Rabia HAJILA

LA STATISTIQUE

LA PARTIE CACHEE DU PATRIMOINE CULTUREL

Cours et exercices corrigés

Tome I

Rabia HAJILA

Enseignant chercheur

A l'Institut National des Sciences

de l'Archéologie et du Patrimoine

- Rabat -

LA STATISTIQUE

LA PARTIE CACHEE DU PATRIMOINE CULTUREL

Cours et exercices corrigés

Tome I

A Selma

A Salmane

Ces cadeaux de Dieu

«Un homme qui meurt est un malheur,

cent hommes qui meurent une catastrophe,
mille hommes qui meurent une statistique,

mais grâce à l'archéologie,

l'histoire de ces hommes

ne meurt jamais»

Edité par Gérard Calot

Présentation de l'ouvrage aux lecteurs

A qui est destiné cet ouvrage ?

Cet ouvrage est destiné essentiellement aux étudiants de l'Institut National des Sciences de l'Archéologie et du Patrimoine, aux chercheurs et aux cadres qui, dans leur activité professionnelle, sont conduits à établir des projets d'étude en archéologie, à les mener ou à les juger, aux enseignants en leur apportant les outils statistiques et les possibilités d'application au patrimoine culturel visant à développer leurs enseignements et leurs recherches, ainsi qu'à tous ceux qui désirent s'initier à la statistique appliquée au patrimoine culturel.

C'est avant tout un ouvrage d'initiation qui vise à l'acquisition de techniques de base, afin de répondre aux difficultés et aux besoins que rencontrent les étudiants pour passer du cours à l'application. L'ouvrage réunit des rappels détaillés de cours visant à familiariser le lecteur avec les notions essentielles, de nombreux exemples d'application, ainsi que des exercices d'application suivis de corrigés succincts, permettant de mettre en pratique et de contrôler les connaissances.

Quels sont ses objectifs ?

Jusqu'à présent, ont été publiés des manuels complets et détaillés de théorie statistique, des travaux de recherche, des ouvrages d'exercices théoriques, ou des ouvrages de vulgarisation pour les économistes, les gestionnaires, les sociologues, les ingénieurs, les universitaires dans des domaines très variés. Le présent ouvrage a des objectifs différents :

- Faire assimiler les techniques de base à des non-spécialistes de la statistique.
- Appliquer ces méthodes aux études archéologiques et aux problèmes du patrimoine culturel.
- Répondre aux besoins de jeunes archéologues marocains et en particulier les préhistoriens.
- Contribuer à l'analyse des données archéologiques recueillies dans le cadre des programmes de recherche qui se déroulaient au Maroc.

Quelles sont les connaissances statistiques requises par les lecteurs ?

Cet ouvrage peut facilement être assimilé par les débutants comme par les étudiants qui n'ont pas une formation mathématique très poussée ou ayant une formation littéraire.

On expose d'abord les grandes lignes de la statistique descriptive, où il s'agit essentiellement de présenter les données sous une forme immédiatement exploitable, en les réduisant à quelques paramètres caractéristiques.

Qu'apporte la statistique à l'archéologie ?

L'analyse descriptive, s'appuie sur l'idée, que les objets archéologiques témoignent des habitudes techniques et du comportement des groupes et des individus qui les fabriquèrent et les utilisèrent et aussi du milieu dans lequel s'inséraient les sociétés humaines. Pour établir les recherches typologiques, il fallait délimiter les traits caractéristiques, les décrire, les regrouper, bref, affronter l'analyse descriptive.

Reconnaissance, fouille et analyse constituent les trois premières étapes de la démarche archéologique que l'archéologue conçoit souvent indépendamment les unes des autres et sans continuité méthodologique explicite.

L'archéologue s'intéresse de plus en plus aux relations entre les sociétés humaines et le milieu naturel, il devient important d'orienter la démarche archéologique vers une plus grande intégration méthodologique de ces trois étapes à l'intérieur d'une approche globale.

Lors d'une reconnaissance, la description d'un site et de son contenu visible peut donc aisément se conformer à un plan quelque peu analytique. Le site est alors défini empiriquement comme « l'espace contenant le regroupement de vestiges le plus distinctement perçu par l'archéologue ».

La fouille permet de décrire les relations et associations observées entre les différents éléments du site : topographie, écologie, structures-témoins, objets-témoins, et autres vestiges. Une fouille idéale vise à l'enregistrement de tous ces éléments et aussi des liaisons que l'on peut observer entre eux, ce qui s'effectue en grande partie par la localisation tridimensionnelle des observations.

L'espace-site défini lors de la reconnaissance peut alors être représenté comme un volume et assimilé « au parallélépipède ayant le maximum de chances de contenir la totalité des vestiges correspondant à un regroupement spatial bien distinct d'autres regroupements voisins ». C'est en cours de fouille ou même seulement d'analyse qu'il est possible de distinguer dans l'espace empirique du site des unités spatiales ayant une signification culturelle : un habitat ou plusieurs habitats juxtaposés ou enchevêtrés par exemple.

La statistique apporte à l'archéologue des possibilités de reconstitution graphique du site, ou plutôt de la disposition des éléments du site les uns par rapport aux autres, sans commune mesure avec ce qu'il pouvait réaliser manuellement. Ces reconstitutions graphiques permettent d'examiner comme par transparence, en totalité ou en partie selon le besoin, le contenu du parallélépipède du site selon n'importe quel angle.

Plan de l'ouvrage :

Cet ouvrage réservé exclusivement à la statistique appliquée au patrimoine culturel fera l'objet d'une série de tomes.

Le 1er tome se limitera essentiellement aux notions de base de la statistique et à la statistique descriptive univariée ou à un seul caractère.

Le 2ème tome sera réservé à la statistique descriptive bivariée ou à deux caractères.

Le 3ème tome traitera l'analyse en composantes principales (ACP) et l'analyse factorielle des correspondances (AFC).

Les tests statistiques, l'ajustement linéaire et la corrélation seront aussi traités dans le 2ème et le 3ème tome.

Les exemples et les exercices proposés dans cet ouvrage se rapportent essentiellement au domaine du patrimoine culturel.

Table des matières

Chapitre I : **Introduction générale**

I- Objet de la statistique :

1) Définition :

La statistique est d'un point de vue théorique une science, une méthode et une technique. Elle permet la collecte des données, l'analyse des données collectées, l'interprétation des données ainsi que la présentation de ces données afin de les rendre lisibles et compréhensibles.

Ainsi la statistique est un domaine des mathématiques qui possède une composante théorique ainsi qu'une composante appliquée. La composante théorique est proche de la théorie des probabilités et forme avec cette dernière, les sciences de l'aléatoire. La statistique appliquée est utilisée dans presque tous les domaines de l'activité humaine.

La statistique appliquée au patrimoine culturel a pour objet la collecte, le traitement et l'interprétation de données d'observation relatives à un groupe d'individus ou d'unités ou d'objets archéologiques dans le but de dégager les caractéristiques ou la répartition de ces objets en fonction de critères d'étude déterminés.

2) Champ d'application de la statistique :

La statistique était présente dès l'antiquité. Les méthodes et outils statistiques sont utilisés dans des domaines très variés tels que :

- En géophysique, pour les prévisions météorologiques, la climatologie, la pollution, les études des rivières et des océans ;
- En démographie : le recensement permet de faire une photographie à un instant donné d'une population et permettra par la suite des sondages dans des échantillons représentatifs ;
- En sciences économiques et sociales, et en économétrie : l'étude du comportement d'un groupe de population ou d'un secteur économique s'appuie sur des statistiques. Les questions environnementales s'appuient également sur des données statistiques ;
- En sociologie : les sources statistiques constituent des matériaux d'enquête, et les méthodes statistiques sont utilisées comme techniques de traitement des données ;
- En marketing : le sondage d'opinion devient un outil pour la décision ou l'investissement ;

- Dans les jeux de hasard et les paris tels que le loto ou les paris équestres, pour "prévoir" les résultats ;
- En physique : l'étude de la mécanique statistique et de la thermodynamique statistique permet de déduire du comportement de particules individuelles un comportement global (passage du microscopique au macroscopique) ;
- En métrologie, pour tout ce qui concerne les systèmes de mesure et les mesures elles-mêmes ;
- En médecine et en psychologie, tant pour le comportement des maladies que leur fréquence ou la validité d'un traitement ou d'un dépistage ;
- En archéologie, appliquée aux vestiges (céramologie, archéozoologie...) ;
- En écologie, pour l'étude des communautés végétales et des écosystèmes ;
- En assurance et en finance (calcul des risques, actuariat, etc.).

3) Etapes de l'enseignement de la statistique :

De manière générale, il existe trois étapes dans l'enseignement de la statistique.

a. La statistique descriptive :

Elle a pour objet de décrire, de dégager l'essentiel et de réaliser des synthèses à l'aide des paramètres caractéristiques et des représentations graphiques.

b. La statistique probabiliste :

Elle a pour objet d'étudier les mécanismes aléatoires à la différence des mécanismes certains qui sont du sort de la statistique descriptive.

c. La statistique mathématique :

Elle permet de prendre des décisions à partir des informations recueillies des mécanismes aussi bien aléatoires que certains.

II- Méthode statistique :

La méthode statistique comporte essentiellement trois phases :

a- Une phase matérielle où il s'agit de rassembler des données, de les regrouper et de les présenter sous forme de tableaux ou graphes ;

b- Une phase analytique qui consiste à réduire les données à un nombre limité de paramètres caractéristiques susceptibles de décrire la série statistique. L'ensemble de ces deux phases constitue l'objet essentiel de la statistique descriptive (ou déductive) dont les résultats restent limités aux échantillons étudiés ;

c- Une phase interprétative, et qui permet de déduire des résultats obtenus sur un échantillon des conclusions relatives à l'ensemble de la population d'où est extrait cet échantillon.

III- Enquête statistique :

Une enquête statistique comporte toujours une phase initiale où il s'agit de collecter des renseignements, suivie d'une phase de dépouillement qui consiste à passer des données brutes à des tableaux ou à des graphes qui se prêtent mieux à l'analyse et l'interprétation.

La manière dont l'enquête est effectuée est évidemment très importante. En particulier, si l'on espère déduire des résultats obtenus sur un échantillon des conclusions relatives à toute la population, il convient de s'assurer que l'échantillon est bien représentatif de cette population.

Chapitre II : Notions de base de la statistique

I-Eléments de vocabulaire statistique :

1) Population ou univers statistique :

C'est l'ensemble d'objets ou de personnes sur lequel doit porter l'étude statistique. Elle peut désigner aussi bien :

- Un ensemble d'êtres vivants.

Exemple :

Les hommes du Djebel Irhoud au Maroc.

- Qu'un ensemble d'objets inanimés.

Exemple :

Les pièces archéologiques et les collections paléontologiques découvertes.

2) Echantillon :

C'est la partie de la population statistique sur laquelle porte l'étude.

Exemple :

La Faune du gisement moustérien du Djebel Irhoud au Maroc.

3) Individu ou unité statistique :

C'est tout objet ou personne sur lesquels porte l'étude.

Exemple :

Chaque mammifère du Djebel Irhoud constitue une unité statistique.

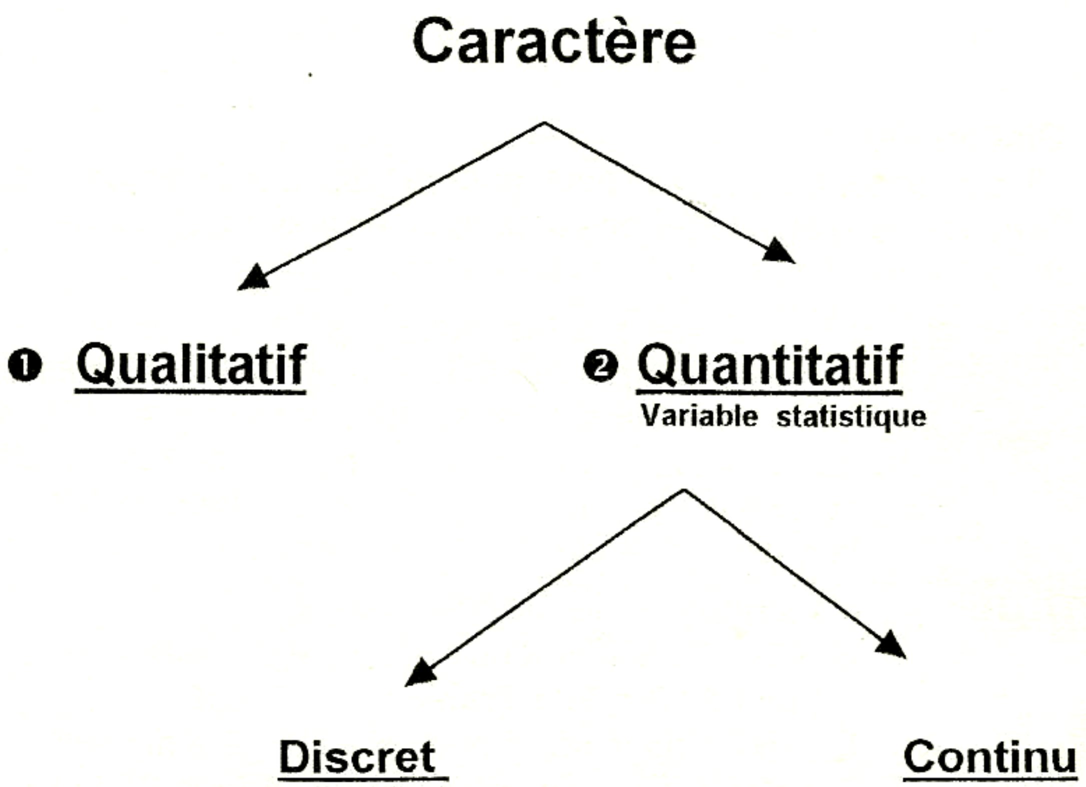
4) Caractère :

C'est l'aspect particulier auquel on s'intéresse, en d'autres termes, c'est le critère que l'on retient pour spécifier la population étudiée.

Exemple :

Les mesures et indices de l'occipital de Temara sur la côte atlantique du Maroc.

Classifications des caractères



On distingue deux grandes catégories de caractères :

a- Caractères qualitatifs :

Un caractère est considéré qualitatif lorsqu'il n'est pas mesurable. C'est-à-dire qu'on ne peut lui associer ni une valeur numérique, ni un ordre naturel.

Exemple :
Les différentes couleurs des silex obtenus dans le site archéologique Ifri n'Ammar dans la région du Rif oriental.

b- Caractères quantitatifs :

Un caractère est considéré quantitatif lorsqu'il est mesurable. C'est-à-dire qu'on peut lui associer une valeur numérique.

Exemple :
L'âge, le poids et la taille des éclats et lames à bord abattu et racloirs trouvés dans la grotte d'Ifri el Baroud dans la région du Rif oriental.

Les caractères quantitatifs peuvent être :

- ✓ Soit discrets (ou discontinus) si les valeurs prises par ces derniers sont des entiers naturels (∈ à ℕ).

Exemple :

Le nombre d'outils soumis à une analyse typologiques dans le site Ifri n'Ammar est de 3182.

- ✓ Soit continus si les valeurs qui sont possibles d'être prises par le caractère sont en nombre infini (∈ à ℝ).

Exemple :

Les dimensions des coquilles perforées du paléolithique moyen d'Ifri n'Ammar.

Par définition, le passage d'une variable qualitative à celle quantitative n'est jamais possible. Cependant, on peut tenter de quantifier certains caractères qualitatifs.

5) Modalités :

Ce sont les différentes situations possibles du caractère étudié.

Exemple :

Les dépôts fins du profil de Sidi Er'hama en Basse Moulouya représentent 4 modalités du caractère taille des sables :

a. Les gravillons à taille supérieur à 2 mm.
b. Les sables grossiers à taille comprise entre 2 et 0,63 mm.
c. Les sables moyens compris entre 0,63 et 0,2 mm.
d. Les sables fins compris entre 0,2 et 0,063 mm.

Les modalités d'un même caractère sont à la fois incompatibles et exhaustives : chaque individu de la population présente une et une seule modalité du caractère envisagé.

6) Effectif n_i ou fréquence absolue :

C'est le nombre de fois que se répète chaque modalité du caractère. L'effectif total de la population N (ou la taille de la population) est la somme de k effectifs n_i (n_1, n_2, …, n_i, …, n_k) correspondants à chacune des modalités de la variable x (x_1, x_2, …, x_i, …, x_k). En effet :

$N = n_1 + n_2 + n_i + \ldots + n_k$

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

7) *Fréquence f_i ou fréquence relative :*

C'est le pourcentage des individus ayant la modalité (i) dans la population étudiée. En d'autres termes la fréquence relative c'est le rapport entre l'effectif n_i et l'effectif total N.

$$f_i = \frac{n_i}{N}$$

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + \ldots + f_k = 1$$

Caractère C	Effectif n	Fréquence f
C_1 C_2 . . . C_k	n_1 n_2 . . . n_k	$f_1=n_1/N$ $f_2=n_2/N$ . . . $f_k=n_k/N$
Total	$N=n_1+n_2+\ldots+n_k$	$f_i=f_1+f_2+\ldots+f_k=1$

Exemple :

Les résultats d'une analyse numismatique des pièces de monnaie sont présentés sous forme d'un tableau réparti selon le lieu de frappe de chaque pièce.

Atelier de frappe	Effectif n_i	Fréquence f_i (en%)
Fès	8	22,22
Andalousie	4	11,11
Tlemcen	3	8,33
Sans indication de l'atelier de frappe	21	58,34
Total	36	100

Interprétation :

Le tableau nous montre que 8 pièces de monnaie sont frappées à Fès et 58,34% des pièces de monnaie étudiées ne présentent aucune indication de l'atelier de frappe.

8) *Série statistique :*

C'est l'ensemble des valeurs du caractère.

Exemple :

Les dépôts alluviaux holocènes du profil Sidi Er'hama en Basse Moulouya.

9) *Sondage :*

Consiste à déterminer un échantillon représentatif, de manière telle que les résultats statistiques trouvés sur cet échantillon soient de ceux que l'on aurait obtenus si l'on avait étudié la population entière.

Exemple :

En 1996, un sondage a été effectué à l'intérieur de la grotte Ifri el Baroud pour évaluer le remplissage du site.

II-Statistique et archéologie :

1) *Définition :*

L'archéologie est une discipline ayant pour objectif la découverte, l'analyse et l'interprétation d'objets ou traces humaines des civilisations anciennes.

L'archéologue ne se contente pas de simplement regarder la forme d'un objet pour décider à quoi il servait (sa fonction). Mais, il commence par le regarder attentivement, le mesurer, mener des recherches et poser des questions à son sujet, par exemple :

- De quelle époque date-t-il ?
- De quoi est-il fait ?
- D'où vient-il ?
- Quelle dimension a-t-il ?
- Quelle forme a-t-il ?

Les réponses aux questions posées par l'archéologue sont fournies par la statistique qui a pour but la collecte, l'analyse et l'interprétation des observations relatives au phénomène étudié.

2) *Exemple de variables qualitatives en archéologie :*

a- Modalités multiples :

Exemple :

Les 5 modalités de céramiques du haut Moyen-âge sont :

- Céramique à pâte rose et à bandeaux cannelés.
- Céramique à pâte beige chamois.
- Céramique à surface lissée.
- Céramique à engobe rouge.
- Céramique modelée.

b- Cas particulier des variables codées en présence/absence :

Exemple :

Le vase campaniforme d'Ain Fouarat se compose du :
- Col : présent/absent.
- Panse : présent/absent.
- Fond : présent/absent.

3) Exemple de variables quantitatives en archéologie :

- Effectifs (*exemple* : nombre d'objets trouvés par site).
- Fréquences (*exemple* : pourcentage d'objets).
- Mensurations (*exemple* : taille des objets).
- Rapports, indices (*exemple* : proportion).

