

Université Mohamed Premier

STATISTIQUE DESCRIPTIVE

Première année sciences économiques et gestion
Semestre 1

Professeur Moustapha Faizi

Enseignant à la Faculté des Sciences Juridiques,
Économiques et Sociales d'Oujda

Année Universitaire :
2021-2022

Avant-propos

La STATISTIQUE DESCRIPTIVE est le premier cours de statistique qui est programmé pour votre formation au sein de la filière sciences économiques et gestion. On verra trois chapitres, le premier est le plus important, il va s'intéresser à la description de données issues d'une seule variable (l'âge d'un client, prix d'un bien, indemnisation d'un assuré...) observée dans une population (ensemble de clients ou de biens ou d'assurés d'une compagnie). Un deuxième chapitre sera dédié à l'étude de deux variables, on cherchera la possibilité d'une liaison, on proposera le modèle classique de la régression linéaire. Le dernier chapitre donnera les principaux indices statistiques utilisés dans l'évaluation des taux de variations des prix et des volumes.

Des séries d'exercices solutionnées seront proposées parallèlement au cours pour compléter l'apprentissage.

La plupart des calculs qui sont faits dans ce cours sont réalisés avec le logiciel Excel de Microsoft.

Notations et abréviations

n_i	<i>Effectif</i>
N_i	<i>Effectif cumulé</i>
f_i	<i>Fréquence</i>
F_i	<i>Fréquence cumulée</i>
Mo	<i>Mode</i>
Me	<i>Médiane</i>
μ	<i>Moyenne d'une population</i>
$\bar{x}$ ou $\bar{x}_n$	<i>Moyenne d'un échantillon</i>
σ^2	<i>Variance d'une population</i>
s^2 ou s_n^2	<i>Variance d'un échantillon</i>
σ	<i>Écart-type d'une population</i>
s ou s_n	<i>Écart-type d'un échantillon</i>

Sommaire

Chapitre 1: Analyse d'une série de donnée univariée	4
1. Introduction	4
1.1 Exemple	4
2. Terminologie liée à la statistique :	6
3. Présentation des résultats : Tableau-Synthèse	10
3.1 Cas d'un caractère qualitatif	10
3.2 Cas d'un caractère quantitatif discret	11
3.3 Cas d'un caractère quantitatif continu	11
3.3.1 Procédé de construction des classes dans le cas continu	12
4. Représentations graphiques	17
4.1 Diagramme à bandes rectangulaires	17
4.2 Diagramme à secteurs circulaires	19
4.3 Histogramme	20
4.4 Polygone	21
4.5 L'ogive	22
5. Calcul des mesures caractéristiques	23
5.1 Mesures de tendance centrale	24
5.1.1 Le mode	24
5.1.2 La médiane	24
5.1.3 La moyenne arithmétique	28
5.1.4 Autres moyennes	33
5.2 Mesures de dispersion	36
5.2.1 L'étendue	37
5.2.2 La variance	37
5.2.3 L'écart type	39
5.2.4 Écart absolu moyen	39
5.2.5 Coefficient de variation	40
5.3 Mesures de position	41
5.3.1 Les quantiles	41
5.3.2 Intervalle et écart interquartile	42
5.3.3 Construction d'un diagramme en boîte à moustache (Box plot)	44
5.3.4 La cote Z	45
5.4 Mesures de forme	47
5.4.1 Représentation des données selon une courbe normale	47
5.4.2 Coefficients d'asymétrie et d'aplatissement	48
5.5 Mesures de concentration	49
5.5.1 Courbe de Lorenz (1876-1959)	50
5.5.2 Indice de Gini	51
Bibliographie	53

Chapitre 1: ANALYSE D'UNE SÉRIE DE DONNÉE UNIVARIÉE

1. Introduction

Nous nous intéressons ici à l'étude de données d'une série statistique, cette série peut être le résultat d'une enquête sur un phénomène de l'économie ou plus généralement des sciences sociales. Enquêter ici est dans le sens où on cherche des informations auprès d'une population. Le statisticien ou l'économiste statisticien doit travailler avec les données de l'enquête afin de les résumer par des méthodes statistiques¹. La manipulation des données ou des statistiques est quelques fois difficile, il faudra être attentif au choix des méthodes statistiques. Les interprétations des résultats de l'enquête doivent être soigneusement écrites sans distraction ou manipulation. Voici un exemple qui illustre cela.

