage de l'observation médiane en calculant valablement la ½ somme des fréquences
3- Me est la valeur correspondante de x_i prise par l'observation médiane

Application à la série N°2 (nombre d'enfants/ménages)

La détermination de Me est approchée par excès ou par défaut. L'observation médianée est la 40e puisque $\frac{N}{2} = \frac{80}{2} = 40$ aura pour valeur un enfant (par défaut). En effet, la 40e observation est plus proche de la 36e que de la 56e d'où la médiane dans notre cas d'espèce est la variable de 36 Me = 1.

Le cumul dans ce cas est descendant jusqu'à la borne unique comprise.

Lecture des effectifs cumulés

- 12 ménages ont 0 enfant et le 12e ménage n'a pas d'enfant
- 36 ménages ont jusqu'à 1 enfant (soit 0 ou 1 enfant) et le 36e ménage a justement 1 enfant
- 56 ménages ont jusqu'à 2 enfants (soit 0,1 ou 2) et le 56e ménage possède 2 enfants
- 6e ligne : 80 ménages ont jusqu'à 5 enfants (1, 2, 3, 4, 5) et le 80e ménage a précisément jusqu'à 5 enfants.

Remarque : on peut également envisager un cumul ascendant (plus de) la borne unique comprise.
Exemple : 12 ménages ont plus de 4 enfants c'est-à-dire 4 enfants ou 5 enfants.

Propriété de la médiane
La médiane possède deux (2) propriétés :
1- La médiane divise l'histogramme en deux (2) aires égales
2- La somme des écarts estimés en valeur absolue de tous les termes d'une série à la médiane est minimum

VII- Distribution uni modale et pluri modale

En général, les distributions de fréquence ne comportent qu'un seul mode (uni modal). Elles peuvent parfois présenter deux (2) ou trois (3) modes, elles sont alors (bimodales ou tri modales).

a- Cas d'une distribution bimodale

Lorsque la courbe de fréquence enregistre deux (2) modes (absolu et relatif), c'est que la distribution étudiée est la résultante de deux (2) sous population composante homogène.
Remarque : le mode n'a de signification concrète que pour les courbes uni modales. De plus, sa détermination est liée à l'intervalle de classe choisie. Enfin, l'amplitude de classe doit être constante pour toute la distribution. Lors de l'élaboration d'une série statistique, le choix de l'intervalle de classe est donc déterminant pour les valeurs calculées des paramètres de position.

VIII- Détermination du mode par le calcul

Pour les distributions de fréquence uni modale (variable continue avec classe), on repère préalablement la classe modale d'amplitude a et dont les deux (2) bornes sont notées respectivement L_1 = limite < ; L_2 = limite >. La classe modale possède l'effectif (ou la fréquence) le plus élevé. Le calcul du mode repose sur une interpolation consistant à appliquer une des formules suivantes :
$Mo = L_1 + a (d_1/d_1+d_2)$
$Mo = L_2 – a (d_1/d_1+d_2)$

d_1 différence entre les effectifs de la classe modale et de la classe précédente.
d_2 différence entre les effectifs de la classe modale et de la classe suivante

a=200 – 175=25 a=25
classe modale = [225-250[
d_1=18-8=10 d_1=10

d_2=18-11=7 d_2=7

$Mo=L_1+a\ (d_1/d_1+d_2) = 225 + 25\ (10/10+7)$

$= 225 + 25\ (10/17) = 225 + 250/17$

$= 225 + 14{,}7 = 239{,}7$

$Mo=L_2-a\ (d1/d1+d2) = 250-25\ (7/10+7)$

$= 250-175/17 = 250-10{,}30$

$Mo = 239{,}7$

Détermination de la médiane par calcul

Pour les distributions dont les observations sont groupées par classe, le calcul de la médiane repose également sur une interpolation et se réalise en trois (3) temps :

1- Repérage de l'observation médiane en calculant préalablement la demi somme des effectifs absolus (S_{xi}) /2 et la demi somme des effectifs relatifs (S_{fi}) /2.
2- Repérage de l'intervalle dans lequel se situe la médiane sachant que l'amplitude d'intervalle est égale à 1.
3- Détermination de la médiane par interpolation.