Exemple :

En céramique, à partir des mesures de la hauteur des vases, on peut proposer un classement en trois tailles :

- petit,
- moyen,
- grand.

4) Nature de l'information en archéologie :

a- Les critères intrinsèques :

Les propriétés mêmes de l'objet serviront à l'analyse du corpus.

Exemple :

Sa forme, sa couleur, sa matière.

b- Les critères extrinsèques :

Ils serviront à l'interprétation du résultat du classement.

Exemple :

Le lieu de découverte, la fonction ou la période chronologique (date des événements historiques).

III- Exercices d'application :

Exercice 1 :

Parmi les caractères suivants lesquels sont qualitatifs, quantitatifs discrets ou quantitatifs continus :

- La largeur d'une grotte.
- Le poids d'une pièce de monnaie.
- La forme d'un vase.
- Le nombre d'éclats.
- Le lieu de découverte.
- Le type de mosaïque.
- La structure des pierres.
- Le nombre de grottes.
- Les dimensions d'un outillage.

Réponse :

- Caractère qualitatif :
  - La forme d'un vase.
  - Le lieu de découverte.
  - Le type de mosaïque.
  - La structure des pierres.
- Caractère quantitatif discret :
  - Le nombre d'éclats.
  - Le nombre de grottes.
- Caractère quantitatif continu :
  - La largeur d'une grotte.
  - Le poids d'une pièce de monnaie.
  - Les dimensions d'un outillage

Exercice 2 :

Dans les cas suivants, définir la population étudiée, le caractère étudié et les valeurs possibles du caractère :

1) Le poids des pièces de monnaie almohades en argent.
2) La couleur des pâtes des amphores tingitanes.

Réponse :

1) a. La population étudiée : L'ensemble des pièces de monnaie almohades.

   b. Le caractère étudié : Le poids des pièces de monnaie en argent.

   c. Le type du caractère : Il est quantitatif continu et prend ses valeurs dans un intervalle de l'ensemble $\mathbb{R}$.

2) a. La population étudiée : L'ensemble des amphores tingitanes.

b. Le caractère étudié : La couleur des pâtes des amphores tingitanes.

c. Le type du caractère : Il est qualitatif et prend ses valeurs dans un intervalle de l'ensemble $\mathbb{N}$ des entiers naturels.

Exercice 3 :

A la fin d'un stage à Sijilmassa, le directeur de l'INSAP demande aux étudiants de la 3ème année le degré de satisfaction concernant leur stage. Les réponses ont été ainsi :

TS, TS, S, S, S, MS, MS, S, S, S, PS, PS, TS, MS, S, PS, TS, PS, S, PS.
Très satisfait (TS), satisfait (S), moyennement satisfait (MS), pas satisfait (PS).

Préciser :

a- La population étudiée.
b- Le caractère étudié.
c- La nature du caractère.
d- Le nombre de modalités du caractère.

Réponse :

a- La population étudiée : Les étudiants de la 3ème année de l'INSAP.
b- Le caractère étudié : Le degré de satisfaction.
c- La nature du caractère : Il est qualitatif.
d- Le nombre de modalités du caractère : 4 modalités : TS, S, MS et PS.

Exercice 4 :

On s'intéresse à l'analyse de 36 pièces de monnaie anciennes selon le lieu de frappe. Les résultats sont ainsi :

- 8 pièces frappées à Fès.
- 4 pièces frappées en Andalousie.
- 3 pièces frappées à Tlemcen.
- 21 pièces sans indication de l'atelier de frappe.

Préciser :

a- La population étudiée.
b- Le caractère étudié.
c- La nature du caractère.
d- Le nombre de modalités du caractère.

Réponse :

a- La population étudiée : Les 36 pièces de monnaies anciennes.
b- Le caractère étudié : Le lieu de frappe.
c- La nature du caractère : Il est qualitatif.
d- Le nombre de modalités du caractère : 4 modalités.

Chapitre III : Analyse statistique univariée ou variables statistiques à une dimension

Section 1 : Représentation graphique :

Les représentations graphiques ont l'avantage d'offrir une meilleure vue d'ensemble de la série statistique. Elles permettent par simple lecture, de voir les caractéristiques essentielles de la série, et aussi de comparer des séries différentes.

1) Caractère qualitatif :

Dans le cas d'un caractère qualitatif, on utilise principalement trois types de représentation graphique :

- ✓ Le diagramme en bâtons.
- ✓ L'histogramme.
- ✓ Le diagramme circulaire.

Lorsque le caractère étudié est la répartition géographique d'une population, la représentation graphique est un cartogramme.

a. Diagramme en bâtons :

Les modalités du caractère sont portées en abscisses (sur l'axe des x), les effectifs ou les fréquences correspondantes sont représentées en ordonnées (sur l'axe des y).

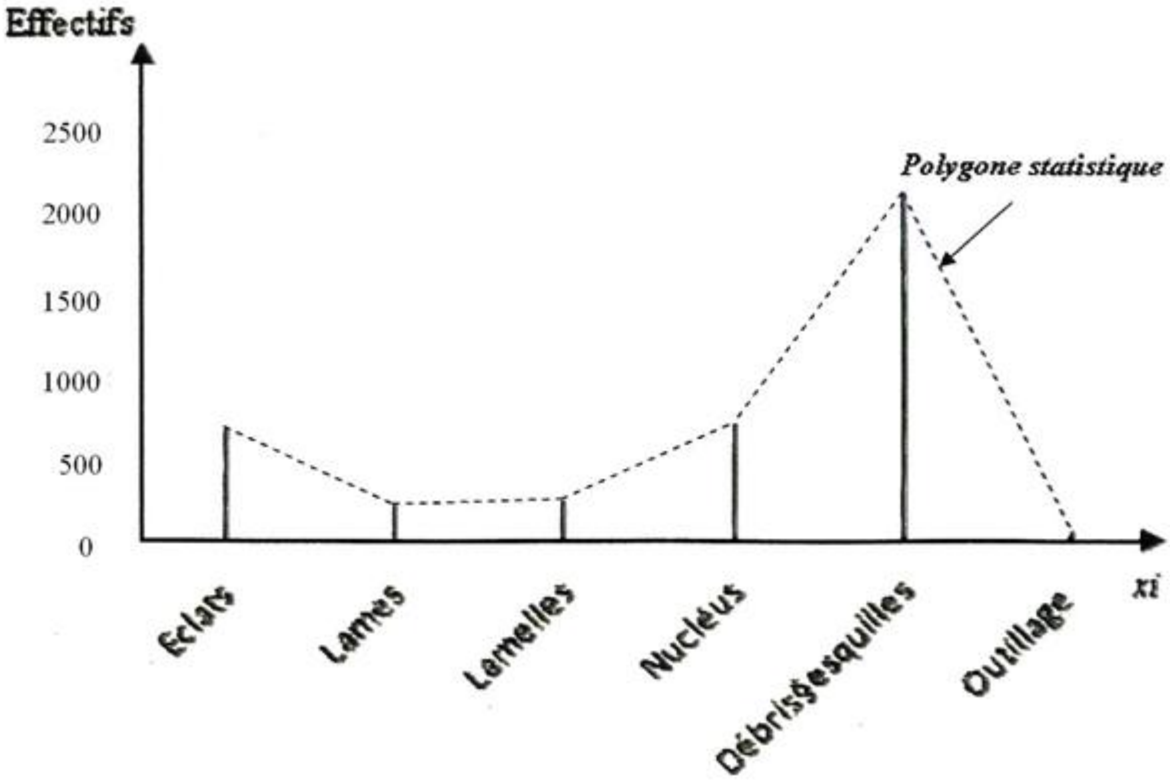
Si l'on joint les sommets des bâtons, on obtient le polygone statistique.

Exemple :

L'effectif total du mobilier lithique recueilli dans le site d'Ifri Ouberrid dans la commune d'Ain Elleuh se présente dans le tableau suivant :

Pièces	Effectifs
Eclats	690
Lames	222
Lamelles	249
Nucléus	716
Débris§esquilles	2129
Outillage	45
Total	4051

Le caractère étudié, le mobilier lithique recueilli dans le site d'Ifri Ouberrid, est un caractère qualitatif. La représentation graphique de cette distribution en diagramme en bâtons est la suivante :



b. Histogramme :

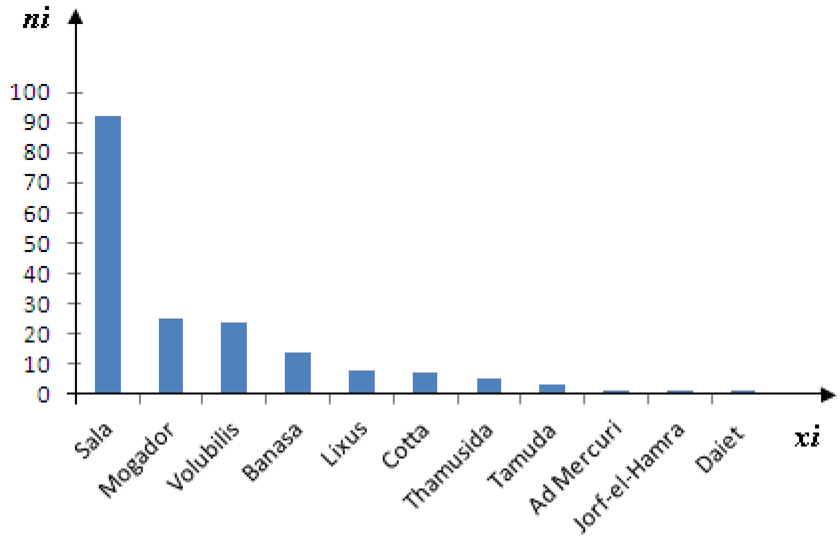
Nous portons en abscisses les modalités, et en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs ou aux fréquences de chaque modalité.

Exemple :

Les 181 marques de potiers italiques, répertoriées sur toute l'étendue du Maroc antique se répartissent entre les onze sites de la manière suivante :

Sites antiques	Effectifs
Sala	92
Mogador	25
Volubilis	24
Banasa	14
Lixus	8
Cotta	7
Thamusida	5
Tamuda	3
Ad Mercuri	1
Jorf-el-Hamra	1
Daiet	1
Total	181

La représentation graphique de cette distribution en histogramme est la suivante :



c. Diagramme circulaire :

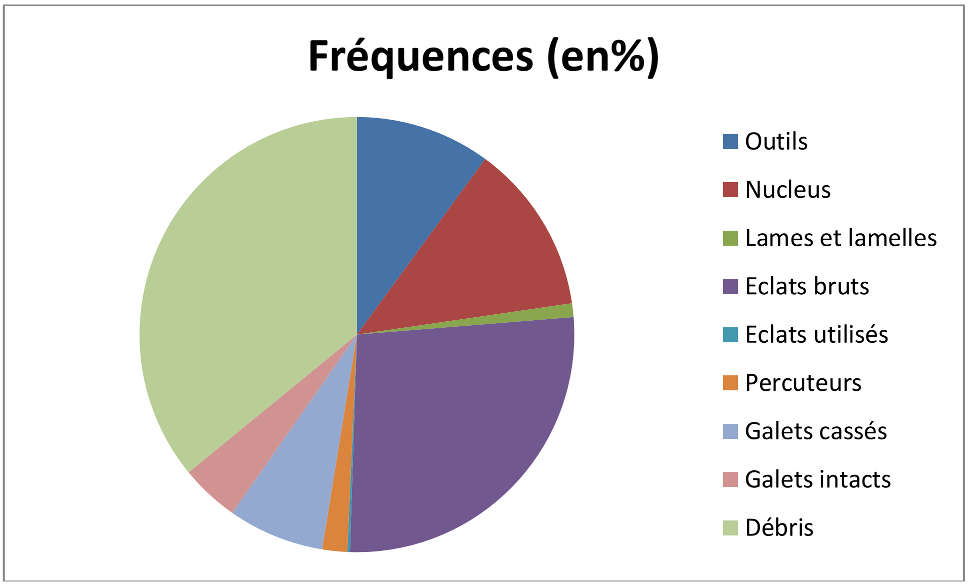
Un diagramme circulaire consiste à partager un disque en secteurs correspondants aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence de la modalité.

Ces digrammes conviennent très bien pour des données politiques ou socio-économiques.

Exemple :

L'inventaire des objets recueillis en 1979 dans le site du Chaperon rouge I, dans la partie sud-ouest du plateau de la Mamora, au sud de Rabat, se présente dans le tableau suivant :

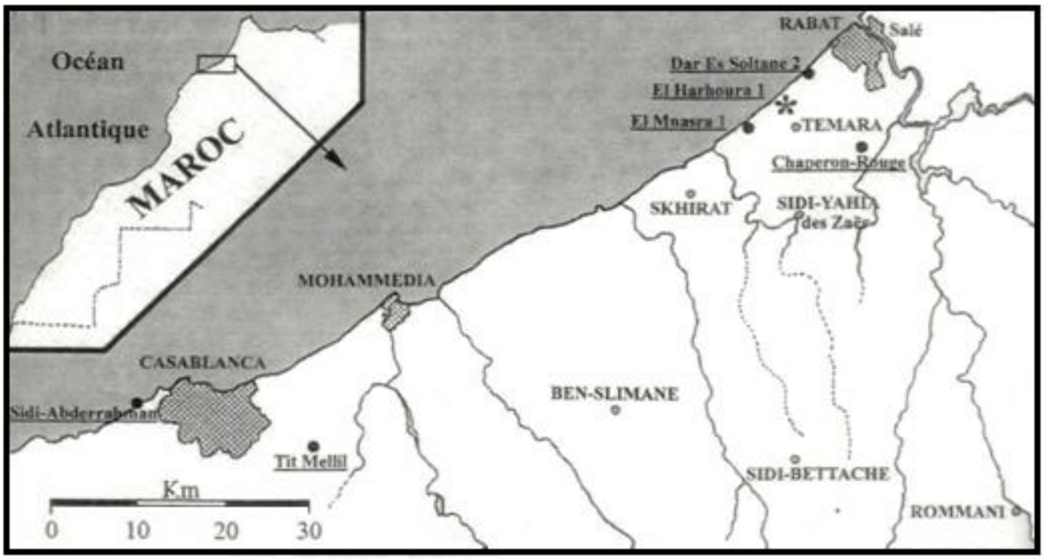
Objets	Effectifs	Fréquences (en%)
Outils	108	10,05
Nucleus	136	12,65
Lames et lamelles	11	1,02
Eclats bruts	288	26,79
Eclats utilisés	2	0,19
Percuteurs	20	1,86
Galets cassés	77	7,16
Galets intacts	47	4,37
Débris	386	35,91
Total	1075	100



a. Cartogramme :

Un cartogramme est une carte géographique dont les secteurs géographiques sont colorés avec une couleur différente suivant l'effectif ou suivant la fréquence du caractère étudié.

Exemple :



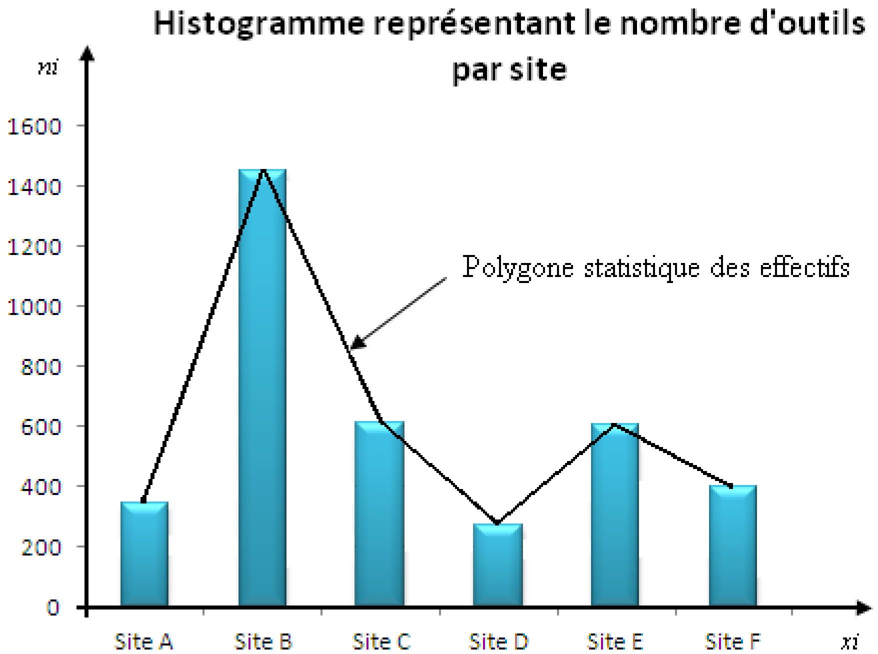
Localisation géographique de la grotte d'El Harhoura 1

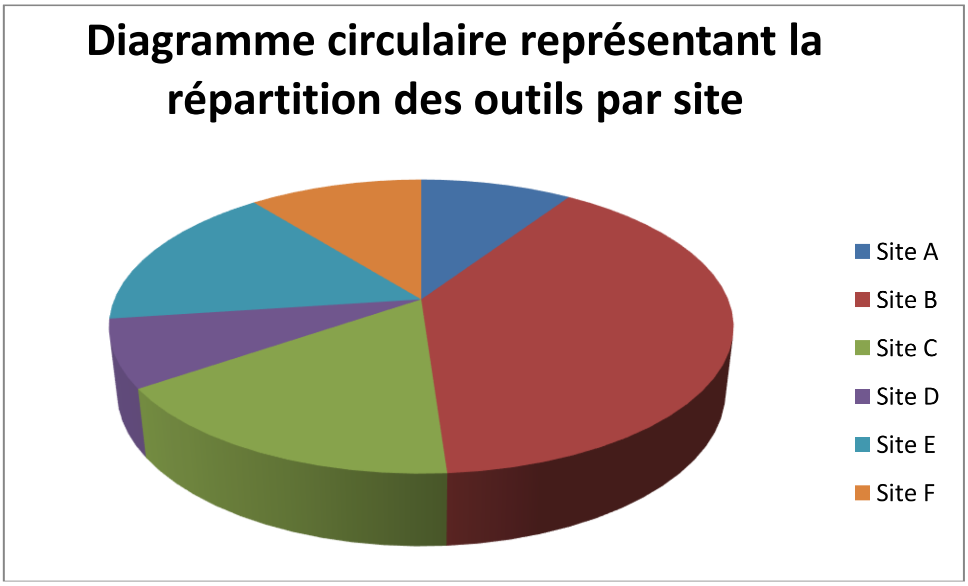
Exemple :

Les stations préhistoriques de surface des rives du Tensift, dans la région de Marrakech, ont livré un important outillage. Les outils recueillis dans les six sites découverts en 1976 sont présentés dans le tableau suivant :

Sites	Nombre d'outils n_i	Fréquences f_i (en %)
Site A	345	9,38
Site B	1453	39,52
Site C	609	16,56
Site D	273	7,43
Site E	602	16,37
Site F	395	10,74
Total	3677	100

L'axe des abscisses n'a pas de signification numérique. Les barres sont étiquetées en fonction de la modalité qu'elles représentent, et leurs longueurs sont proportionnelles à l'effectif de la modalité.





2) Caractère quantitatif :

Dans ce cas, la variable statistique peut être discrète ou continue.

Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

- ✓ Le diagramme différentiel correspond à une représentation des effectifs ou des fréquences.
- ✓ Le diagramme intégral correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

a. Variable statistique discrète :

- ➢ Diagramme différentiel : diagramme en bâtons, des effectifs ou des fréquences. La différence avec le cas qualitatif consiste en ce que les abscisses ici sont les valeurs de la variable statistique.
- ➢ Diagramme intégral : courbe en escaliers des effectifs cumulés ou des fréquences cumulées.