1.1 Exemple

Dans un établissement universitaire, on se pose la question si la sélection des étudiants à l'inscription est à l'avantage des hommes ou non ?

Sur l'ensemble de l'établissement, on a les résultats suivants :

	H	F
Candidats	1198	449
Sélectionnés	534	113
Proportion	44.6%	25.2%

Conclusion : la sélection est à l'avantage des hommes

On reprend les résultats selon les deux départements A et B :

	H	F
Département A	512/825	89/108
Proportion	62%	82.4%
Département B	22/373	24/341
Proportion	6%	7%

Conclusion : la sélection est à l'avantage des femmes

- Les deux démarches sont correctes
- Rien ne permet de dire que l'une est plus raisonnable que l'autre !

Il faut distinguer la bonne interprétation à travers la bonne méthode

Définition de la STATISTIQUE

¹ Un économiste statisticien est considéré comme un métier, il peut être chargé des prévisions des ventes ou d'achats, pour cela, il doit être capable de mener des études de marché en utilisant les outils statistiques

La statistique est une méthode scientifique dont l'objet est

- de *recueillir*,
- d'organiser et de résumer
- et d'*analyser*

les données d'une enquête, d'une étude ou d'une expérience, puis

- de *tirer* les conclusions logiques
- et de *prendre les décisions* qui s'imposent à partir des analyses effectuées.

EXEMPLE :

Le gouvernement marocain veut réorienter sa politique sociale. Il demande au HCP (HAUT-COMMISSARIAT AU PLAN) la structure de la population :

- âge
- revenu
- catégorie socioprofessionnelle
- état civil
- etc.

Le HCP doit procéder aux étapes suivantes :

1. Recueillir les informations utiles
2. Organiser ces données
3. Les résumer si nécessaire

1. La collecte de données :

De quelle information avons-nous besoin ?

- C'est l'objet de l'étude

Qui doit-on examiner ?

- Qui est/sont le/s sujet/s (Population, échantillon)

Comment faire ?

- C'est la méthode (sondages, entretiens)

2. Organisation des données : nous *présentons* ici les données sous forme de tableaux ou de graphiques

Partons d'un échantillon de 1000 personnes, nous avons donc 1000 revenus.

Nous pouvons présenter ces chiffres par le biais de

- Tableau de synthèse
- Histogramme

Et si l'information est trop riche, on ne tire que des données caractéristiques, c.à.d.

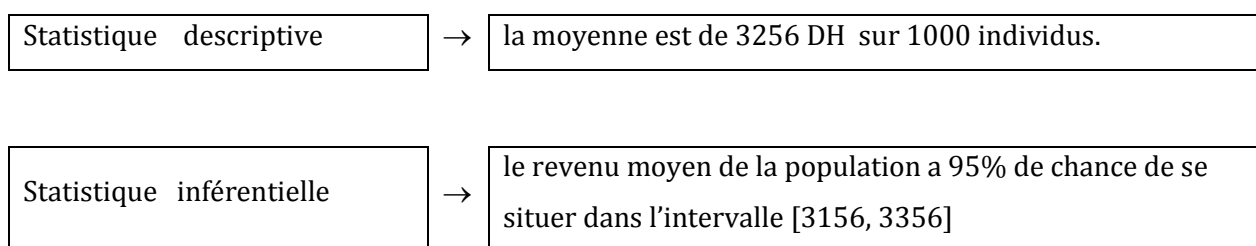
3. Résumer l'information : On n'en choisi que quelques paramètres significatifs (par exemple mode, médiane, moyenne, etc)

Ces trois points sont du domaine de la statistique descriptive.

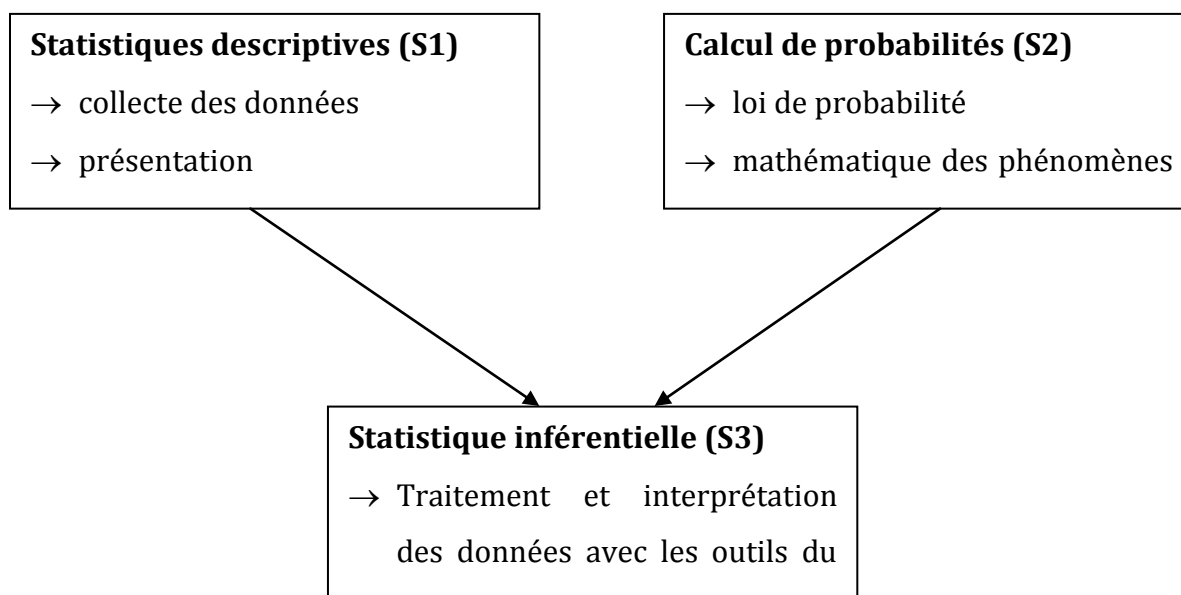
Si la moyenne des revenus échantillonnés est par exemple de 3256 DH.

Que représente cette valeur par rapport à l'ensemble de la population ?

C'est cela qui importe vraiment pour le gouvernement. Nous entrons là dans le domaine de la statistique inférentielle (dite aussi statistique inductive ou statistique mathématique) (cours du semestre 3)



Une étude de statistique peut se résumer ainsi :



2. Terminologie liée à la statistique :

Définition 1 : La **population** est l'ensemble de tous les individus concernés lors de l'étude d'un sujet particulier.

Exemple 1 : Dans le milieu rural, pour établir la proportion des enfants qui meurent avant l'âge de 5 ans, on considère la population des enfants du milieu rural dont la tranche d'âge se situe entre 0 et 5 ans (strictement).

Remarque : une population peut être

- finie (une population de personnes) ou
- infinie (les produits fabriqués en série sans arrêt)

Exemple 2 : Dans une entreprise de fabrication de batteries pour des Smartphones model A1, A2 et A3, on veut étudier la durée de consommation de la batterie chargée de model A1, on considère la population de toutes les batteries des Smartphones model A1 fabriquées par l'entreprise.

Définition 2 : On appelle **échantillon** toute *partie* ou *sous-ensemble* de la population.

Exemple 1 : Pour connaître la réaction des conducteurs de voitures vis-à-vis de la hausse du carburant, on interroge au hasard des automobilistes.

Exemple 2 : Pour établir le taux de pourcentage des pièces de mauvaise qualité fabriquées par une machine, on tire au hasard dans un lot fabriqué par cette machine un échantillon.

Définition 3 : On appelle **individu** *chaque élément* de la population ou de l'échantillon.

On peut utiliser le terme « **unité statistique** » quand l'élément représente

- un objet
- un événement
- une entité précise comme l'entreprise
- un groupe de personnes ou d'objets

Exemples :

- Chaque automobiliste dans l'exemple précédent est un individu
- Toute pièce fabriquée par la machine est une unité statistique
- Dans le recensement national marocain de 2014, on veut établir le nombre moyen d'enfants par famille dans les régions rurales, chaque famille dans ces régions est une unité statistique.