Exemple :

Le tableau ci-après présente la teneur en argent des pièces de monnaie anciennes (en pourcentage massique) selon le numéro de la pièce analysée :

Numéro de la pièce analysée	Teneur en argent
1	87,38
2	84,87
3	88,67
4	88,39
5	87,95
6	88,12
7	86,89
8	89,33
9	92,28
10	89,36
Total	883,24

On demande de :

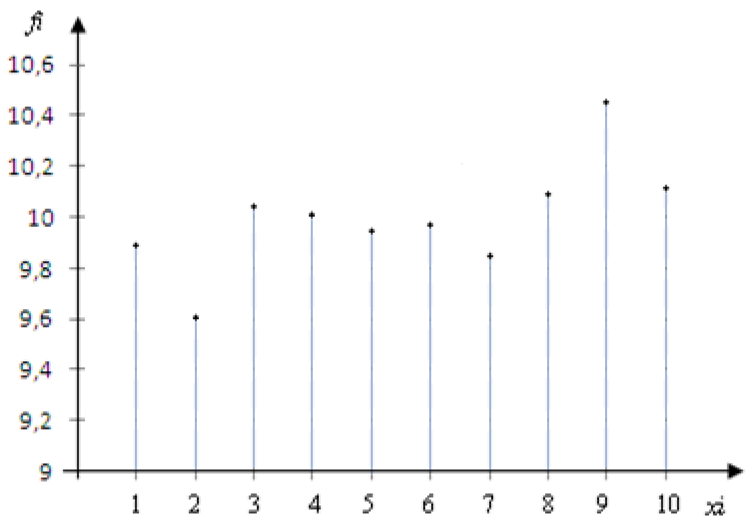
- Déterminer la population étudiée.
- Déterminer le caractère étudié.
- Représenter le diagramme différentiel adéquat.
- Représenter le diagramme intégral adéquat.

Solution :

- La population étudiée est la teneur en argent des pièces de monnaie analysée.
- Le caractère observé est le numéro correspondant à la référence d'analyse de la pièce de monnaie. C'est un caractère quantitatif et la variable statistique correspondante, qui ne peut prendre que des valeurs entières de 1 à 10, est discrète.
- La représentation graphique différentielle adéquate est le diagramme en bâtons. La représentation des effectifs est identique à celle des fréquences ; seule change l'échelle verticale.

Numéro de la pièce analysée x_i	Teneur en argent n_i	Fréquence f_i en %	Fréquence cumulée en %
1	87,38	09,89	09,89
2	84,87	09,61	19,50
3	88,67	10,04	29,54
4	88,39	10,01	39,55
5	87,95	09,96	49,51
6	88,12	09,97	59,48
7	86,89	09,84	69,32
8	89,33	10,11	79,43
9	92,28	10,45	89,88
10	89,36	10,12	100,00
Total	883,24	100	---

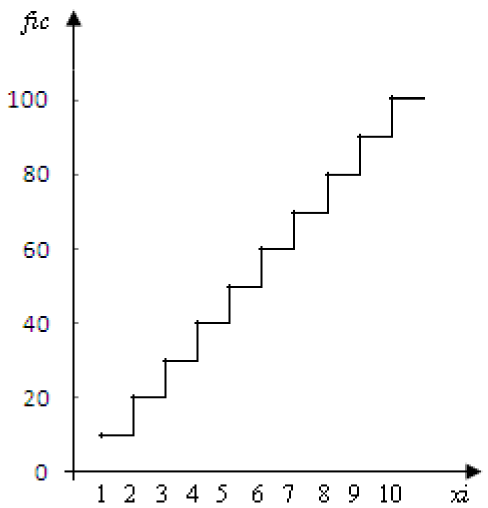
<u>Diagramme en bâtons</u>



A chaque valeur x_i de la variable, portée en abscisse, on fait correspondre un segment vertical de longueur proportionnelle à la fréquence f_i de cette valeur.

- La représentation graphique intégrale adéquate est la courbe en escalier : les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe en escalier.

Courbe en escalier



b. Variable statistique continue :

Dans ce cas les observations sont regroupées en classes.

Chaque classe possède une certaine amplitude, qui est la longueur de l'intervalle définissant la classe.

Le rapport entre l'effectif d'une classe et son amplitude s'appelle la densité d'effectif.

Le rapport entre la fréquence d'une classe et son amplitude s'appelle la densité de fréquence.

- Diagramme différentiel : histogramme des densités. Nous portons en abscisse les classes représentant les modalités et en ordonnées des rectangles dont la longueur est proportionnelle à la densité d'effectif ou à la densité de fréquence.
- Diagramme intégrale : courbe cumulative des effectifs ou des fréquences.

Exemple :

Soit le tableau suivant qui représente la distribution de 120 dirhams d'époque Almohade à Sidi el Mokhfi au Rharb du Maroc en fonction de leurs poids.

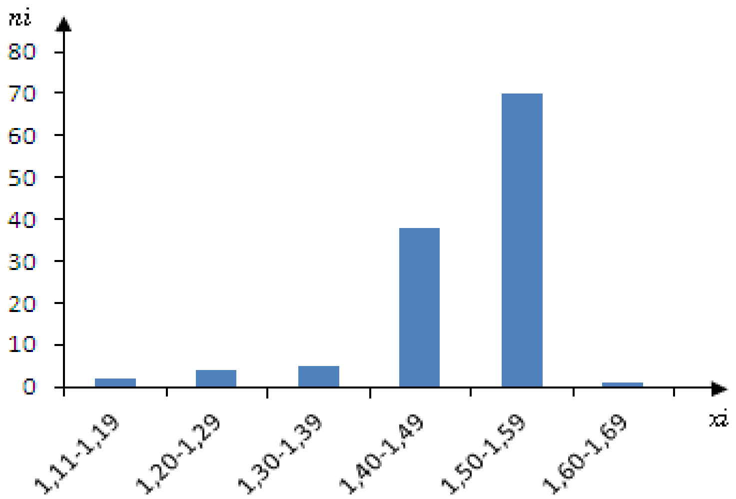
Poids (en g) x_i	Nombre de pièces de monnaie n_i	f_i (en %)
1,11-1,19	2	1,67
1,20-1,29	4	3,33
1,30-1,39	5	4,17
1,40-1,49	38	31,67
1,50-1,59	70	58,33
1,60-1,69	1	0,83
Total	120	100

On demande de :

- Déterminer la population étudiée.
- Déterminer le caractère étudié.
- Représenter le diagramme différentiel adéquat.
- Représenter le diagramme intégral adéquat.

Solution :

- La population étudiée est celle de 120 dirhams d'époque Almohade à Sidi el Mokhfi au Rharb du Maroc.
- Le caractère observé est le poids de ces dirhams. C'est un caractère quantitatif et la variable statistique correspondante est continue.
- La représentation graphique différentielle adéquate est l'histogramme.



Histogramme représentant la distribution de 120 dirhams d'époque Almohade à Sidi el Mokhfi au Rharb du Maroc en fonction de leurs poids.

Section 2 : Paramètres caractéristiques :

Le but de l'étude statistique est aussi de résumer des données par des paramètres. Il existe trois types de paramètres :

- ✓ Paramètres de position ou de tendance centrale.
- ✓ Paramètres de dispersion.
- ✓ Paramètres de forme.

Paragraphe 1 : Paramètres de position :

Les paramètres de position permettent de savoir autour de quelles valeurs se situent les valeurs de la variable statistique étudiée.

Les principaux paramètres de position sont :

- Le mode.
- La médiane.
- La moyenne.

A. Le mode :

Le mode, noté M_o, est la valeur du caractère à laquelle correspond le plus grand effectif ou la plus grande fréquence.

La détermination du mode varie selon la nature de la variable statistique étudiée.

i. ***Dans le cas d'une variable discrète*** **:**

La détermination du mode est facile et immédiate. Ainsi dans le tableau statistique des fréquences ou des effectifs, le mode est la valeur ou la modalité qui a le plus grand effectif ou la plus grande fréquence.

Remarque : Une série statistique peut avoir :

- Un seul mode, elle est dite dans ce cas une série unimodale.
- Deux modes, elle est dite dans ce cas une série bimodale.
- Plusieurs modes, elle est dite dans ce cas multimodale ou plurimodale.

Graphiquement, le mode correspond au bâton le plus élevé.

Exemple :

Soit la série statistique suivante relative à la répartition de l'outillage recueilli dans six sites archéologiques du Grand Casablanca :

Nombre de sites x_i	Effectifs n_i
1	190
2 ←	**1065**
3	185
4	367
5	451
6	242
Total	2500

La valeur 2 dont l'effectif est le plus grand est le mode de la série considérée.

Cette série n'a qu'un seul mode, c'est une série unimodale.

ii. *Dans le cas d'une variable continue* :

Pour déterminer le mode, on définit tout d'abord la classe modale dont l'effectif est le plus élevée, ensuite on détermine la valeur modale soit graphiquement soit algébriquement.

Exemple :

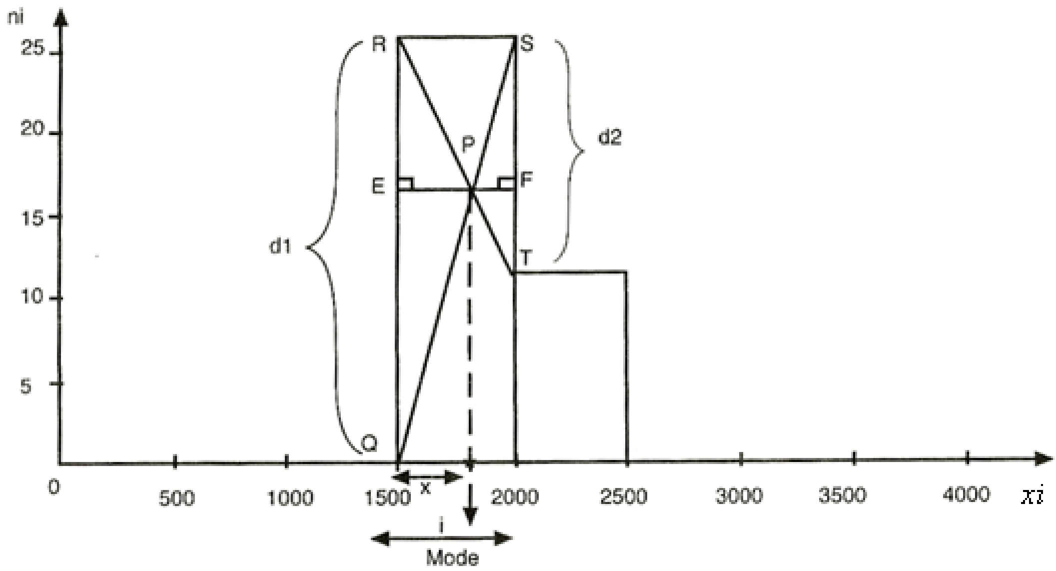
Le tableau suivant donne la répartition des lames recueillies parmi les objets trouvés dans un site archéologique :

Objets x_i	Effectifs n_i
1500-2000 ←	**26**
2000-2500	12
2500-3000	11
3000-3500	6
3500 et plus	5
Total	60

La classe modale dont l'effectif est le plus élevé est la classe : $[1500 - 2000[$.

a- <u>Détermination graphique du mode :</u>

Elle consiste à tracer l'histogramme, tout en se limitant à sa partie qui donne la classe modale et les deux classes adjacentes à celle-ci. En ce référant à l'exemple précédent, il vient :



Les triangles PQR et PST sont semblables et on peut écrire :

$$\frac{\overline{EP}}{\overline{QR}} = \frac{\overline{PF}}{\overline{TS}}$$

Le mode $M_0 = L_1 + x$
L_1 : Limite inférieure de la classe modale.

Déterminons x :

Soit i : Intervalle de la classe modale.

$$\frac{x}{d_1} = \frac{i-x}{d_2} \quad \Rightarrow d_2\, x = d_1(i - x)$$
$$= d_1 i - d_1\, x$$
$$\Rightarrow x\,(d_2 + d_1) = d_1 i$$
$$\Rightarrow x = \frac{d_1 i}{d_2+d_1}$$

Par suite :

$$M_o = L_1 + \frac{d_1 i}{d_1+d_2}$$

Par deux diagrammes, on joint les deux sommets du rectangle modal aux points d'intersection respectifs des deux hauteurs de ce rectangle avec les bases supérieures des deux rectangles qui l'encadrent. Le point d'intersection de la hauteur, passant par le point commun aux deux diagonales et l'axe des abscisses correspond au mode. La valeur modale est obtenue par lecture directe sur l'axe des abscisses.

b- Détermination algébrique du mode :

Une fois on a défini la classe modale, on détermine le mode en appliquant la formule démontrée plus haut :

$M_o = L_1 + \frac{d_1}{d_1+d_2} \times i$

Avec :

L_1 : Limite inférieure de la classe modale.

d_1 : Différence entre l'effectif ou la fréquence de la classe modale et l'effectif ou la fréquence de la classe qui précède la classe modale.

d_2 : Différence entre l'effectif ou la fréquence de la classe modale et l'effectif ou la fréquence de la classe post-modale.

i : Différence entre la limite supérieure et la limite inférieure de la classe modale.

Application :

Dans le cas de l'exemple précédent on a :

$M_o = L_1 + \frac{d_1}{d_1+d_2} \times i$

Avec :

L_1 : limite inférieure de la classe modale $[1500 - 2000[$.

$L_1 = 1500$

$d_1 = 26$

$d_2 = 26 – 12 = 14$

i = Différence entre la limite supérieure et la limite inférieure de la classe modale.

$i = 2000 – 1500 = 500$

D'où :

$M_o = 1500 + \frac{26}{26+14} \times 500$

$M_o = 1825$.

B. La médiane :

La médiane est une valeur de la variable statistique qui partage la population étudiée en deux effectifs égaux, telle que la moitié de ces effectifs lui soit inférieure et l'autre moitié lui soit supérieure.

La détermination de la médiane suppose que les valeurs prises par la variable soient rangées dans un ordre croissant.

1) Cas d'une variable discrète :

a- Lorsque le nombre d'observations de la variable est impair :

Dans ce cas le nombre d'observations est N=2n+1, la médiane M_e correspond à l'observation de rang (n+1).

Exemple :

Dans un site archéologique, un préhistorien a détecté des outils de tailles (en cm) suivantes :

15, 10, 9, 6, 11, 12, 5, 16, 7.

Rangeons par ordre croissant les tailles d'outils détectés :

5, 6, 7, 9, $\boxed{\mathbf{10}}$, 11, 12, 15, 16.

On détermine le rang (n+1), on a :

$$N = 2n + 1 = 9 \quad \Rightarrow n = 4$$

$$\Rightarrow n + 1 = 5$$

$$\Rightarrow M_e = 10 \text{ cm}$$

Donc la médiane M_e correspond à la $5^{ème}$ observation c'est-à-dire la taille de 10 cm.

Signification :

4 outils ont une taille inférieure à 10 cm et 4 autres outils ont une taille supérieure à 10 cm.

b- Lorsque le nombre d'observations de la variable est pair :

Dans ce cas N = 2n, la médiane M_e correspond à la demi-somme des observations de rang n et n+1.

Exemple :

Eliminons de l'exemple précédent la taille 5, il reste les tailles suivantes :

6, 7, 9, $\boxed{\mathbf{10}}$, $\boxed{\mathbf{11}}$, 12, 15, 16.

Le rang de la médiane est :

$$N = 2n = 8 \Rightarrow n = 4 \text{ et } n + 1 = 5$$

$$\Rightarrow M_e = \frac{10+11}{2} = 10{,}5 \text{ cm}$$

Remarque :

La médiane peut être indéterminée dans le cas suivant :

Supposons la série de tailles des outils suivante :

6, 7, 9, 11, 11, 12, 13.

$N = 2n + 1 = 7 \Rightarrow n = 3$

Rang de la médiane : $n + 1 = 4 \Rightarrow M_e = 11$

Cette valeur n'est pas la médiane, car on a : 3 outils qui ont une taille inférieure à 11cm et seulement 2 outils qui ont une taille supérieure à 11 cm. Dans, ce cas, il est préférable de dire que la médiane n'existe pas. La taille 10 peut être considérée comme caractéristique du centre de la série.

2) Cas d'une variable continue :

a- Détermination algébrique de la médiane :

Exemple :

Le tableau suivant donne la répartition des sites archéologiques fouillés selon les pièces de monnaies trouvées:

Pièces de monnaie x_i	Effectifs n_i	Effectifs cumulés croissants $n_i c\nearrow$	Effectifs cumulés décroissants $n_i c\searrow$
1500-2000	26	26	60
2000-2500	12	38	34
2500-3000	11	49	22
3000-3500	6	55	11
3500 et plus	5	60	5
Total	60	--	--

Le calcul de la médiane s'effectue en deux étapes :

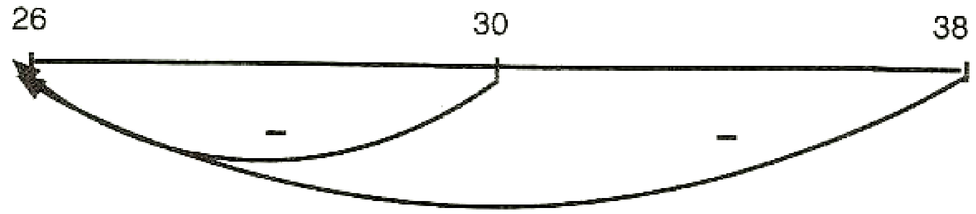
1ère étape : Détermination du rang de la médiane :

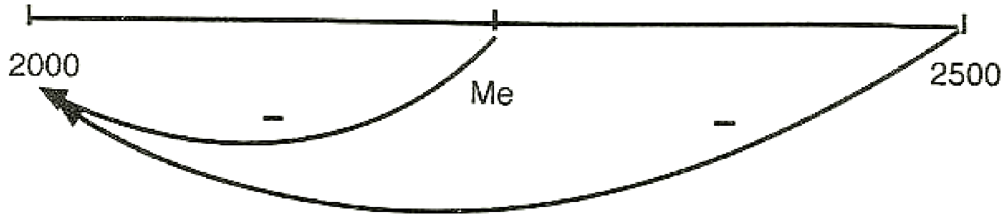
Rang de la M_e : $\frac{\sum n_i}{2} = \frac{60}{2} = 30$ ou $\frac{\sum f_i}{2} = \frac{1}{2} = 0{,}5$

La médiane correspond aux pièces de monnaie du 30ème site archéologique.

2ème étape : Détermination de la classe M_e :

En consultant la colonne des effectifs cumulés croissants, on voit que la médiane appartient à la classe $[2000 - 2500[$ et en appliquant le schéma suivant :





$\frac{M_e - 2000}{2500-2000} = \frac{30-26}{38-26} \quad \Rightarrow \frac{M_e - 2000}{500} = \frac{4}{12}$

$\Rightarrow M_e = \frac{4 \times 500}{12} + 2000$

$\Rightarrow M_e = 2000 + \frac{2000}{12}$

$\Rightarrow M_e = 2000 + 166{,}67 = 2166{,}67$.

Ce qui donne la formule suivante :

$$M_e = L_1 + [(L_2 - L_1)\, \frac{\frac{\sum n_i}{2} - \sum 1n_i}{\sum 2n_i - \sum 1n_i}\,]$$

Avec :

L_1 : Limite inférieure de la classe médiane.

L_2 : Limite supérieure de la classe médiane.

$\sum 1n_i$: Effectif cumulé croissant de la classe précédent la classe médiane.

$\sum 2n_i$: Effectif cumulé croissant de la classe médiane.

Signification :

✓ 50% des sites archéologiques disposent de moins de 2166,67 pièces de monnaie.