Définition 4 : La **taille** représente le nombre d'individus d'un échantillon (on notera n) ou d'une population (on notera N).

Définition 5 : Le **caractère** (ou **variable**) est l'aspect particulier que l'on désire étudier.

Exemple : Dans une enquête concernant une population de personnes, on peut s'intéresser à

- L'âge
- La taille
- Le sexe
- Le niveau d'éducation
- L'opinion sur un sujet donné
-

Définition 6 : On appelle **modalités** (**réalisation** ou **état** possible du caractère) les différentes possibilités que peut présenter un caractère.

Exemple 1 : Le nombre d'enfants par famille est un caractère de modalités les nombres entiers suivants 0, 1, 2, 3, 4,...

Exemple 2 : L'appréciation que peut retenir une clientèle vis-à-vis d'un service ou d'un produit (comestible ou pas) est un caractère qui peut avoir plusieurs modalités, on donne l'exemple d'un produit alimentaire, électroménager et d'un service :

Pour un produit alimentaire :

☐ Excellent ☐ Très bon ☐ Assez bon ☐ Moyen ☐ Pas bon

Pour un produit électroménager :

☐ Très utile ☐ Utile ☐ Moins utile

Pour un produit service :

☐ Excellent ☐ Important ☐ Banal ☐ Long ☐ Rapide ☐ Convivial ☐ Mal entretenu,...

Définition 7 : On dit d'un caractère qu'il est **qualitatif** si ses modalités ne s'expriment pas par un nombre.

Exemple : L'opinion, le sexe, ou la couleur des cheveux sont des caractères qualitatifs.

Remarque : on distingue deux cas pour un caractère qualitatif

- le cas où les modalités représentent un certain **accroissement** significatif
- le cas du caractère où les modalités n'ont aucune relation avec le sens des accroissements.

Dans le premier cas le caractère est dit **ordinal** tandis que le deuxième est dit **nominal**

Exemple 2 : La série suivante détermine la durée de vie dans un échantillon d'ampoules (elle est exprimée en heures) :

200	210	205	195	170	180	190	206	185	212	196	198
198	199	207	203	205	196	201	189				

Une série statistique peut être traitée de plusieurs façons différentes, en général on distingue

- le résumé des données sous forme de tableau,
- la représentation graphique
- le calcul des mesures caractéristiques dans le cas de caractères numériques (une mesure caractéristique peut être la moyenne, la médiane, ...)

3. Présentation des résultats : Tableau-Synthèse

3.1 Cas d'un caractère qualitatif

En distinguant les modalités du caractère observé, chacune d'elles constitue une classe particulière.

En comptant le nombre de données de la série qui se trouve dans la même classe, on constitue ce qu'on appelle « **Effectif** ».

Si le caractère a k modalités distinctes, les effectifs associés seront respectivement $n_1, n_2, \dots, n_k$.

Si n est la taille de l'échantillon, on doit avoir

$$n_1 + n_2 + \dots + n_k = n$$

Effectifs : représentent les **distributions** des individus ou des unités statistiques sur les différentes modalités.

On parle aussi de **fréquences** des classes :

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

Qu'on peut exprimer par des pourcentages (il s'agit d'un formatage), en multipliant les fréquences par 100

$$100 \frac{n_1}{n}, 100 \frac{n_2}{n}, \dots, 100 \frac{n_k}{n}$$

Remarque : une fréquence $f_i \in [0,1]$ se confond à la **probabilité** qu'un individu appartienne à la $i^{\text{ème}}$ classe.

3.2 Cas d'un caractère quantitatif discret

Définition : Un caractère quantitatif est discret si l'ensemble des valeurs qu'il peut prendre est fini ou dénombrable. Le plus souvent, ces valeurs sont entières.

Remarque : si le caractère a un nombre petit de modalités, on peut former des classes comme dans le cas du caractère qualitatif.

Par contre si le nombre de modalités est très grand, on traite le cas comme pour le cas du caractère continu.