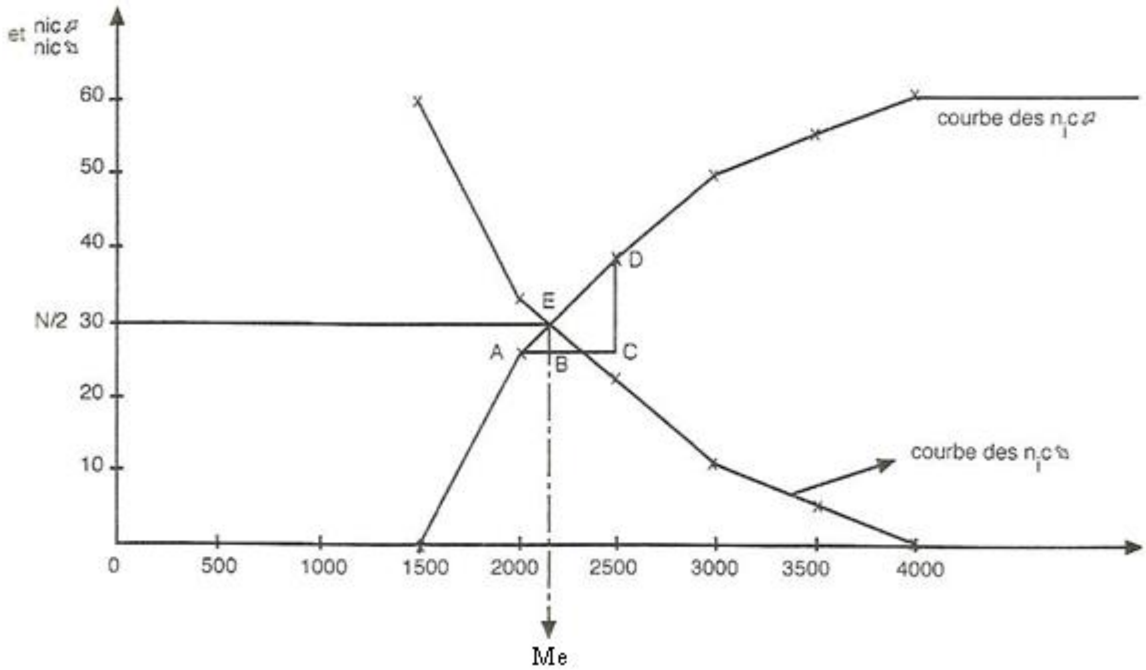
✓ 50% des sites archéologiques disposent de plus de 2166,67 pièces de monnaie.

b- Détermination graphique de la médiane :

La détermination graphique de la médiane consiste à tracer la courbe des fréquences cumulées croissantes (ou des effectifs cumulés croissants) et la courbe des fréquences cumulées décroissantes (ou des effectifs cumulés décroissants), la médiane est l'abscisse du point d'intersection de ces deux courbes (avec la droite d'ordonnée égale à $\frac{\sum n_i}{2}$ ou $\frac{\sum f_i}{2}$ qui est parallèle à l'axe des abscisses).

Exemple :

Reprenons l'exemple précédent et traçons la courbe des effectifs cumulés croissants $n_i c\nearrow$ et la courbe des effectifs cumulés décroissants $n_i c\searrow$.



Remarque : On peut se limiter à tracer uniquement la courbe cumulative croissante ou la courbe des $n_i c\nearrow$, dans ce cas, la médiane est l'abscisse du point d'intersection de cette courbe avec la droite parallèle à l'axe des abscisses, d'ordonnée $\frac{N}{2}$ ou 0,5.

En se référant à la courbe des effectifs cumulés croissants $n_i c\nearrow$, les deux triangles ADC et AEB sont semblables. Les propriétés des triangles semblables nous permettent d'écrire :

$$\frac{\overline{AB}}{\overline{AC}} = \frac{\overline{BE}}{\overline{CD}} \Rightarrow \frac{M_e - 2000}{2500-2000} = \frac{30-26}{38-26}$$

$$\Rightarrow M_e = 2166,67$$

C. La moyenne :

1) La moyenne arithmétique :

a- Définition :

On appelle moyenne arithmétique de la variable x, notée $\overline{x}$, le nombre obtenu en divisant la somme de toutes les valeurs prises par cette variable par le nombre total d'observations (ou effectif total). De cette définition découle deux formules de la moyenne arithmétique :

- *Formule simple* : Elle s'applique au cas où chaque valeur x_i n'a été observée qu'une seule fois. Elle s'écrit :

$$\overline{x} = \frac{x_1+x_2+\cdots+xk}{n_1+n_2+\cdots+nk} = \frac{\sum_{i=1}^{k} x_i}{N}$$

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i}{N}$$

Exemple :

Supposons que lors d'une analyse des pièces de monnaie almohades, un chercheur a recueilli cinq pièces de monnaie frappées en Andalousie de dimensions différentes (en mm) :

8-10-11-12-9

La moyenne est :

$$\overline{x} = \frac{\sum_{i=1}^{5} x_i}{N} = \frac{8+10+11+12+9}{5} = \frac{50}{5} = 10$$

$\overline{x}$ = 10 mm

La dimension moyenne de ces cinq pièces de monnaie frappées en Andalousie est de 10 mm.

- *Formule pondérée* : Elle s'applique au cas où les valeurs de la variable ont été observées plusieurs fois, cela revient à pondérer chaque valeur de la variable par l'effectif qui lui correspond. Elle s'écrit :

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}$$

Exemple :

Soit le tableau suivant relatif aux outils recueillis lors d'une fouille archéologique suivant le nombre de nucléus.

Outils x_i	nucléus n_i	$n_i x_i$
10	5	50
09	4	36
12	4	48
14	2	28
15	2	30
Total	$\sum ni$ = 17	192

La moyenne est :

$$\overline{x} = \frac{\sum_{i=1}^{5} n_i x_i}{\sum_{i=1}^{5} n_i} = \frac{192}{17} = 11{,}29$$

Le nombre moyen des nucléus parmi les outils recueillis est de 11,29.

b- Méthodes de calcul de la moyenne arithmétique :

- *Cas d'une variable discrète :*

Calcul direct de $\overline{x}$ partir des fréquences. Comme on sait que :

$f_i = \frac{n_i}{N} \quad \Rightarrow \quad n_i = N\, f_i$

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i}\ \overline{x} = \frac{\sum_{i=1}^{k} (N f_i) x_i}{N} = \frac{N \sum_{i=1}^{k} f_i x_i}{N}$$

$$\overline{x} = \sum_{i=1}^{k} f_i x_i$$

Exemple :

Soit la série statistique suivante relative à la distribution des gravures dans six secteurs d'un site de gravures rupestres :

Secteurs x_i	Gravures n_i	$n_i\, x_i$	$f_i\, x_i$
1	20	20	0,20
2	35	70	0,70
3	23	69	0,69
4	12	48	0,48
5	7	35	0,35
6	3	18	0,18
Total	100	260	2,60

$$\overline{x} = \frac{\sum_{i=1}^{6} n_i x_i}{\sum_{i=1}^{6} n_i} = \frac{260}{100} = 2{,}6$$

$\Rightarrow$ 2,6 est le nombre moyen de gravures rupestres.

- *Cas d'une variable continue :*

Calcul direct de $\overline{x}$ à partir de la formule de définition :

$$\overline{x} = \frac{\sum_{i=1}^{k} n_i x_i}{\sum_{i=1}^{k} n_i} \quad \text{où} \quad \begin{cases} n_i \textit{ désigne l'effectif de la classe } i \\ \textit{et} \\ x_i \textit{ désigne le centre de la classe } i \end{cases}$$

Le centre de chaque classe x_i s'obtient de la manière suivante :

$$x_i = \frac{L_1 + L_2}{2} \quad \text{où} \quad \begin{cases} L_1 \textit{ désigne la limite inférieure de la classe } i \\ \textit{et} \\ L_2 \textit{ désigne la limite supérieure de la classe } i \end{cases}$$

Exemple :

Le tableau suivant donne la répartition de 80 sites archéologiques d'une certaine région du royaume selon le nombre d'objets trouvés pendant une période donnée :

Nombre d'objets trouvés	Sites archéologiques n_i	Centres de classes x_i	$n_i\,x_i$
200-300	10	250	2500
300-400	18	350	6300
400-500	30	450	13500
500-600	12	550	6600
600-700	6	650	3900
700-800	4	750	3000
Total	80	---	35800

$$\overline{x} = \frac{\sum_{i=1}^{6} n_i x_i}{\sum_{i=1}^{6} n_i} = \frac{35800}{80} = 447{,}5$$

C'est le nombre moyen d'objets trouvés dans les 80 sites archéologiques.

2) La moyenne géométrique :

a- Définition :

Soit une variable statistique x pouvant prendre N valeurs : x_1, x_2,…, x_n ; on appelle moyenne géométrique de ces valeurs la racine $N^{\text{ième}}$ de leur produit, elle se note G.

La moyenne géométrique a pour expression :

$$G = \sqrt[N]{x_1 \times x_2 \times \ldots \times x_n} = [\prod_{i=1}^{n} x_i]^{1/n}$$

$$G = \sqrt[\sum n_i]{x_1^{n_1} \times x_2^{n_2} \times \ldots \times x_n^{nk}} = [\prod_{i=1}^{k} x_i^{n_i}]^{1/\sum n_i}$$

b- Calcul de la moyenne géométrique :

Le calcul de G est toujours long et pénible, c'est pourquoi on a souvent avantage à utiliser les logarithmes :

Formule simple :

$$\log G = \frac{\sum_{i=1}^{n} \log x_i}{N}$$

Formule pondérée :

$$\log G = \frac{\sum_{i=1}^{n} n_i \log x_i}{\sum_{i=1}^{n} n_i}$$

c- Champ d'application de G :

Le calcul de G s'applique lorsque les valeurs de la variable se multiplient, c'est-à-dire suivent une progression géométrique. Elle permet l'estimation du rapport moyen de variation d'une variable évolutive.

Par exemple :

- Le taux de d'accroissement moyen des fouilles archéologiques dans le site Ifri n'Ammar.
- Le taux de croissance moyen des événements hydrologiques enregistrés dans les dépôts de la Basse Moulouya.

Exemple :

Reprenons l'exemple précédent relatif à la distribution des gravures dans six secteurs d'un site de gravures rupestres :

Secteurs	Gravures n_i	$\log x_i$	$n_i \log x_i$
1	20	0	0
2	35	0,30103	10,53605
3	23	0,47712	10,97376
4	12	0,60206	7,22472
5	7	0,69897	4,89279
6	3	0,77815	2,33445
Total	100	---	35,96177

La moyenne géométrique est :

$$\log G = \frac{\sum_{i=1}^{6} n_i \log x_i}{\sum_{i=1}^{6} n_i} = \frac{35{,}96177}{100} = 0{,}3596177$$

$$\Rightarrow G = 10^{0{,}3596177} = 2{,}28885 \approx 2{,}29$$

3) La moyenne harmonique :

a- Définition :

On appelle moyenne harmonique, notée H de k valeurs : x_1, x_2, …, x_k prises par une variable x, le nombre dont l'inverse est égal à la moyenne arithmétique des inverses des valeurs de cette variable.

Formule simple :

$$\frac{1}{H} = \frac{1}{N}\sum_{i=1}^{k}\frac{1}{x_i} \Rightarrow H = \frac{N}{\sum_{i=1}^{k}\frac{1}{x_i}}$$

Formule pondérée :

$$\frac{1}{H} = \frac{\sum_{i=1}^{k}\frac{n_i}{x_i}}{\sum_{i=1}^{k} n_i} \Rightarrow H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i}}$$

b- Champ d'application de H :

La moyenne harmonique est utilisée dans les questions traitant des taux, des prix, des vitesses, …, c'est-à-dire les cas où la grandeur varie selon une raison inverse d'une autre valeur.

Exemple :

Reprenons l'exemple précédent relatif à la distribution des gravures dans six secteurs d'un site de gravures rupestres :

Secteurs x_i	Gravures n_i	$1/x_i$	n_i/x_i
1	20	1	20
2	35	0,5	17,5
3	23	0,3333	7,6659
4	12	0,25	3
5	7	0,20	1,4
6	3	0,1667	0,5
Total	100	---	50,0659

La moyenne harmonique est :

$$\frac{1}{H} = \frac{\sum_{i=1}^{6}\frac{n_i}{x_i}}{\sum_{i=1}^{6} n_i} = \frac{50{,}0659}{100}$$

$$\Rightarrow H = \frac{N}{\sum_{i=1}^{6}\frac{n_i}{x_i}} = \frac{100}{50{,}0659} = 1{,}9973$$

$$\Rightarrow H \approx 2$$

4) La moyenne quadratique :

a- Définition :

On appelle moyenne quadratique des valeurs $x_1, x_2, \ldots, x_k$ prises par une variable x, le nombre Q tel que son carré est égal à la moyenne arithmétique des carrées de ces valeurs.

Formule simple :

$$Q^2 = \frac{\sum_1^k x_i^2}{N} = \frac{\sum_1^k x_i^2}{\sum_i^k n_i}$$

$$\Rightarrow Q = \sqrt{\frac{\sum_1^k x_i^2}{\sum_1^k n_i}}$$

Formule pondérée :

$$Q^2 = \frac{\sum_1^k n_i x_i^2}{N} = \frac{\sum_1^k n_i x_i^2}{\sum_i^k n_i}$$

$$\Rightarrow Q = \sqrt{\frac{\sum_1^k n_i x_i^2}{\sum_1^k n_i}}$$

b- Champ d'application de Q :

La moyenne quadratique est utilisée lorsque la variable x est élevée au carré, elle trouve son application dans le calcul de la variance.

Remarque :

Quelque soient les valeurs observées on montre que les moyennes : arithmétique, géométrique, harmonique et quadratique doivent vérifier toujours l'inégalité suivante :

$$H < G < \bar{x} < Q$$

Exemple :

Reprenons l'exemple précédent relatif à la distribution des gravures dans six secteurs d'un site de gravures rupestres :

Secteurs x_i	Gravures n_i	x_i^2	$n_i\, x_i^2$
1	20	1	20
2	35	4	140
3	23	9	207
4	12	16	192
5	7	25	175
6	3	36	108
Total	100	---	842

La moyenne quadratique est :

$$Q^2 = \frac{\sum_1^6 n_i x_i^2}{N} = \frac{\sum_1^k n_i x_i^2}{\sum_1^6 n_i} = \frac{842}{100} = 8{,}42$$

Donc : $Q = \sqrt{8{,}42} = 2{,}90$

On vérifie que : $H < G < \bar{x} < Q$

On a bien : $2 < 2{,}29 < 2{,}6 < 2{,}90$

Exercice d'application :

On donne le nombre d'outils recueillis par archéologue :
-20 archéologues n'ont recueilli aucun outil.
-40 archéologues ont recueilli 1 outil.
-60 archéologues ont recueilli 2 outils.
-40 archéologues ont recueilli 3 outils.
-20 archéologues ont recueilli 4 outils.

On demande de :
1) Dresser le tableau statistique.
2) Calculer la médiane, le mode et la moyenne arithmétique.
3) Déterminer graphiquement la médiane.

Solution :

1. Le caractère étudié est le nombre d'outils recueillis. Le tableau statistique se présente comme suit :

Outils x_i	n_i	$f_i \nearrow$ %	f_i c $\nearrow$ %	$n_i\, x_i$	$n_i\, x_i^2$
0	20	11,11	11,11	0	--
1	40	22,22	33,33	40	40
2	60	33,34	66,67	120	240
3	40	22,22	88,89	120	360
4	20	11,11	100,00	80	320
Total	180	100	---	360	960

2. a. Calcul de la médiane :

Le caractère étudié est discret. Le nombre d'observations est pair et est égal à 180.

Dans ce cas le rang de la médiane est :

N = 2n d'où n = 90ème observation.

La médiane M_e correspond à la demi-somme de la nième et la (n+1)ième observation.

On aura : $M_e = \frac{2+2}{2} = \frac{4}{2}$ = 2 outils.

On peut donc dire que 50% des archéologues ont recueilli moins de 2 outils et 50% ont recueilli au moins 2 outils.

b. Calcul du mode :

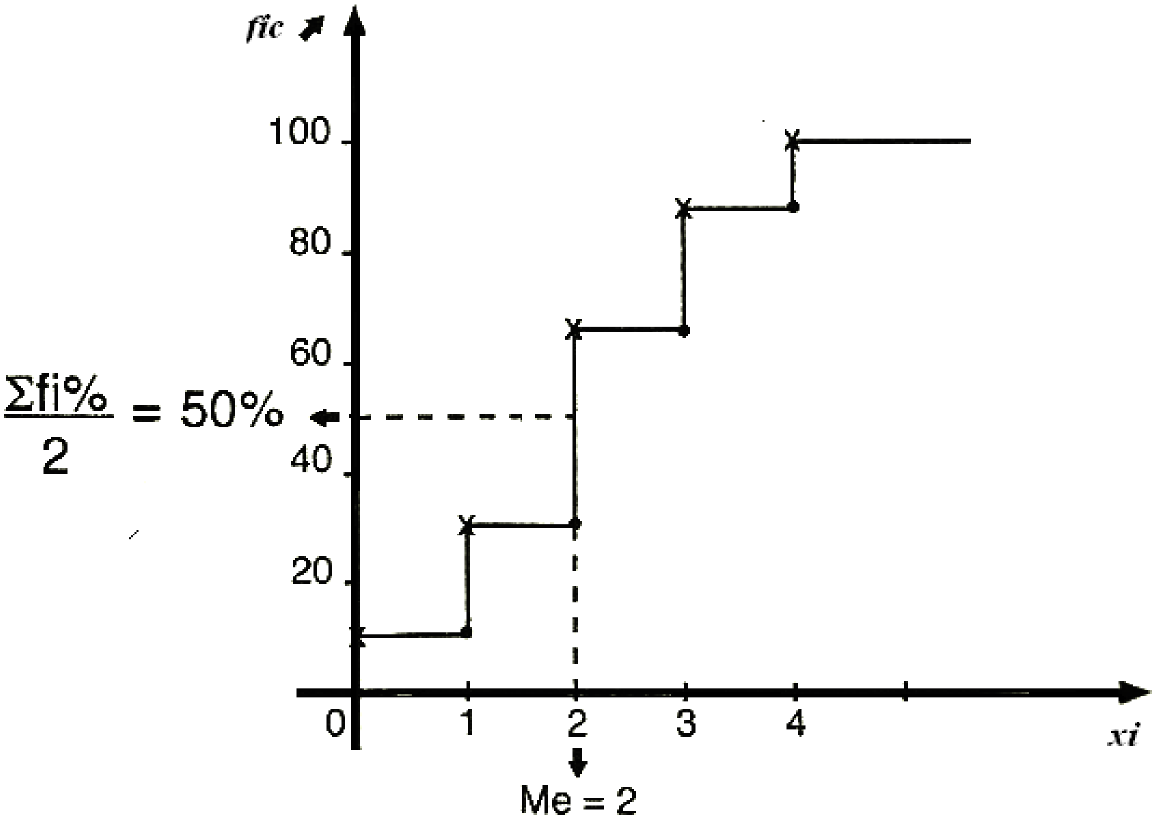
Le mode correspond à la modalité la plus observée.
M_o = 2 outils

c. Calcul de la moyenne arithmétique :

$\bar{x} = \frac{\sum_{i=1}^{5} n_i x_i}{\sum_{i=1}^{5} n_i} = \frac{360}{180}$ = 2 outils

Chaque archéologue a recueilli en moyenne 2 outils.

3. Détermination graphique de la médiane par la courbe cumulative croissante des fréquences.



Paragraphe 2 : Paramètres de dispersion :

Les caractéristiques de tendance centrale ne sont pas toujours suffisantes pour caractériser une série statistique car deux séries statistiques, peuvent avoir le même mode, la même médiane, la même moyenne et, cependant, être distribuées différemment comme le montre l'exemple suivant :

Exemple :

Soient deux sites S_1 et S_2 disposant d'un certain nombre d'objets découverts par un archéologue pendant 7 mois.

Site S_1 : 78-79-80-$\boxed{80}$-80-81-82.

Site S_2 : 30-35-80-$\boxed{80}$-80-125-130.

On constate que ces deux séries ont la même médiane égale à 80 ; la même moyenne égale à 80 et le même mode égal à 80. Cependant elles ne sont pas comparables ; elles sont donc différentes par la dispersion de leurs valeurs.

Dans le site S_1, les objets sont resserrés ou concentrés autour de la valeur centrale 80 ; tandis que dans le site S_2 les objets sont plus étalés, plus dispersés par rapport à cette valeur centrale typique.

Ainsi pour donner une représentation plus suffisante du phénomène étudié, il est important de mesurer la dispersion entre les différentes valeurs prises par une variable statistique et leur valeur centrale.

1) L'étendue :

Définition :

L'étendue d'une série statistique notée E est la différence entre la plus grande valeur et la plus petite valeur de la variable statistique.

Exemple :

1) Soit la série d'outils suivante :
   8-12-15-20-25-30-40-60
   $E = 60 - 8$
   $E = 52$

L'étendue est donc égale à 52.

2) Reprenons le tableau donnant la répartition de 80 sites archéologiques d'une certaine région du royaume selon le nombre d'objets trouvés pendant une période donnée.

Nombre d'objets trouvés	Sites archéologiques n_i
200-300	10
300-400	18
400-500	30
500-600	12
600-700	6
700-800	4
Total	80

L'étendue ou domaine de variation est :

$E = 800 - 200$

$E = 600$ objets.

2) Les quantiles :

Définition :

Les quantiles sont des paramètres qui nous renseignent sur la dispersion des valeurs de la variable autour de la médiane.

a- Les quartiles :

Les quartiles sont des valeurs du caractère qui partagent la série statistique en quatre parties égales ; ils sont au nombre de trois que l'on désigne par Q_1, Q_2 et Q_3.

- Le premier quartile : Q_1 est la valeur du caractère telle que 25% des observations lui soient inférieures et 75% des observations lui soient supérieures.
- Le deuxième quartile : Q_2 correspond à la médiane.
- Le troisième quartile : Q_3 est la valeur du caractère telle que 75% des observations lui soient inférieures et 25% des observations lui soient supérieures.

L'intervalle interquartile ou l'écart interquartile est la différence entre Q_3 et Q_1, on le note :

$$I = Q_3 - Q_1$$

Cet intervalle contient 50% des observations, c'est un indicateur très important en matière de dispersion.

Détermination des quartiles :
Comme la médiane, les quartiles se déterminent à partir des effectifs (ou des fréquences) cumulés, soit par le calcul, soit graphiquement.

Détermination des quartiles par le calcul :

Exemple :
A partir de l'exemple précédent relatif à la répartition de 80 sites archéologiques selon le nombre d'objets trouvés, calculons les quartiles et l'intervalle interquartile.

Nombre d'objets trouvés	Sites archéologiques n_i
200-300	10
300-400	18
400-500	30
500-600	12
600-700	6
700-800	4
Total	80

Calculons Q_1 : On suit trois étapes :

- Rang de $Q_1 = \frac{\sum_1^6 n_i}{4} = \frac{80}{4} = 20$

- Classe contenant Q_1 = [300 - 400[

- L'interpolation linéaire donne :
  $Q_1 = 300 + [(400 - 300)\,\frac{20-10}{28-10}]$
  $Q_1 = 300 + (100 \times \frac{10}{18})$
  $Q_1 = 355{,}55$

  Soit Q_1 = 355,55 objets.

Interprétation :

25% des sites archéologiques disposent d'un nombre inférieur à 355,55 objets, alors que 75% des sites disposent d'un nombre supérieur à 355,55 objets archéologiques.

Calculons Q_2 : On suit trois étapes :

- Rang de Q_2 : $N = 2n \Rightarrow n = \frac{80}{2} = 40$

- Classe contenant Q_2 = [400 - 500[

- L'interpolation linéaire donne :
  $Q_2 = M_e = 400 + [(500 - 400)\,\frac{40-28}{58-28}]$
  $Q_2 = 400 + (100 \times \frac{12}{30})$
  $Q_2 = 440$

Soit Q_2 = 440 objets.

Signification :

50% des sites archéologiques disposent d'un nombre inférieur à 440 objets et 50% des sites disposent d'un nombre supérieur à 440 objets archéologiques.

Calculons Q_3 : On suit trois étapes :

- Rang de $Q_3 = \frac{\sum_1^6 n_i \times 3}{4} = 60$

- Classe contenant Q_3 = [500 - 600[

- L'interpolation linéaire donne :
  $Q_3 = 500 + [(600 - 500)\,\frac{60-58}{70-58}]$
  $Q_3 = 500 + (100 \times \frac{2}{12})$
  $Q_3 = 516{,}67$

  Soit Q_3 = 516,67 objets.

Signification :

75% des sites archéologiques disposent d'un nombre inférieur à 515,67 objets et 25% des sites disposent d'un nombre supérieur à 516,67 objets archéologiques.

L'intervalle interquartile :
$I_q = Q_3 - Q_1$
$I_q = 516{,}67 - 355{,}55$
$I_q = 161{,}12$

Soit I_q= 161,12 objets.

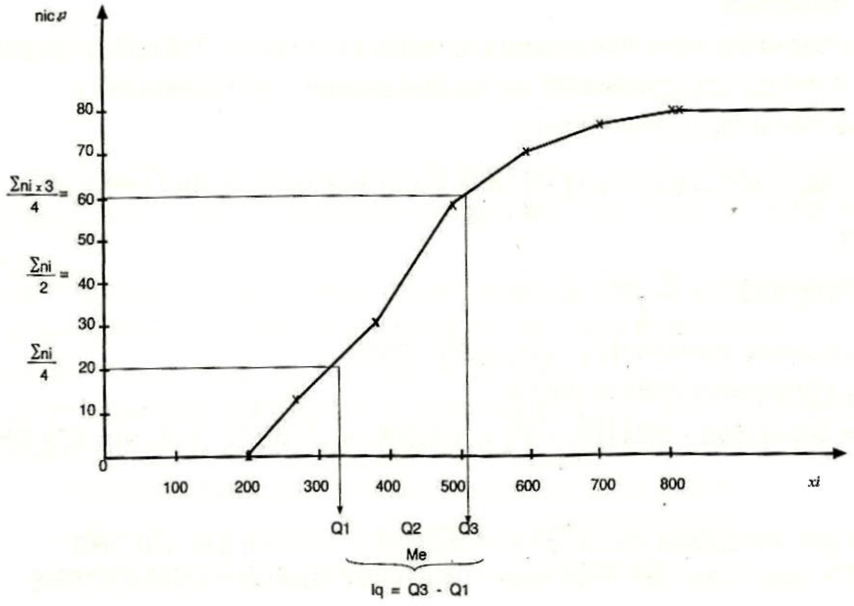
Détermination graphique des quartiles et l'intervalle interquartile :

On trace la courbe des fréquences cumulées croissantes ou des effectifs cumulés croissants.

Exemple :

Reprenons l'exemple précédent relatif à la répartition de 80 sites archéologiques selon le nombre d'objets trouvés.

Nombre d'objets trouvés	Sites archéologiques n_i
200-300	10
300-400	18
400-500	30
500-600	12
600-700	6
700-800	4
Total	80



b- Les déciles :

Les déciles sont des valeurs qui partagent la série statistique en 10 parties égales. Ils sont au nombre de 9 : D_1, D_2, …, D_9 ; les plus caractéristiques sont le 1er décile D_1 et le 9ème décile D_9 :

- Le 1er décile ou D_1 est la valeur du caractère telle que 10% des observations lui soient inférieur et 90% des observations lui soient supérieur.
- Le 9ème décile ou D_9 est la valeur du caractère telle que 90% des observations lui soient inférieur et 10% lui soient supérieur.

L'intervalle interdécile est la différence entre D_9 et D_1, il contient 80% des observations.

$$I_D = D_9 - D_1$$

Détermination des déciles :

Les déciles se déterminent de la même manière que les quartiles à partir des effectifs (ou des fréquences) cumulés.

Exemple :

Calcul de D_1 : On suit trois étapes :

- Rang de $D_1 = \frac{\sum_1^6 n_i}{10} = \frac{80}{10} = 8$

- Classe contenant D_1 = [200 - 300[

- L'interpolation linéaire donne :
  $D_1 = 200 + [(300 - 200)\,\frac{8-0}{10-0}]$
  $D_1 = 200 + (100 \times \frac{8}{10})$
  $D_1 = 280$

  Soit $D_1 = 280$ objets.

Signification :
10% des sites archéologiques disposent d'un nombre inférieur à 280 objets archéologiques.

Calcul de D_9 : On suit trois étapes :

- Rang de $D_9 = \frac{\sum_1^6 n_i \times 9}{10} = \frac{80\times 9}{10} = 72$

- Classe contenant D_9 = [600 - 700[

- L'interpolation linéaire donne :
  $D_9 = 600 + [(700 - 600)\,\frac{72-70}{76-70}]$
  $D_9 = 600 + (100 \times \frac{2}{6})$
  $D_9 = 633{,}33$

  Soit $D_9 = 633{,}33$ objets.

Signification :
90% des sites archéologiques disposent d'un nombre inférieur à 633,33 objets archéologiques.

Intervalle interdécile est :

$I_D = D_9 - D_1$
$I_D = 633{,}33 - 280$
$I_D = 353{,}33$

Soit $I_D = 353{,}33$ objets.

3) L'écart absolu moyen :

Définition :

L'écart absolu moyen peut être calculé soit par rapport à la médiane (rarement utilisé), soit par rapport à la moyenne (le plus utilisé).

L'écart absolu moyen par rapport à la moyenne arithmétique, noté e$\bar{x}$ est la moyenne arithmétique des écarts absolus entre les valeurs xi de la variable x et la moyenne arithmétique $\bar{x}$. Il a pour expression :

$$e\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i |\ x_i - \bar{x}\ |$$

Exemple :

Soit la série statistique suivante relatif aux découvertes de céramiques campaniformes au Maroc selon le nombre de fragments découverts :

Nombre de fragments	n_i	x_i	$n_i\ x_i$	x_i - $\bar{x}$	$n_i|\ x_i - \bar{x}\ |$
08-10	22	9	198	3,06	67,32
10-12	28	11	308	1,06	29,68
12-14	30	13	390	0,94	28,2
14-16	15	15	225	2,94	44,1
16-18	5	17	85	4,94	24,7
Total	100	--	1206	---	194

Calculons d'abord $\bar{x}$:

$$\bar{x} = \frac{\sum_1^5 n_i x_i}{\sum_1^5 n_i} = \frac{1206}{100} = 12{,}06$$

Ensuite on calcule e$\bar{x}$:

$$e\bar{x} = \frac{1}{N}\sum_{i=1}^{5} n_i |\ x_i - \bar{x}\ |$$

$$e\bar{x} = \frac{194}{100} = 1{,}94$$

Remarque :

L'écart absolu moyen fait appel à des valeurs absolues, c'est pourquoi on lui préfère l'écart-type.

4) La variance :

Définition :

La variance est la moyenne arithmétique des carrées des écarts des valeurs de la variable à leur moyenne arithmétique, on la désigne par V(x) ou σ^2.

$$V(x) = \sigma^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{\sum_{i=1}^{k} n_i} \text{ (en fonction des effectifs)}$$

Ou :

$$\sigma^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2 \text{ (en fonction des fréquences)}$$

5) L'écart-type :

Définition :

L'écart-type est la racine carrée positive de la variance.

$$\sigma_x = \sqrt{V(x)} \quad \text{ou} \quad \sigma_x = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^2}{\sum_{i=1}^{k} n_i}}$$

Calcul de la variance et de l'écart-type :

Exemple :

Reprenons l'exemple précédent relatif aux découvertes de céramiques campaniformes au Maroc selon le nombre de fragments découverts :

Nombre de fragments	n_i	x_i	$n_i x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2$
08-10	22	9	198	-3,06	9,36	205,92
10-12	28	11	308	-1,06	1,12	31,36
12-14	30	13	390	0,94	0,88	26,4
14-16	15	15	225	2,94	8,64	129,40
16-18	5	17	85	4,94	24,40	122
Total	100	--	1206	---	---	515,08

Calculons $\overline{x}$:

$$\overline{x} = \frac{\sum_{i=1}^{5} n_i x_i}{\sum_{i=1}^{5} n_i} = \frac{1206}{100} = 12,06$$

$$V(x) = \frac{\sum_{i=1}^{5} n_i (x_i - \overline{x})^2}{\sum_{i=1}^{5} n_i} = \frac{518,08}{100} = 5,15$$

$\sigma^2 = \sqrt{5{,}15} \approx 2{,}27$

6) Le coefficient de variation :

Définition :

Le coefficient de variation noté C.V est défini comme le rapport de l'écart type à la moyenne, il sert à comparer la dispersion de deux séries statistiques dont les valeurs sont de grandeurs différentes. Il a pour formule :

$$\text{C.V} = \frac{\sigma^2}{\overline{x}}$$

C'est un nombre sans dimension, il est toujours exprimé en pourcentage (en %).

Exemple :

A partir de l'exemple précédent relatif aux découvertes de céramiques campaniformes au Maroc selon le nombre de fragments découverts, calculons le coefficient de variation :

$$\text{C.V} = \frac{\sigma^2}{\overline{x}} = \frac{2{,}27}{12{,}06} = 0{,}1882$$

Soit : C.V = 18,82 %

Signification :

Les fragments découverts ne sont pas très dispersés puisque la dispersion est simplement de 18,82%.

7) Les moments :

Définition :

On appelle moment d'ordre r par rapport à une constante x_o (ou centré sur x_o), la quantité :

$$m_r = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - x_o)^r$$

On définit deux sortes de moments :

- ✓ Les moments non centrés ou moments simples, dans ce cas $x_o = 0$ et :

$$m_r = \frac{1}{N}\sum_{i=1}^{k} n_i\,x_i^r = \sum_{i=1}^{k} f_i x_i^r$$

- ✓ Les moments centrés ou moments par rapport à la moyenne arithmétique qu'on note :

$$\mu_r = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^r \text{ pour r = 0,1, …,4}$$

Ces moments sont donnés par les formules figurant dans le tableau suivant :

r	x_o=0 (moments simples ou non centrés)	$x_o=\overline{x}$ (moments centrés)
0	$m_0 = 1$	$\mu_o = 1$
1	$m_1 = \frac{1}{N}\sum_{i=1}^{k} n_i\, x_i = \overline{x}$	$\mu_1 = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - \overline{x}) = 0$
2	$m_2 = \frac{1}{N}\sum_{i=1}^{k} n_i\, x_i^2 = Q^2$	$\mu_2 = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^2 = V(x)$
3	$m_3 = \frac{1}{N}\sum_{i=1}^{k} n_i\, x_i^3$	$\mu_3 = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^3$
4	$m_4 = \frac{1}{N}\sum_{i=1}^{k} n_i\, x_i^4$	$\mu_4 = \frac{1}{N}\sum_{i=1}^{k} n_i\,(x_i - \overline{x})^4$

Exercice d'application :

On donne la durée de vie de 1000 dinosaures découverts dans le village de Tazouda dans les montagnes du Haut-Atlas marocain.

Durée de vie des dinosaures (en 1000 ans)	Nombre de dinosaures
[0 – 0,5[	25
[0,5 – 1[	60
[1 – 1,5[	265
[1,5 – 2[	450
[2 – 2,5[	135
[2,5 – 3[	35
[3 – 3,5[	30
Total	1000

Sachant que des éléphants découverts près de Tan-Tan au Sud-Ouest du Maroc ont une durée de vie moyenne de 1800 ans, et que cette variable s'écarte de sa moyenne, en moyenne, de 108 ans.

On demande de :

1) Comparer les deux coefficients de variation relatifs à la durée de vie des dinosaures et des éléphants.

2) Quel est le pourcentage de l'étendue contenant 80% des observations centrales ?

Solution :

La variable étudiée est la durée de vie des dinosaures et des éléphants.

Les données relatives à la durée de vie des dinosaures sont présentées dans le tableau suivant :

Durée de vie des dinosaures (en 1000 ans)	Nombre de dinosaures n_i	Centres des classes x_i	$n_i\, x_i$	$n_i(x_i - \bar{x})^2$	$n_i c\nearrow$
[0 – 0,5[	25	0,25	6,25	50,23	25
[0,5 – 1[	60	0,75	45,00	50,51	85
[1 – 1,5[	265	1,25	331,25	46,19	350
[1,5 – 2[	450	1,75	787,50	3,06	800
[2 – 2,5[	135	2,25	303,75	45,80	935
[2,5 – 3[	35	2,75	96,25	41,01	970
[3 – 3,5[	30	3,25	97,50	75,13	1000
Total	1000	---	1667,5	311,93	---

1) On calcule la moyenne et l'écart-type :

a- La moyenne :

- Pour les dinosaures :

$$\bar{x}_D = \frac{\sum_{i=1}^{7} n_i x_i}{\sum_{i=1}^{7} n_i} = \frac{1667,5}{1000} = 1,6675$$

Soit $\bar{x}_D$ = 1667,5 ans

En moyenne la durée de vie des dinosaures du Haut-Atlas marocain est de 1667,5 ans.

- Pour les éléphants :

D'après les données de l'exercice on a :

$\bar{x}_E$ = 1800 ans

En moyenne la durée de vie des éléphants découverts près de Tan-Tan au Sud-Ouest du Maroc est de 1800 ans.

b- La variance et l'écart-type :

- Pour les dinosaures :

$$V(x)_D = \frac{\sum_{i=1}^{7} n_i (x_i - \bar{x})^2}{\sum_{i=1}^{7} n_i}$$
$$= \frac{311,93}{1000} = 0,31$$

Par suite :

$$\sigma(x)_D = \sqrt{V(x)D} = \sqrt{0,31} = 0,55677$$

Soit :

$$\sigma(x)_D = 556{,}77 \text{ ans}$$

c- Le coefficient de variation :

- Pour les dinosaures :

$$CV_D = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{0{,}55677}{1{,}6675} \times 100 = 33{,}39\%$$

- Pour les éléphants :

$$CV_E = \frac{\sigma_x}{\bar{x}} \times 100 = \frac{108}{1800} \times 100 = 6\%$$

Puisque $CV_E < CV_D$; la durée de vie des éléphants découverts près de Tan-Tan au Sud-Ouest du Maroc est plus homogènes et moins dispersée que celle des dinosaures du Haut-Atlas marocain.

2) Calcul du pourcentage de l'étendue contenant 80% des observations centrales : Il s'agit de l'intervalle :

$I_D = D_9 – D_1$ qui contient 80% des observations.
Calculons alors D_1 et D_9 (1^{er} et $9^{ème}$ décile).

- Calcul de D_1 :

$$\text{* Rang de } D_1 = \frac{\sum_{i=1}^{7} n_i}{10} = \frac{1000}{10} = 100$$

* Classe contenant D_1 = [1 – 1,5[

* Interpolation linéaire :

$$D_1 = 1 + [(1{,}5 - 1) \times \frac{100-85}{350-85}]$$

$$D_1 = 1 + (0{,}5 \times \frac{15}{265})$$

$$D_1 = 1{,}03$$

- Calcul de D_9 :

$$\text{* Rang de } D_9 = \frac{\sum_{i=1}^{7} n_i \times 9}{10} = \frac{1000 \times 9}{10} = 900$$

* Classe contenant D_9 = [2 – 2,5[

* Interpolation linéaire :

$D_9 = 2 + [(2,5 - 2) \times \frac{900-800}{935-800}]$

$D_9 = 2 + (0,5 \times \frac{100}{135})$

$D_9 = 2,37$

D'où : $I_D = D_9 - D_1 = 2,37 - 1,03 = 1,34$

Soit : I_D = 1340 ans.