Exemple : Le nombre d'enfants par famille dans une région est représenté dans la série suivante :

2	2	3	1	0	4	0	1	3	6	5	1	5	6	4	2	2
5	2	3	4	0	4	3	2	4	3	5	1	3	2	2	3	2

On résume ces données dans le tableau suivant :

Répartition de 34 familles d'une région selon le nombre d'enfants

Nombre d'enfants	Effectifs n_i	Fréquences f_i	Pourcentage $100f_i$
0	3	0.09	8.82
1	4	0.12	11.77
2	9	0.26	26.47
3	7	0.20	20.59
4	5	0.15	14.71
5	4	0.12	11.76
6	2	0.06	5.88
Total	34 (= n)	1.00	100.00

3.3 Cas d'un caractère quantitatif continu

Définition : Un caractère quantitatif est dit continu s'il peut prendre théoriquement n'importe quelle valeur dans un intervalle donné de nombres réels.

Exemple : La taille et le poids d'un individu.

Construire une distribution d'effectifs

- Constituer un certain nombre de classes.

- Toute classe représente un intervalle délimité par ses deux bornes appelés limites de classes.
- Deux classes ont une intersection vide ; c.à.d. qu'une donnée appartient à une seule classe.
- Les classes sont contiguës et d'égale largeur

3.3.1 Procédé de construction des classes dans le cas continu

Étape 1 : Déterminer l'étendue de la série

Soit x_i la $i^{\text{ème}}$ donnée observée.

Posons e l'étendu de la série :

$$e = \max_{1 \leq i \leq n} x_i - \min_{1 \leq i \leq n} x_i$$

Étape 2 : Déterminer le nombre de classes

Déterminer par une règle mathématique, on donne ici la formule de Sturges et on prend *la valeur entière la plus proche de*

$$1 + 3.322 \log_{10}(n)$$

où n est le nombre de données dans la série

Par exemple, si $n = 30$, la formule de Sturges donne la valeur de 5.907, l'entier le plus proche est 6 :

Nombre de classes selon la taille n de la série
(formule de Sturges)

Valeur de n	Nombre de classes
$12 \leq n < 23$	5
$23 \leq n < 46$	6
$46 \leq n < 91$	7
$91 \leq n < 181$	8
$181 \leq n < 362$	9
$362 \leq n < 724$	10
$724 \leq n < 1448$	11

Étape 3 : Déterminer l'amplitude des classes

L'amplitude des classes sera définit par

$$c = \frac{e}{k}$$

avec k le nombre de classes. Le nombre c ne contient pas plus de décimales que les données de la série, pour cela il suffit de tronquer le quotient $\frac{e}{k}$ et d'ajouter 1 au dernier chiffre retenu afin de se doter d'une amplitude suffisante pour que toutes les données puissent être contenues dans les classes.

Exemple 1 : On a calculé le quotient $\frac{e}{k}$ pour cinq séries statistiques, on a eut les résultats suivants : 0.8912 ; 4.2310 ; 11.1123 ; 6.3146 ; 7.0000

Si les données de la série contiennent deux chiffres après la virgule, l'amplitude des classes sera : 0.90 ; 4.24 ; 11.12 ; 6.32 ; 7.01

Si les données de la série contiennent seulement un chiffre après la virgule, l'amplitude des classes sera : 0.9 ; 4.3 ; 11.2 ; 6.4 ; 7.1

Si les données de la série sont entières, alors l'amplitude des classes sera : 1 ; 5 ; 12 ; 7 ; 8

Exemple 2 : La série suivante représente l'ancienneté du personnel cadre d'une entreprise

Ancienneté du personnel cadre en années

9.4	8.4	12.0	16.3	11.9	16.4	9.8	7.0	11.5	12.6
8.3	8.0	9.5	12.1	11.0	14.1	13.1	7.1	12.6	12.1
11.0	12.2	14.0	9.4	10.2	13.4	7.3	14.6	11.1	10.3
11.2	11.0	11.4	15.4	12.5	10.5	10.0	11.9	13.2	15.6
16.3	11.2	11.1	12.8	10.6	10.5	15.0	10.2	13.1	12.0
13.8	10.3	9.8	12.4	11.4	10.4	8.2	9.3	11.6	9.4

$n = 60$; $\text{Max } x_i = 16.4$; $\text{Min } x_i = 7.0$; l'étendue de la série est $e = 9.4$.