Le pourcentage de l'étendue contenant 80% des observations centrales est :

$$\frac{\text{ID}}{E} \times 100 = \frac{1,34}{3,5-0} \times 100 = 38,28\%$$

Paragraphe 3 : Paramètres de forme :

Définition :

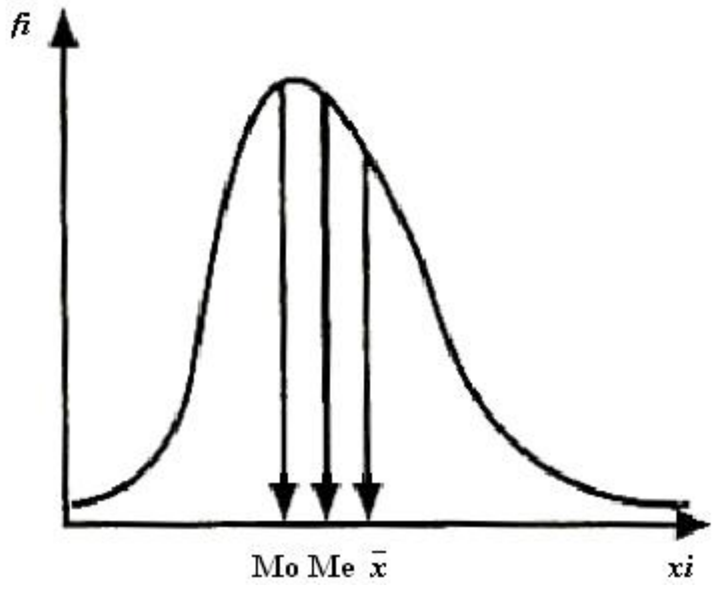
Les paramètres de forme nous renseignent sur la forme d'une distribution statistique, généralement on s'intéresse à la symétrie et à l'aplatissement.

- ✓ Une distribution statistique est dite symétrique si les observations se répartissent de façon égale de part et d'autre de la valeur centrale. Dans le cas contraire, la distribution est dite asymétrique ou dissymétrique.
- ✓ Une distribution est plus ou moins aplatie suivant que les fréquences des valeurs voisines des valeurs centrales sont plus ou moins élevées par rapport aux autres. L'aplatissement d'une courbe de fréquence est jugé par comparaison avec la courbe théorique des fréquences d'une loi dite normale ou loi de Laplace-Gauss.

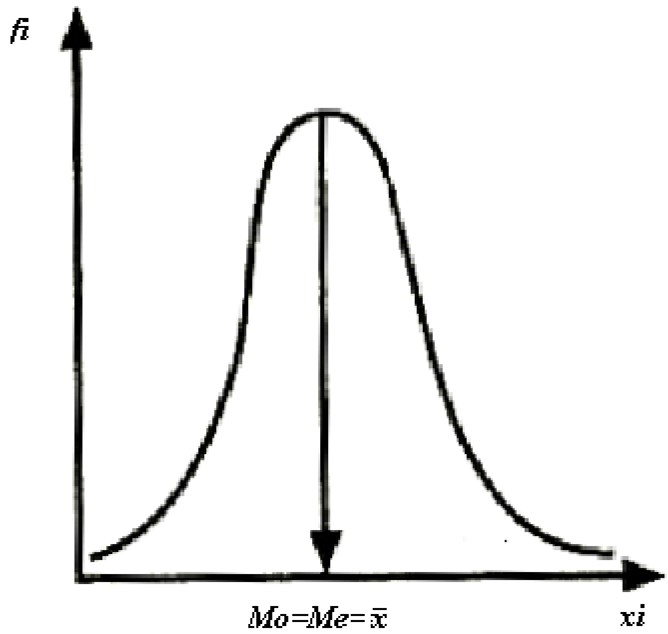
1) Coefficients d'asymétrie :

Une distribution parfaitement symétrique possède les propriétés suivantes :

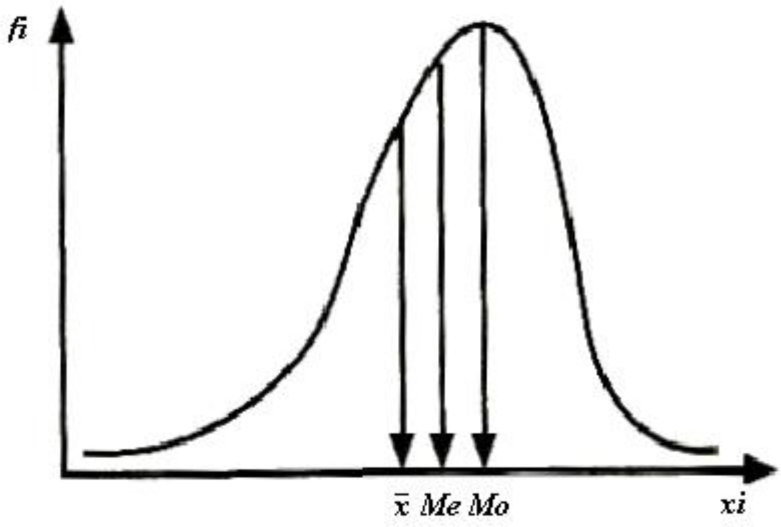
Moyenne = Médiane = Mode



Etalement à droite ou coefficient positif : Dans ce cas : $M_o < M_e < \bar{x}$



Symétrie ou coefficient nul : Dans ce cas : $M_o = M_e = \bar{x}$



Etalement à gauche ou coefficient à gauche : Dans ce cas : $M_o > M_e > \overline{x}$

Les principaux coefficients d'asymétrie sont les suivants :

a- Coefficient de Yule :

$$S_y = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1} \quad \text{ou} \quad S_y = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

Le signe du coefficient nous renseigne sur le comportement de la symétrie, en effet :

- Si $S_y = 0$, on dit que la courbe est symétrique.
- Si $S_y > 0$, on dit que le coefficient est positif et l'étalement est à droite.
- Si $S_y < 0$, on dit que le coefficient est négatif et l'étalement est à gauche.

b- Coefficient de Kelley :

$$S_k = M_e - \frac{D_1 + D_9}{2}$$

De la même manière, le signe du coefficient nous renseigne sur le comportement de la symétrie, en effet :

- Si $S_k = 0$, on dit que la distribution est symétrique.
- Si $S_k > 0$, on dit que le coefficient est positif et la distribution est asymétrique à droite.
- Si $S_k < 0$, on dit que le coefficient est négatif et la distribution est asymétrique à gauche.

c- Coefficient de Fisher :

$$S_F = \frac{\mu_3}{\sigma x^3} = \frac{\frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \overline{x})^3}{\sigma x^3}$$

Avec μ_3 : moment centre d'ordre 3.

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \overline{x})^3.$$

On interprète de la même manière selon le signe du coefficient. En effet :

- Si $S_F = 0$, on dit que la courbe est symétrique.
- Si $S_F > 0$, on dit que le coefficient est positif et l'étalement est à droite.
- Si $S_F < 0$, on dit que le coefficient est négatif et l'étalement est à gauche.

d- Coefficient de Pearson :

$$S_p = \frac{\overline{x} - M_\circ}{\sigma x} \quad \text{ou} \quad S_p = 3\,\frac{\overline{x} - Me}{\sigma x}$$

On interprète de la même manière selon le signe du coefficient. En effet :

- Si $S_p = 0$, on dit que la distribution est symétrique.
- Si $S_p > 0$, on dit que le coefficient est positif et la distribution est asymétrique à droite.
- Si $S_p < 0$, on dit que le coefficient est négatif et la distribution est asymétrique à gauche.

2) Coefficient d'aplatissement :

Pour mesurer le degré d'aplatissement d'une courbe on utilise :

a- Le coefficient de Fisher :

$$P_F = \frac{\mu_4}{\sigma x^4} - 3$$

Ce coefficient s'interprète de la façon suivante :

- Si $P_F < 0 \Rightarrow$ la distribution est plus aplatie que la normale.
- Si $P_F = 0 \Rightarrow$ la distribution est normale.
- Si $P_F > 0 \Rightarrow$ la distribution est moins aplatie que la normale.

b- Le coefficient de Pearson :

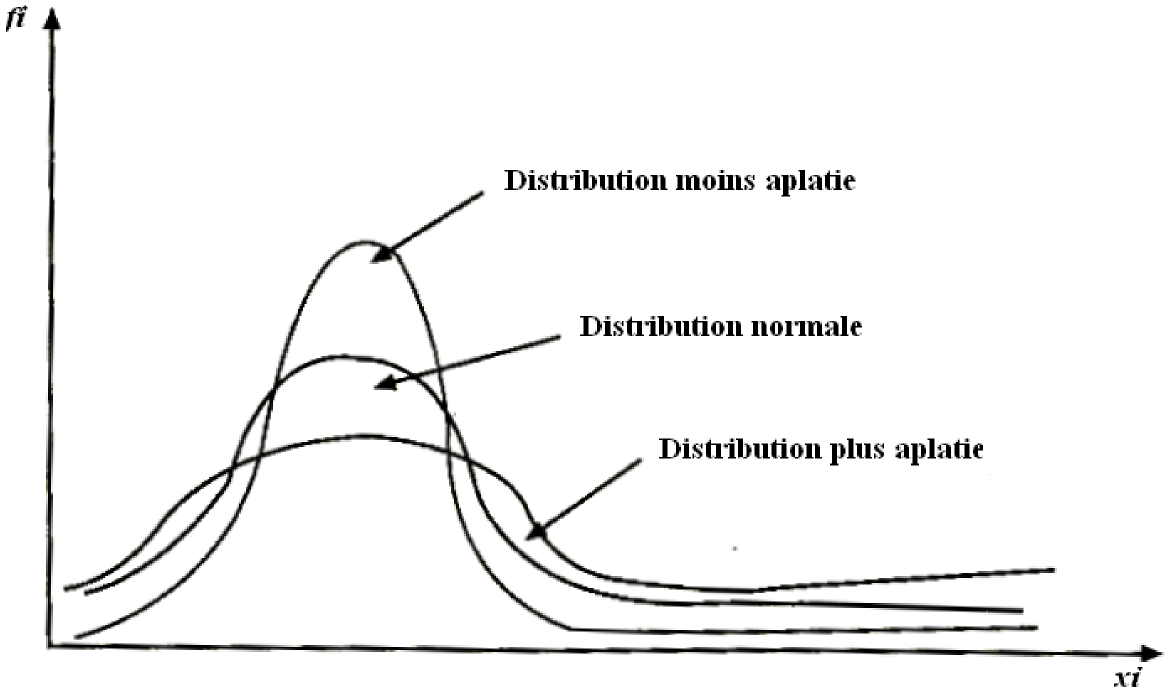
Pearson propose le coefficient suivant :

$$P_p = \frac{\mu_4}{\sigma x^4}$$

Ce coefficient s'interprète de la façon suivante :

- Si $P_p < 3 \Rightarrow$ la distribution est plus aplatie que la normale.
- Si $P_p = 3 \Rightarrow$ la distribution est normale.
- Si $P_p > 3 \Rightarrow$ la distribution est moins aplatie que la normale.

De manière générale la représentation graphique des polygones des fréquences simples pourra revêtir une des trois cas de figures suivantes :



Exemple :

Soit la distribution statistique suivante relative aux pièces de monnaie découvertes dans un site archéologique.

Pièces de monnaie x_i	f_i en %
10-12	04
12-16	20
16-20	30
20-24	20
24-26	08
26-28	06
28-40	12
Total	100

Appréciez la forme de la distribution en question.

x_i	f_i	c_i	$c_i f_i$	$f_i(c_i-\bar{x})^2$	$f_i(c_i-\bar{x})^3$	$f_i(c_i-\bar{x})^4$	$f_i\ c\nearrow$
10-12	0,04	11	0,44	3,79	-36,96	359,99	0,04
12-16	0,20	14	2,8	9,08	-61,24	412,73	0,24
16-20	0,30	18	5,4	2,25	-6,17	16,91	0,54
20-24	0,20	22	4,4	0,32	0,40	0,50	0,74
24-26	0,08	25	2	1,45	6,18	26,35	0,82
26-28	0,06	27	1,62	2,35	14,72	92,14	0,88
28-40	0,12	34	4,08	21,1	279,78	3709,84	1
Total	1,00	--	20,74	40,34	196,71	4618,46	--

L'étude de la forme d'une distribution renvoie à l'étude de deux aspects :

1) L'asymétrie :

a- Le coefficient de Pearson :

$$S_p = \frac{\bar{x} - M_{\circ}}{\sigma x} = \frac{20{,}74-18}{6{,}35} = 0{,}43 > 0$$

Nous avons :

$$\bar{x} = \sum_{i=1}^{7} f_i c_i = 20{,}74$$

et $M_o = 16 + \frac{(0{,}30-0{,}20)\times 4}{(0{,}30-0{,}20)+(0{,}30-0{,}2)}$

$= 16 + \frac{0{,}4}{0{,}20}$

D'où : $M_o = 18$

et $V(x) = \sum_{i=1}^{7} f_i(x_i - \bar{x})^2 = 40{,}34$

d'où : $\sigma_x = \sqrt{V(x)} = 6{,}35$

b- Le coefficient de Fisher :

$$S_F = \frac{\mu_3}{\sigma x^3} = \frac{\sum_{i=1}^{7} f_i(c_i - \bar{x})^3}{\sigma x^3} = \frac{196{,}71}{(6{,}35)^3} \approx 0{,}77 > 0$$

c- Le coefficient de Yule :

$$S_Y = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Calculons Q_1 :

- Rang de $Q_1 = \frac{\sum_{i=1}^{7} f_i}{4} = \frac{1}{4} = 0{,}25$
- Classe contenant $Q_1 = [16 - 20[$
- Interpolation linéaire : $Q_1 = 16 + 4\,\frac{0{,}25-0{,}24}{0{,}54-0{,}24} = 16{,}13$

Calculons Q_2 :

- Rang de $Q_2 = \frac{\sum_{i=1}^{7} f_i}{2} = \frac{1}{2} = 0{,}5$
- Classe contenant $Q_2 = [16 - 20[$
- Interpolation linéaire : $Q_2 = 16 + 4\,\frac{0{,}5-0{,}24}{0{,}54-0{,}24} \approx 19{,}47$

Calculons Q_3 :

- Rang de $Q_3 = \frac{\sum_{i=1}^{7} f_i \times 3}{4} = \frac{3}{4} = 0{,}75$
- Classe contenant $Q_3 = [24 - 26[$
- Interpolation linéaire : $Q_3 = 24 + 2\,\frac{0{,}75-0{,}74}{0{,}82-0{,}74} = 24{,}25$

D'où : $S_Y = \frac{Q_3+Q_1-2Q_2}{Q_3-Q_1} = \frac{24,25+16,13-2(19,47)}{24,25-16,13}$

$S_Y = \frac{1,44}{8,12} = 0,18 > 0$

Interprétation :

Tous les coefficients indiquent une assez importante asymétrie à droite de la distribution statistique étudiée.

2) L'aplatissement :

Calculons le coefficient de Fisher :

$$P_F = \frac{\mu_4}{\sigma x^4} - 3 = \frac{\sum_{i=1}^{7} f_i (c_i - \bar{x})^4}{\sigma x^4} - 3$$

$$P_F = \frac{4618,46}{1625,90} - 3 = -0,16 < 0$$

Interprétation :

La distribution statistique étudiée est aplatie par rapport à la normale.

Section 3 : Caractéristiques de concentration :

L'étude de la concentration permet de caractériser la structure d'un grand nombre d'entités dont les modalités du caractère sont susceptibles d'être additionnées.

On étudie la concentration chaque fois qu'on peut remarquer une disparité ou inégalité de répartition entre les unités de la population considérée.

A- Détermination graphique de la concentration :

1) Courbe de concentration :

La visualisation graphique de la concentration consiste à porter sur un même graphique dans un repère orthonormé :

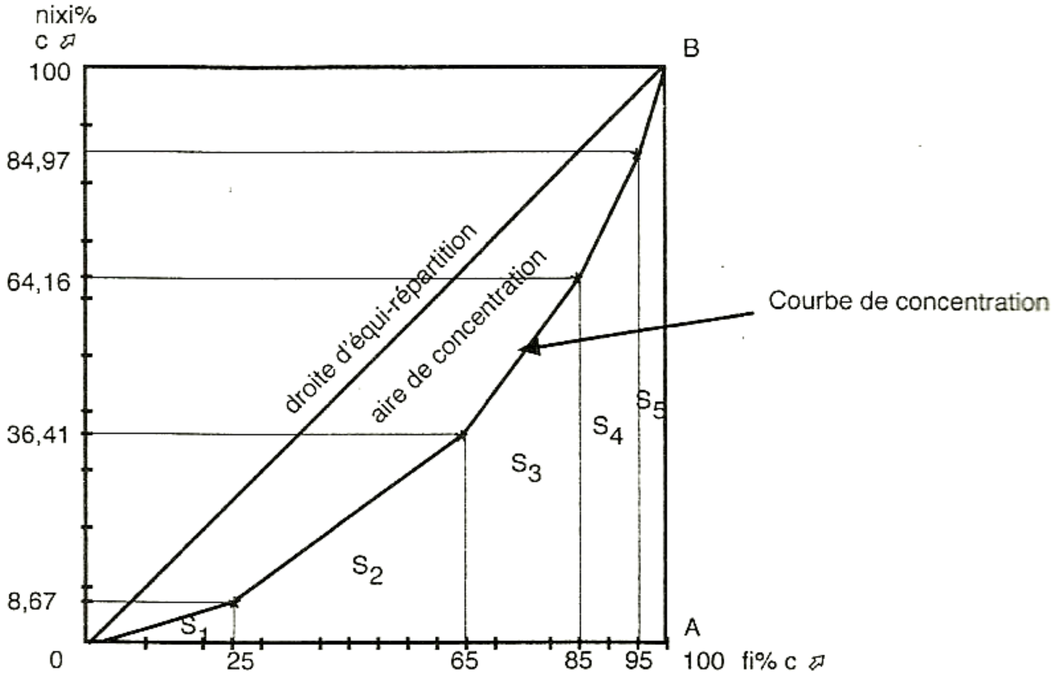
a. En abscisse la fréquence relative cumulée croissante f_i en pourcentage $f_i c \nearrow \%$.

b. En ordonnée, les fréquences relatives cumulées croissantes en pourcentage de la masse totale $n_i x_i c \nearrow \%$.

Exemple :

Soit la distribution suivante relative à la répartition de 80 sites selon le nombre d'outils :

Sites	Nombre d'outils	x_i	f_i %	f_i c↗%	$n_i x_i$	$n_i x_i$ %	$n_i x_i$c↗%	$n_i x_i$ c ↗
10-20	20	15	25	25	300	8,67	8,67	300
20-40	32	30	40	65	960	27,74	36,41	1260
40-80	16	60	20	85	960	27,75	64,16	2220
80-100	8	90	10	95	720	20,81	84,97	2940
100-160	4	130	5	100	520	15,03	100,00	3640
Total	80	---	100	---	3460	100,00	---	---



Interprétation :

La courbe obtenue est dite courbe de concentration ou courbe de Lorentz. Ainsi, plus la courbe de concentration est proche de l'axe des abscisses plus la concentration est forte, et plus la courbe de concentration est proche de la première bissectrice plus la concentration est faible. La concentration est nulle lorsque la courbe se confond avec la diagonale. Il est donc nécessaire de mesurer l'intensité de la concentration par le coefficient de concentration ou indice de Gini.

2) **Indice de Gini :**

L'indice de concentration de Gini est le rapport de l'aire de concentration (la surface délimitée par la courbe de concentration et la première bissectrice) à l'aire du triangle OAB.

$I_c = \frac{Aire\ de\ concentration}{Aire\ du\ triangle\ OAB}$

L'indice de Gini est un nombre sans dimension, toujours compris entre 0 (dans ce cas la concentration est nulle) et 1 (dans ce cas la concentration est totale). Il est exprimé en pourcentage.

L'aire de concentration est la différence entre l'aire du triangle OAB et l'aire S délimitée par la courbe et les deux cotés OA et AB.

$$\text{L'aire du triangle OAB} = \frac{base \times hauteur}{2} = \frac{100 \times 100}{2} = \frac{10000}{2} = 5000$$

L'aire S est égale à la somme des aires du triangle et les différents trapèzes résultant de la construction de la courbe.

$S = S_1 + S_2 + S_3 + S_4 + S_5$

Calculons S_1, S_2, S_3, S_4 et S_5 :

$S_1 = \frac{1}{2}(25 \times 8{,}67) = 108{,}375$
$S_2 = \frac{1}{2}(8{,}67 + 36{,}41)(65 - 25) = 901{,}6$
$S_3 = \frac{1}{2}(36{,}41 + 64{,}16)(85 - 65) = 1005{,}7$
$S_4 = \frac{1}{2}(64{,}16 + 84{,}97)(95 - 85) = 745{,}65$
$S_5 = \frac{1}{2}(84{,}97 + 100)(100 - 95) = 462{,}425$

$S = S_1 + S_2 + S_3 + S_4 + S_5$
$S = 108{,}375 + 901{,}6 + 1005{,}7 + 745{,}65 + 462{,}425$
$S = 3223{,}75$

D'où :

$$I_c = \frac{5000 - 3223{,}75}{5000} = 0{,}355 \quad \Rightarrow \quad I_c = 35{,}5\%$$

Interprétation :
La concentration des outils est relativement faible.

B- Détermination algébrique de la concentration :

1) Détermination de la médiale :

a- Définition :

La médiale partage la somme des $n_i x_i$ en deux parties égales.

b- Calcul de la médiale :

Calculons la médiale à partir de l'exemple précédent relatif à la répartition de 80 sites selon le nombre d'outils :

Calcul de M_{le} : On suit trois étapes :

- Rang de $M_{le} = \frac{\sum_1^5 n_i x_i}{2} = \frac{3460}{2} = 1730$

- Classe contenant M_{le} = [40 - 80[

- L'interpolation linéaire donne :
  $M_{le} = 40 + [40\ (\frac{1730-1260}{2220-1260})]$
  $M_{le} = 40 + (40 \times \frac{470}{960})$
  $M_{le} = 59,58$

  Soit M_{le} = 59,58 outils.

Signification :
La moitié de la masse des outils est partagé par des sites dont le nombre d'outils est inférieur à 59,58 outils et l'autre moitié de la masse des outils est partagé par des sites dont le nombre d'outils est supérieur à 59,58 outils.

2) Détermination de la concentration :

La détermination de la concentration est calculée par l'écart entre la médiale et la médiane, noté par :

$$\Delta M = M_{le} - M_e$$

Cet écart est toujours positif du fait que $M_{le} \geq M_e$ puisque le calcul de la médiale est fait en fonction de la masse des valeurs.

Une fois cet écart calculé, on le compare au domaine de variation ou à l'étendue de la série étudiée. Trois cas sont possibles :

- Lorsque ΔM est supérieur à l'étendue, on dit que la concentration est forte.
- Lorsque ΔM est inférieur à l'étendue, on dit que la concentration est faible.
- Lorsque ΔM est nul, on dit que la médiale M_{le} est égale à la médiane M_e.

Calculons ΔM à partir de l'exemple précédent relatif à la répartition de 80 sites selon le nombre d'outils :

Calcul de M_e : On suit trois étapes :

- Rang de $M_e = \frac{\sum_1^5 f_i}{2} = \frac{100}{2} = 50$
- Classe contenant M_e = [20 - 40[
- L'interpolation linéaire donne :

  $M_e = 20 + [(40 - 20)\ (\frac{50-25}{65\ -25})]$

  $M_e = 20 + (20 \times \frac{25}{40})$

  $M_e = 32{,}5$

Par suite l'écart : $\Delta M = M_{le} - M_e$

$\Delta M = 59{,}58 - 32{,}5$

$\Delta M = 27{,}08$

Calculons l'étendue E :

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

$E = 160 - 10$

$E = 150$

Comparons ΔM à l'étendue E :

$\frac{\Delta M}{E} = \frac{27{,}08}{150} = 0{,}18$ soit 18%

Signification :

D'après cet indicateur, la concentration des outils dans ces sites est donc faible.

Exercices d'application

Exercice n°1 :

Supposons la série statistique suivante relative à la répartition de 40 sites archéologiques selon le nombre d'objets recueillis par site :

0 - 4 - 4 - 1 - 3 - 4 - 1- 1- 5 - 5 - 0 - 0 - 2 - 2 - 3 - 3 - 4 - 4 - 3 - 3 - 6 - 6 - 3 - 3 - 5 - 2 - 2 - 3 - 2 - 2 - 1 - 1- 0 - 6 - 4 - 1 - 1 - 2 - 2 - 5.

1) Préciser le caractère étudié, sa nature et le nombre de modalités.
2) Dresser le tableau statistique.
3) Représenter le diagramme en bâtons des effectifs.
4) Représenter la courbe en escalier des fréquences cumulées croissantes.

Réponse :

1) -Le caractère étudié est le nombre d'objets recueillis par site archéologique.

-La variable étudiée est quantitative discrète.

-Le nombre de modalités prises par le caractère étudié est égal à 7. En effet :

$x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 5$, $x_6 = 6$, $x_7 = 7$.

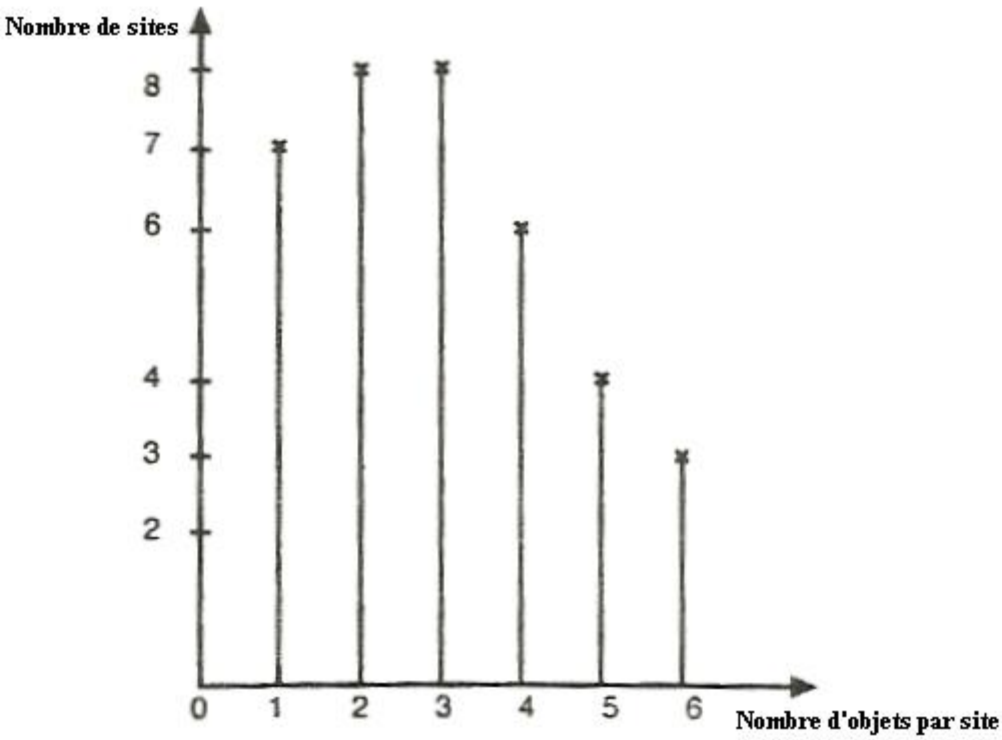
A chacune de ces valeurs correspond un effectif n_i :

$n_1 = 4$, $n_2 = 7$, $n_3 = 8$, $n_4 = 8$, $n_5 = 6$, $n_6 = 4$, $n_7 = 3$.

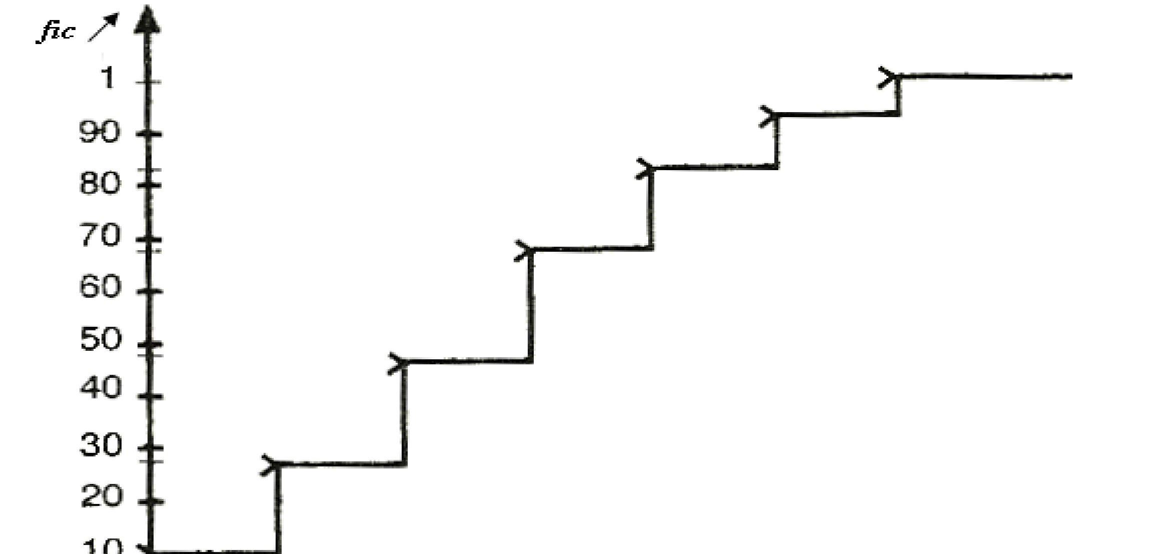
2) Le tableau statistique se présente comme suit :

Nombre d'objets par site x_i	Nombre de sites archéologique n_i	Fréquence f_i en %	Fréquence cumulée croissante $f_i c\nearrow$
0	4	10	10
1	7	17,5	27,5
2	8	20	47,5
3	8	20	67,5
4	6	15	82,5
5	4	10	92,5
6	3	7,5	100
Total	40	100	---

3) Représentation du diagramme en bâtons des effectifs.



4) Représentation de la courbe en escalier des fréquences cumulées croissantes.



Exercice n°2 :

Supposons que le décompte de l'outillage de la grotte Velozzo, près de Dar Bou Azza à Casablanca se présente comme suit :

3 grattoirs, 5 perçoirs, 7 troncatures, 4 microlithes géométriques, 4 éclats et lames à bord abattu, 2 lamelles à bord abattu.

On demande de :

1) Dresser le tableau statistique en calculant les fréquences.
2) Spécifier le caractère étudié.
3) Représenter graphiquement cette distribution.

Réponse :

1) On présente ces données dans un tableau statistique comprenant trois colonnes. La première colonne comporte les modalités du caractère, la deuxième comporte les effectifs et la troisième comporte les fréquences.

Modalités x_i	Nombre d'outils n_i	Fréquence f_i en %
Grattoirs	3	12
Perçoirs	5	20
Troncatures	7	28
Microlithes géométriques	4	16
Eclats et lames à bord abattu	4	16
Lamelles à bord abattu	2	8
Total	25	100

A chaque modalité *i* correspond un effectif noté n_i. L'effectif total est :

$$N = \sum_{i=1}^{6} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$$
$$N = 3 + 5 + 7 + 4 + 4 + 2 = 25 \text{ outils.}$$

A chaque modalité *i* correspond une fréquence notée f_i.

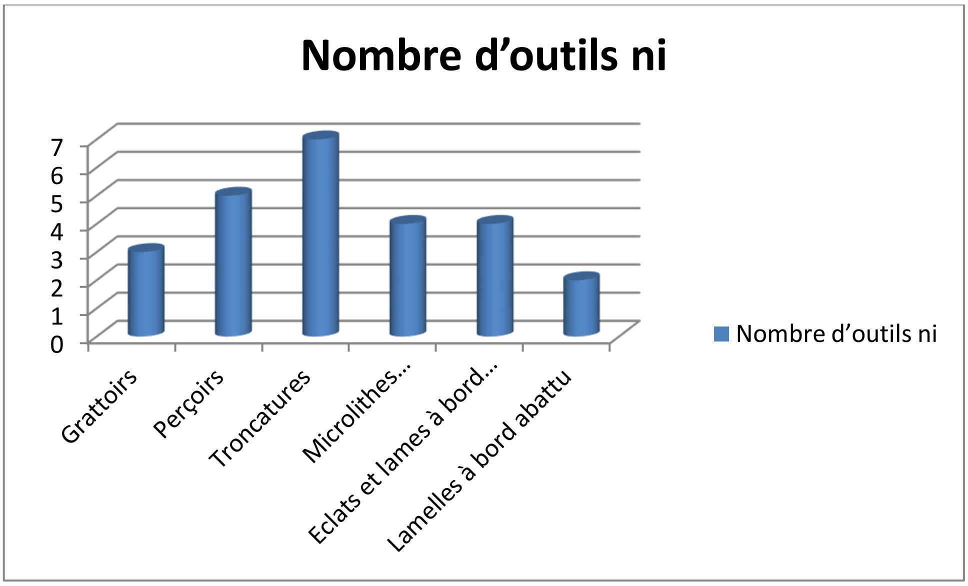
$$f_i = \frac{ni}{N} \times 100$$

$$\begin{aligned}\sum_{i=1}^{6} fi &= f_1 + f_2 + f_3 + f_4 + f_5 + f_6 \\ &= \left(\frac{n_1}{N} + \frac{n_2}{N} + \frac{n_3}{N} + \frac{n_4}{N} + \frac{n_5}{N} + \frac{n_6}{N}\right) \times 100 \\ &= \left(\frac{3}{25} + \frac{5}{25} + \frac{7}{25} + \frac{4}{25} + \frac{4}{25} + \frac{2}{25}\right) \times 100 \\ &= 12 + 20 + 28 + 16 + 16 + 8 \\ &= 100\%\end{aligned}$$

2) Le caractère étudié est le décompte de l'outillage de la grotte Velozzo.

3) Cette distribution peut être représentée par un histogramme comme elle peut être représentée par un diagramme circulaire.

L'histogramme peut être représenté en portant sur l'axe des abscisses les modalités x_i, et sur l'axe des ordonnées des rectangles dont la hauteur est proportionnelle à l'effectif n_i correspondant à chaque modalité.



Pour représenter le diagramme circulaire, on partage un cercle en secteurs dont les aires sont proportionnelles aux pourcentages donnés.

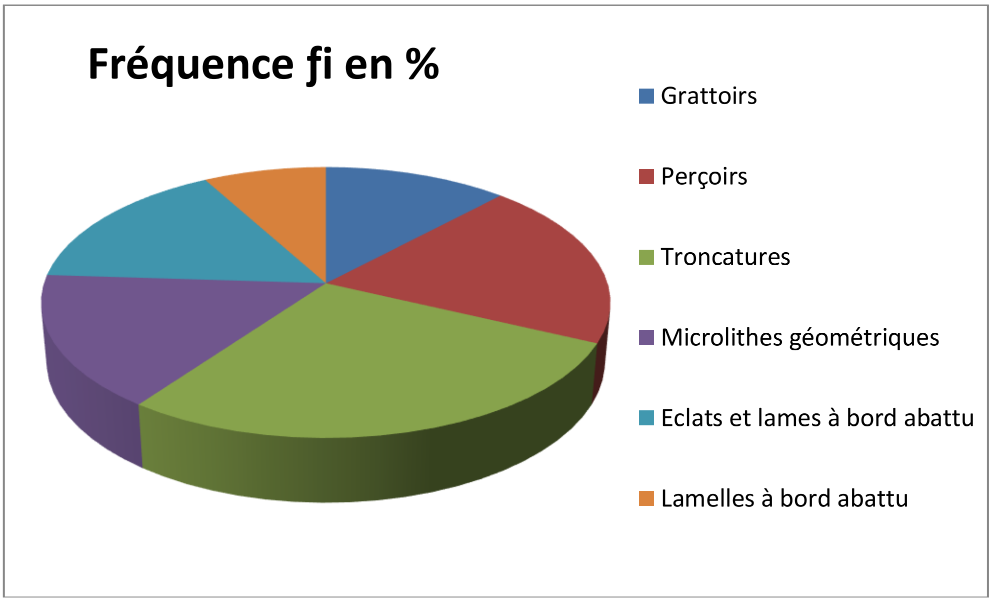
A chaque modalité du caractère correspond un secteur angulaire ou angle au centre appelé α_i dont la mesure en degrés est : $\alpha_i = 360° \times f_i$.

$\alpha_1 = 360° \times 0{,}12 = 43{,}2$ $\alpha_3 = 360° \times 0{,}28 = 100{,}8$ $\alpha_5 = 360° \times 0{,}16 = 57{,}6$

$\alpha_2 = 360° \times 0{,}20 = 72$ $\alpha_4 = 360° \times 0{,}16 = 57{,}6$ $\alpha_6 = 360° \times 0{,}08 = 28{,}8$

Après avoir calculé les angles α_i, on s'assure que leur somme est bien égale à 360° puis on trace le diagramme circulaire.

$$\begin{aligned}\sum_{i=1}^{6} \alpha_i &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 \\ &= 43{,}2 + 72 + 100{,}8 + 57{,}6 + 57{,}6 + 28{,}8 \\ &= 360°\end{aligned}$$



Exercice n°3 :

Lors d'une fouille archéologique, un archéologue a détecté 10 pièces de monnaie circulaire ayant pour rayon :

- 2 cm pour les 4 premiers.
- 3 cm pour les 3 suivantes.
- 4 cm pour les 3 dernières.

1) Préciser la variable étudiée et sa nature.
2) Dresser le tableau statistique.
3) Calculer le rayon moyen, le rayon modal et le rayon médian.

Réponse :

1) La variable étudiée est le rayon de la pièce de monnaie (en cm).
   Sa nature est quantitative discrète.
2) Le tableau statistique relatif à cette distribution se présente comme suit :

Rayon x_i en cm	Nombre de pièces n_i	$n_i\, x_i$
2	4	8
3	3	9
4	3	12
Total	10	29

3) - Calcul du rayon moyen :

$$\bar{x} = \frac{\sum_{i=1}^{3} n_i\, x_i}{\sum_{i=1}^{3} n_i} = \frac{29}{10} = 2{,}9 \text{ cm.}$$

- Calcul du rayon modal :

Dans le cas d'une variable discrète le mode correspond à la valeur qui a le plus grand effectif.

$$M_o = 2 \text{ cm}$$

- Calcul du rayon médian :

Dans le cas d'une variable discrète et lorsque le nombre d'observation est pair ; c'est-à-dire :

$$N = 2\,n = 10 \quad \Rightarrow \quad n = \frac{10}{2} = 5 \quad \text{et} \quad n + 1 = 6$$

Essayons d'étaler les 10 pièces de monnaie :

$$2,\ 2,\ 2,\ 2,\ \boxed{3},\ \boxed{3},\ 3,\ 4,\ 4,\ 4.$$

D'où :

$$M_e = \frac{3+3}{2} = \frac{6}{2} = 3 \text{ cm.}$$

Exercice n°4 :

Pour tester les connaissances des 20 étudiants de la 2ème année INSAP en céramique islamique, on a proposé un contrôle dont les notes obtenues sont sur 20 et se présentent comme suit :

11-9-10-13-6-7-11-11-5-7-10-11-14-15-9-12-12-11-5-12.