Le nombre de classes est égal à 7 ; $\frac{e}{k} = 9.4/7 = 1.34$ et l'amplitude des classes sera $c = 1.4$.

Étape 4 : Construire les intervalles

On fixe tout d'abord la limite inférieure du 1^{er} intervalle, on peut choisir pour cette limite la plus petite mesure de la série, ou une autre valeur inférieure mais proche.

En additionnant l'amplitude à cette valeur, on obtient la limite supérieure de la classe.

Pour les autres classes, la limite supérieure coïncide avec la limite inférieure de la classe suivante, puis on additionne à cette limite l'amplitude pour obtenir la limite supérieure de la nouvelle classe.

Exemple : On revient sur l'exemple 2 de la série qui représente l'ancienneté du personnel cadre d'une entreprise.

La limite inférieure du 1^{er} intervalle coïncidera la valeur 7.0 qui est la plus petite valeur de la série et puisque l'amplitude est $c = 1.4$; on aura comme premier intervalle $[7.0 ; 8.4[$.

Par convention, pour que toute donnée appartienne à une seule classe, les intervalles ainsi constitués seront fermés à gauche et ouverts à droite.

Les intervalles qui suivront le 1^{er} intervalle seront les suivants :

$[8.4 ; 9.8[$ $[9.8 ; 11.2[$ $[11.2 ; 12.6[$ $[12.6 ; 14[$ $[14 ; 15.4[$ $[15.4 ; 16.8[$

et on s'arrête à la 7^{ème} classe qui contiendra la plus grande valeur.

En effet il faut s'assurer toujours que toutes les données sont incluses dans l'ensemble des classes.

Étape 5 : Effectif et fréquence des classes

Il suffira ensuite de pointer les données de la série dans l'intervalle approprié, on peut ainsi calculer au besoin la fréquence ou le pourcentage.

Exemple :

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Effectif n_i	Fréquence f_i	Pourcentage %
$[7.0 ; 8.4[$	6	0.10	10.00
$[8.4 ; 9.8[$	6	0.10	10.00
$[9.8 ; 11.2[$	16	0.27	26.67
$[11.2 ; 12.6[$	15	0.25	25.00
$[12.6 ; 14[$	8	0.13	13.33
$[14 ; 15.4[$	4	0.07	6.67
$[15.4 ; 16.8[$	5	0.08	8.33
Total	60	1.00	100.00

Pour calculer certaines mesures caractéristiques de la distribution, on associe à chaque classe un centre et un effectif cumulé.

Le centre d'une classe est le point milieu de l'intervalle la constituant, il est donné par :

$$\frac{\text{limite inférieure} + \text{limite supérieure}}{2}$$

L'effectif cumulé d'une classe est la somme de l'effectif de cette classe avec les effectifs des classes qui l'ont précédé. On note N_i la fréquence cumulée de la i^{ème} classe.

$$N_i = n_1 + n_2 + \dots + n_i$$

Elle correspond au nombre de données de la série dont la valeur est inférieure à la limite supérieure de la classe.

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Centre m_i	Effectif n_i	Effectif cumulé N_i
[7.0 ; 8.4[	7.7	6	6
[8.4 ; 9.8[	9.1	6	12
[9.8 ; 11.2[	10.5	16	28
[11.2 ; 12.6[	11.9	15	43
[12.6 ; 14[	13.3	8	51
[14 ; 15.4[	14.7	4	55
[15.4 ; 16.8[	16.1	5	60
Total		60	

Remarque : Dans certain cas, on a intérêt à construire un ou deux classes plus larges que les autres, en général la première ou la dernière. Par exemple, dans une enquête on a interrogé les citoyens d'une ville, la répartition des individus selon les groupes d'âge est la suivante :

Âge : A en année	Effectif n_i
$A < 20$	...
$20 \leq A < 30$	...
$30 \leq A < 40$	...
$40 \leq A < 60$	...
$60 \leq A$	...

Exemples à résoudre

Faites l'exercice de résoudre les deux exemples suivants :

Exemple 1 : Les données suivantes représentent l'âge de 45 personnes choisies au hasard dans un collège :

17	16	17	18	18	17	17	19	17
22	18	17	22	18	38	17	17	21
29	22	18	45	18	17	22	19	44
20	20	17	19	19	17	22	17	17
18	22	21	17	20	18	19	23	18

Construire une distribution d'effectif, de fréquence et de pourcentage.

Solution :

L'âge est un caractère quantitatif continu ; on procède à la construction des classes :

$n = 45$; $\text{Max } x_i = 45$; $\text{Min } x_i = 16$; l'étendue est la valeur $e = 29$.

Le nombre de classes est égal à 6 ; $\frac{e}{k} = 29/6 = 4.8$ et l'amplitude des classes sera $c = 5$.

La première classe est $[16 ; 21[$, la dernière est $[41 ; 46[$.

Distribution de 45 collégiens selon leur âge

Âge	Effectif n_i	Fréquence f_i	Pourcentage %
$[16 ; 21[$	32	0.71	71.1
$[21 ; 26[$	9	0.20	20.1
$[26 ; 31[$	1	0.02	2.2
$[31 ; 36[$	0	0.00	0
$[36 ; 41[$	1	0.02	2.2
$[41 ; 46[$	2	0.04	4.5
Total	45	1.00	100

Exemple 2 : Le salaire horaire (en €) des 42 employés à temps partiel d'un magasin à rayons est respectivement :

6.80	6.80	7.45	7.60	7.30	6.80	6.30
6.30	6.80	6.30	6.30	7.10	7.30	6.80
5.55	6.00	6.90	7.50	6.40	5.90	6.05
5.75	5.55	5.55	5.55	8.25	6.90	7.00
8.35	5.75	6.70	5.90	6.25	6.50	7.15
5.55	6.25	5.90	8.00	6.60	5.55	7.50

Construire un tableau de fréquence tout en calculant la moyenne de classe, l'effectif cumulé et la fréquence cumulée.

Solution :

$e = 8.35 - 5.55 = 2.8$: étendu de la série

$k = 6$: nombre de classes

$$\frac{e}{k} = \frac{2.8}{6} = 0.466..$$

On choisira comme amplitude la valeur $c = 0.47$

Distribution des employés selon le salaire horaire

Salaire	Moyenne de classe	Effectif n_i	Effectif cumulé N_i	Fréquence f_i	Fréquence cumulée F_i
[5.55 ; 6.02[	5.79	12	12	0.29	0.29
[6.02 ; 6.49[	6.26	8	20	0.19	0.48
[6.49 ; 6.96[	6.73	10	30	0.24	0.71
[6.96 ; 7.43[	7.20	5	35	0.12	0.83
[7.43 ; 7.90[	7.67	4	39	0.10	0.93
[7.90 ; 8.37[	8.14	3	42	0.07	1.00
Total		42		1.00	

4. Représentations graphiques

La représentation des résultats graphiquement c.à.d. par des dessins et tracés géométriques facilite la description des résultats contenus dans le tableau-synthèse des données.

Il existe plusieurs représentations graphiques d'une même série de données. On peut citer le diagramme à bandes rectangulaires, l'histogramme, le polygone, l'ogive et le diagramme à secteurs circulaires.