1) Préciser la variable étudiée, sa nature et le nombre de ses modalités.
2) Dresser le tableau statistique.
3) Déterminer la note moyenne, la note modale et la note médiane.
4) Quel est le pourcentage des étudiants ayant moins de 10 ?
5) Quel est le pourcentage des étudiants ayant au moins 9 et pas plus de 12 ?

Réponse :

1) La variable étudiée est la note obtenue en céramique islamique. Sa nature est quantitative discrète, composée de 10 modalités.

2) Le tableau statistique relatif à cette distribution se présente comme suit :

Notes x_i sur 20	Etudiants n_i	$n_i\, x_i$
5	2	10
6	1	6
7	2	14
9	2	18
10	2	20
11	5	55
12	3	36
13	1	13
14	1	14
15	1	15
Total	20	201

3) - Calcul de la note moyenne :

$$\bar{x} = \frac{\sum_{i=1}^{10} n_i\, x_i}{\sum_{i=1}^{10} n_i} = \frac{201}{20} = 10{,}05/20.$$

- Calcul de la note modale :

L'effectif le plus élevé est 5 et correspond à la modalité 11, donc :

$$M_o = 11/20.$$

- Calcul de la note médiane :

Nous sommes dans le cas d'une variable discrète où le nombre de modalité est pair :

$$N = 2\,n = 20 \qquad \Rightarrow \qquad n = \frac{20}{2} = 10 \qquad \text{et} \qquad n + 1 = 11$$

On expose les notes obtenues par les 20 étudiants :

5-5-6-7-7-9-9-10-10-$\boxed{11}$-$\boxed{11}$-11-11-11-12-12-12-13-14-15.

Par suite :

$$M_e = \frac{11+11}{2} = \frac{22}{2} = 11/20$$

4) Calculons le pourcentage des étudiants ayant moins de 10/20 :

$$P = \frac{2+1+2+2}{20} \times 100 = \frac{7}{20} \times 100 = 35\%.$$

5) Calculons le pourcentage ayant au moins 9 et pas plus de 12 :

$P = \frac{2+2+5+3}{20} \times 100 = \frac{12}{20} \times 100 = 60\%.$

Exercice n°5 :

Lors d'une fouille archéologique, on a détecté un certain nombre de mobiliers archéologiques selon l'âge et le type de mobilier mentionnés dans le tableau suivant :

Age (en 100 ans)	Mobiliers archéologiques	
	Fragments	Pièces
05-10	10	05
10-15	15	09
15-20	20	11
20-25	25	13
25-30	30	22
Total	100	60

1) Calculer l'âge moyen des fragments et l'âge moyen des pièces.
2) Calculer les 2 variances.
3) Calculer les 2 écarts-types.
4) Calculer les 2 coefficients de variation.

Réponse :

1) Avant de calculer l'âge moyen des fragments et l'âge moyen des pièces, on dresse d'abord le tableau statistique suivant :

Age	Fragments n_f	Pièces n_p	Centre de classes x_i	$n_f\, x_i$	$n_p\, x_i$
[05,10[	10	5	7,5	75	37,5
[10,15[	15	9	12,5	187,5	112,5
[15,20[	20	11	17,5	350	192,5
[20,25[	25	13	22,5	562,5	292,5
[25,30[	30	22	27,5	825	605
Total	100	60	---	2000	1240

- Calcul de l'âge moyen des fragments :

$\bar{x}_f = \frac{\sum_{i=1}^{5} nfx_i}{\sum_{i=1}^{5} nf} = \frac{2000}{100} = 20$ Soit : $\bar{x}_f = 20 \times 100 = 2000$ ans.

L'âge moyen des fragments est de 2000 ans.

- Calcul de l'âge moyen des pièces :

$\bar{x}_p = \frac{\sum_{i=1}^{5} npx_i}{\sum_{i=1}^{5} np} = \frac{1240}{60} = 20{,}67$ Soit : $\bar{x}_p = 20{,}67 \times 100 = 2067$ ans.

L'âge moyen des pièces est de 2067 ans.

2) De la même manière on dresse d'abord le tableau statistique suivant avant de calculer la variance des fragments et la variance des pièces :

Age	Fragments n_f	Pièces n_p	$(x_i-\bar{x}_f)^2$	$n_f(x_i-\bar{x}_f)^2$	$(x_i-\bar{x}_p)^2$	$n_p(x_i-\bar{x}_p)^2$
[05,10[	10	5	156,25	1562,5	173,45	867,25
[10,15[	15	9	56,25	843,75	66,75	600,75
[15,20[	20	11	6,25	125	10,05	110,55
[20,25[	25	13	6,25	156,25	3,35	43,55
[25,30[	30	22	56,25	1687,5	46,65	1026,3
Total	100	60	---	4375	---	2648,4

- Calcul de la variance des fragments :

$$V_{xf} = \frac{\sum_{i=1}^{5} nf(x_i - \bar{x}f)2}{\sum_{i=1}^{5} nf} = \frac{4375}{100} = 43,75$$

- Calcul de la variance des pièces :

$$V_{xp} = \frac{\sum_{i=1}^{5} np(x_i - \bar{x}p)2}{\sum_{i=1}^{5} np} = \frac{2648,4}{60} = 44,14$$

3) Calcul des écarts-types :

- Ecart-type des fragments :

$$\sigma_{xf} = \sqrt{V(x)f} = \sqrt{43,75} = 6,61$$

- Ecart-type des pièces :

$$\sigma_{xp} = \sqrt{V(x)p} = \sqrt{44,14} = 6,64$$

4) Calcul des deux coefficients de variation :

- Coefficient de variation des fragments :

$$CV_f = \frac{\sigma x\mathrm{f}}{\bar{x}\mathrm{f}} \times 100 = \frac{6,61}{20} \times 100 = 33,05\%$$

- Coefficient de variation des pièces :

$$CV_p = \frac{\sigma x\mathrm{p}}{\bar{x}p} \times 100 = \frac{6,64}{20,67} \times 100 = 32,12\%$$

L'âge des pièces est moins dispersée que l'âge des fragments du fait que :

$CV_p < CV_f$

Exercice n°6 :

On donne les sites et gisements préhistoriques du Maroc selon les découvertes des gravures rupestres.

Gravures rupestres	4-10	10-20	20-30	30-70	70-160	160-500
Sites et gisements	400	200	100	40	20	10

1) Calculer les paramètres centraux.
2) Etudier la forme de cette distribution en calculant les coefficients Y_1 et Y_2.
3) Tracer la courbe de concentration et calculer l'indice de GINI.

Réponse :

1) Avant de calculer les paramètres centraux on dresse le tableau statistique relatif à cette distribution.

Gravures rupestres	x_i	n_i	f_i%	f_i% x_i	f_i% cor	$f_i x_i$ %	f_i%c ↗	$f_i x_i$%c ↗	f_i% x_i^2	$f_i\, x_i^3$	$f_i\, x_i^4$
4-10	7	400	52	364	52	17,7	52	17,7	2548	178,36	1248,52
10-20	15	200	26	390	15,6	18,9	78	36,6	5850	877,5	13162,5
20-30	25	100	13	325	7,8	15,8	91	52,4	8125	2031,25	50781,25
30-70	50	40	5,2	260	0,78	12,6	96,2	65	13000	6500	325000
70-160	115	20	2,5	287,5	--	14	98,7	79	33062,5	38021,87	4372515,6
160-500	330	10	1,3	429	--	21	100	100	141570	467181	154169730
Total	---	770	100	2055,5	--	100	---	---	204155,5	514789,99	158932437,87

a. Calcul de la moyenne :

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i\% x_i}{100} = \frac{2055,5}{100} = 20,555 \text{ gravures rupestres.}$$

b. Calcul du mode : Lorsque les classes sont d'*amplitudes inégales*, il faut procéder au calcul des fréquences corrigées pour assurer leur proportionnalité par rapport aux aires des rectangles correspondants. On prend généralement l'amplitude la plus faible comme valeur de référence.

Si l'on désigne par :

- f_i l'effectif de la classe i,
- a_i son amplitude,
- a_o l'amplitude de référence.

La fréquence corrigée f_{ci} sera donnée par :

$$f_{ci} = \frac{f_i}{a_i} \times a_0$$

Application numérique :

$$f_{c_1} = \frac{f_1}{a_1} \times a_o = \frac{52}{6} \times 6 = 52$$

$$f_{c_2} = \frac{f_2}{a_2} \times a_o = \frac{26}{10} \times 6 = 15{,}6$$

$$f_{c_3} = \frac{f_3}{a_3} \times a_o = \frac{13}{10} \times 6 = 7{,}8$$

$$f_{c_4} = \frac{f_4}{a_4} \times a_o = \frac{5{,}2}{40} \times 6 = 0{,}78$$

$$f_{c_5} = \frac{f_5}{a_5} \times a_o = \frac{2{,}5}{90} \times 6 \approx 0$$

$$f_{c_6} = \frac{f_6}{a_6} \times a_o = \frac{1{,}3}{340} \times 6 \approx 0$$

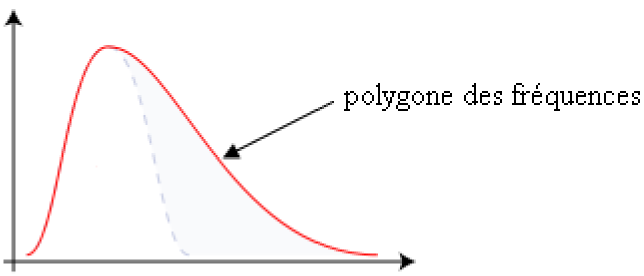
La classe modale dont la fréquence est la plus élevée est la classe [4 - 10[. En utilisant les fréquences en pourcentage corrigées on obtient :

$$M_o = 4 + [\frac{52-15{,}6}{(52-15{,}6)+(52-7{,}8)} \times 6] = 6{,}7096 \text{ gravures rupestres.}$$

c. Calcul de la médiane :

- Rang de la médiane : $\frac{\sum_{i=1}^{6} f_i\%}{2} = \frac{100}{2} = 50\%$
- La classe médiane est [4 -10[.
- $M_e = 4 + [6 \times \frac{50-0}{52-0}] = 9{,}769$ gravures rupestres.

On constate que : $6{,}7096 < 9{,}769 < 20{,}555$ c'est-à-dire que : $M_o < M_e < \bar{x}$ ce qui montre que la distribution des gravures rupestres est étalée à droite graphiquement.



Etalement à droite

2) Etude de la forme :

a- Le coefficient d'asymétrie de Fischer :

Calculons les moments centrés μ_r et les moments simples m_r :

$\mu_r = \sum_{i=1}^{6} f_i(x_i - \bar{x})^r$ et $m_r = \sum_{i=1}^{6} f_i x_i^r$ or $\bar{x} = m_1$

d'où : $\mu_r = \sum_{i=1}^{6} f_i(x_i - m_1)^r$

$m_1 = 20{,}55$; $\mu_2 = m_2 - m_1^2 = 2041{,}55 - (20{,}55)^2 = 1619{,}24$
$m_3 = 514789{,}99$; $m_4 = 158932437{,}87$
$\mu_3 = m_3 - 3m_2\, m_1 + 2\, m_1^3$
$= 514789{,}99 - 3 \times 2041{,}55 \times 20{,}55 + 2 \times (20{,}55)^3$
$= 514789{,}99 - 125861{,}55 + 17356{,}63$
$= 406285$
$\sigma_x = \sqrt{v(x)} = \sqrt{\mu_2} = \sqrt{1619{,}24} = 40{,}24$

Par suite : $\sigma_x^3 = (40{,}24)^3 = 65158{,}92$

Le coefficient d'asymétrie est : $Y_1 = \frac{\mu_3}{\sigma^3} = \frac{406285}{65158{,}92} = 6{,}23$

$Y1 > 0$ la courbe est donc étalée à droite.

b- Le coefficient d'aplatissement :

$Y_2 = \frac{\mu_4}{\sigma^4} - 3$ ou $Y_2 = \frac{\mu_4}{(\mu_2)^2} - 3$

Pour calculer le coefficient d'aplatissement on aura besoin de μ_4 :

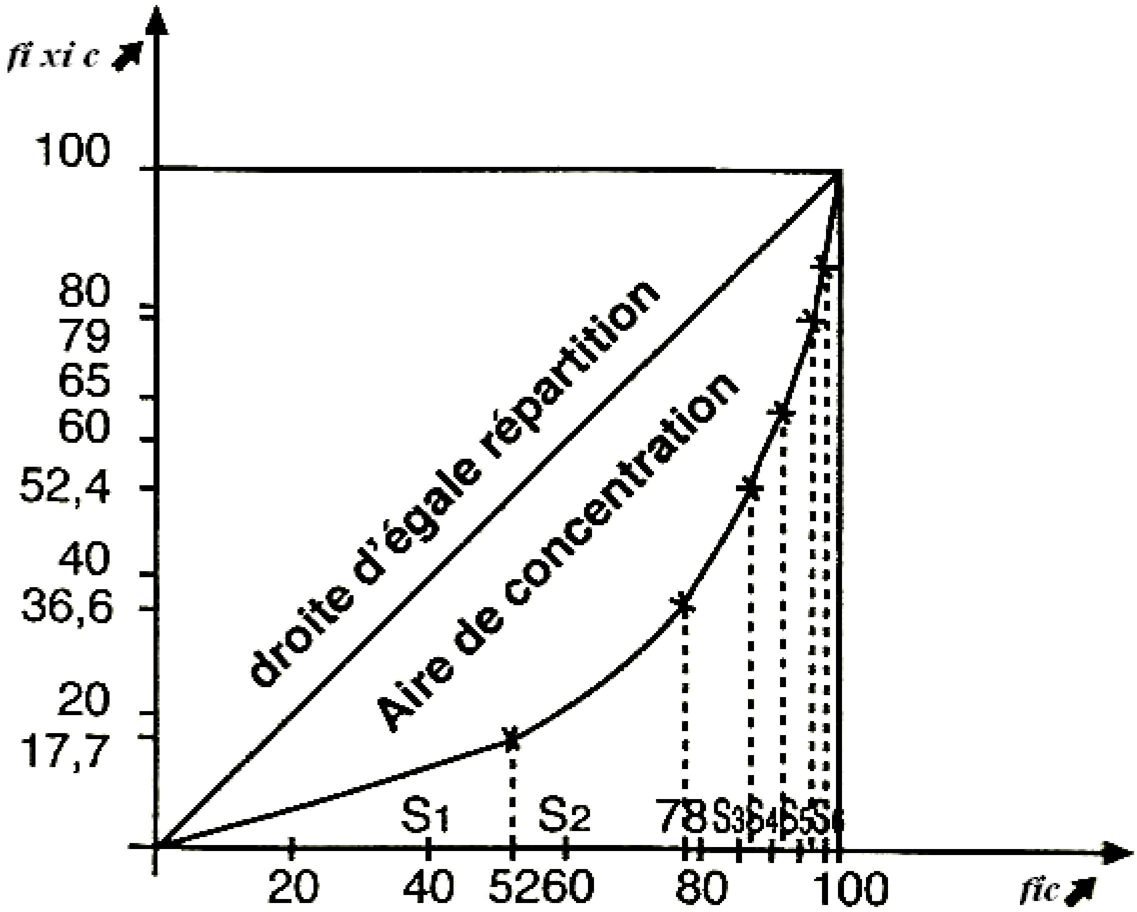
$\mu_4 = m_4 - 4\, m_3\, m_1 + 6\, m_1^2\, m_2 - 3\, m_1^4$
$= 158932437{,}87 - 4 \times 514789{,}99 \times 20{,}55 + 6 \times (20{,}55)^2 \times 2041{,}55 - 3 \times (20{,}55)^4$
$= 158932437{,}87 - 42315737{,}17 + 5172910 - 535018{,}20$
$= 121254592{,}5$

Par suite : $Y_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{(\mu_2)^2} - 3$
$Y_2 = \frac{121254592{,}5}{(40{,}24)^4} - 3 = \frac{121254592{,}5}{(1619{,}29)^2} - 3$
$Y_2 = 43{,}24$

$Y2 > 0$ la courbe est donc moins aplatie que la normale.

3) La courbe de concentration :

Pour tracer la courbe de concentration on porte sur l'axe des abscisses les fréquences en pourcentage cumulées croissantes f_i % c ↗ et en ordonnés on porte les $f_i x_i$ % c ↗.



On calcule les surfaces S_i situées entre la courbe et l'axe OX :

$S_1 = (52 \times 17,7) \times \frac{1}{2} = 460,2$

$S_2 = (17,7 + 36,6) \times \frac{26}{2} = 705,9$

$S_3 = (36,6 + 52,4) \times \frac{13}{2} = 578,5$

$S_4 = (52,4 + 65) \times \frac{5,2}{2} = 305,24$

$S_5 = (65 + 79) \times \frac{2,5}{2} = 180$

$S_6 = (79 + 100) \times \frac{1,3}{2} = 116,35$

$\sum_{i=1}^{6} S_\mathrm{i}$ = $S_1 + S_2 + S_3 + S_4 + S_5 + S_6$

= 460,2 + 705,9 + 578,5 + 305,24 + 180 + 116,35

= 2346,19

L'indice de Gini : $I_G = \frac{5000 - \sum_{i=1}^{6} S_\mathrm{i}}{5000} \times 100 = \frac{5000 - 2346,19}{5000} \times 100 = 53,07\%$

La concentration est donc relativement moyenne.

Bibliographie

- Bulletin d'archéologie marocaine. Tome XXII, 2012.
- Bulletin d'archéologie marocaine. Tome XXI, 2009.
- Bulletin d'archéologie marocaine. Tome XX, 2004.
- Bulletin d'archéologie marocaine. Tome XIX, 2002.
- Bulletin d'archéologie marocaine. Tome XVIII, 1998.
- Bulletin d'archéologie marocaine. Tome XVII, 1987-1988.
- Bulletin d'archéologie marocaine. Tome XV, 1983-1984.
- Bulletin d'archéologie marocaine. Tome XIV, 1981-1982.
- Bulletin d'archéologie marocaine. Tome XII, 1979-1980.
- Bulletin d'archéologie marocaine. Tome XI, 1978.
- Bulletin d'archéologie marocaine. Tome IX, 1973-1975.
- Bulletin d'archéologie marocaine. Tome X, 1976.
- Bulletin d'archéologie marocaine. Tome II, 1957.
- B. Piganiol : Statistique et économétrie. Mémentos Dalloz, 2ème édition 1978.
- E.A.Hassini, B. Kouhlani et M. Benabdesselam : Statistique descriptive : cours et exercices corrigés. 1ère édition 2004.
- Emile Amzallag et Norbert Piccioli : Introduction à la statistique. Exercices corrigés avec rappels détaillés de cours et exemples. Collection « Méthodes » 1978.
- Henri Immediato : cours de statistique. Licence scientifique.
- Michel Janvier : Statistique descriptive avec ou sans tableur. Cours et exercices corrigés. Dunod, Paris, 1999.
- Omar Rajâa : Manuel de statistique descriptive. 1ère édition 2001.
- Philippe Lazar et Daniel Schwartz : Eléments de probabilités et statistique. Médecine-Sciences, Flammarion, 4ème édition 1989.
- Rachid Boutti : Mémento pratique : Statistiques. Diagnostic et prise de décision. Exercices corrigés + Cas professionnels expliqués. Collection expertise, 1ère édition 1996.