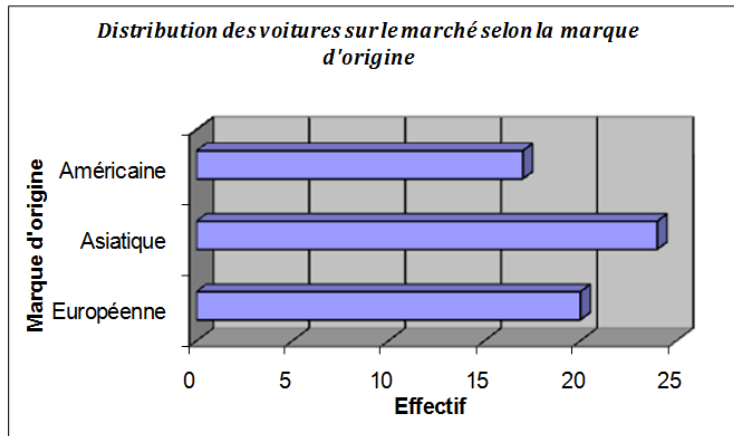
4.1 Diagramme à bandes rectangulaires

Il est particulièrement adapté à la représentation d'un caractère qualitatif ou quantitatif discret, il est constitué par la juxtaposition de bandes verticales ou la superposition de bandes horizontales (le plus souvent pour ce cas là, l'ordre des modalités pour les données nominales est arbitraire).

Exemple1 : dans le but de connaître les tendances du marché de la voiture citadine, un sondage a été effectué auprès d'un certain nombre de concessionnaires. Les résultats sont groupés dans le tableau suivant selon que la voiture est de marque d'origine européenne, asiatique ou américaine :

Marque d'origine	Effectif n_i	Fréquences f_i
Européenne	20	0.33
Asiatique	24	0.39
Américaine	17	0.28
Total	61	1

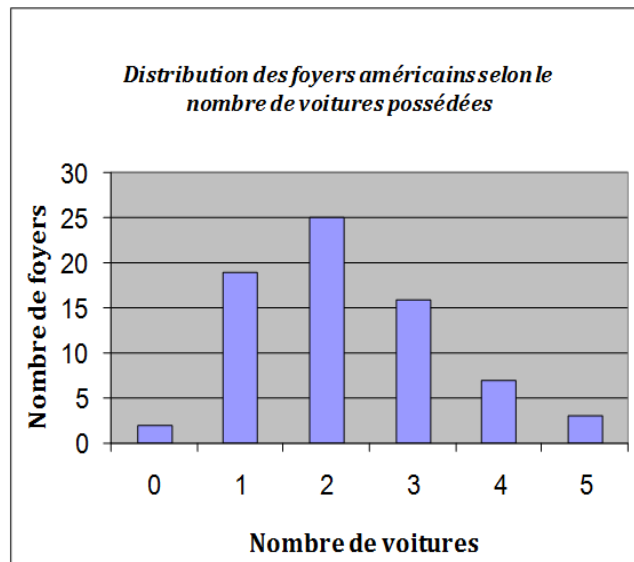
Le caractère observé n'a pas de modalités qu'on peut ordonner, on choisira pour le graphe un diagramme à bandes rectangulaires horizontales.



Exemple2 : On veut connaître le nombre de voitures par foyer américain, on choisit pour cela au hasard 72 foyers représentant la population d'un état américain. Les résultats sont regroupés dans le tableau suivant :

Nombre de voitures	Nombre de foyers n_i
0	02
1	19
2	25
3	16
4	07
5	03
Total	72

Les modalités du caractère étudié sont ordonnées, on choisit le diagramme à bandes rectangulaires verticales.



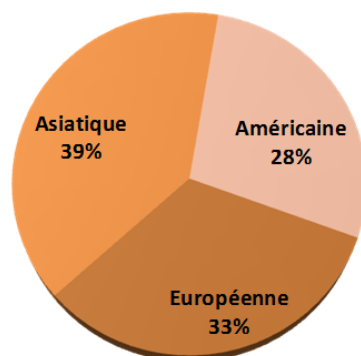
4.2 Diagramme à secteurs circulaires

Ce type de graphique est utilisé le plus souvent dans le cas du caractère qualitatif. On divise un cercle en autant de secteurs circulaires qu'il y'a de classes. Pour déterminer l'angle de chaque secteur on multiplie la fréquence de la classe par 360° .

Reprenons l'exemple 1 :

Marque d'origine	Effectif n_i	Fréquences f_i	Pourcentages	Angle au centre du secteur
Européenne	20	0.33	33%	118.03°
Asiatique	24	0.39	39%	141.64°
Américaine	17	0.28	28%	100.33°
Total	61	1.00	100%	360.00°

Distribution des voitures sur le marché
selon la marque d'origine

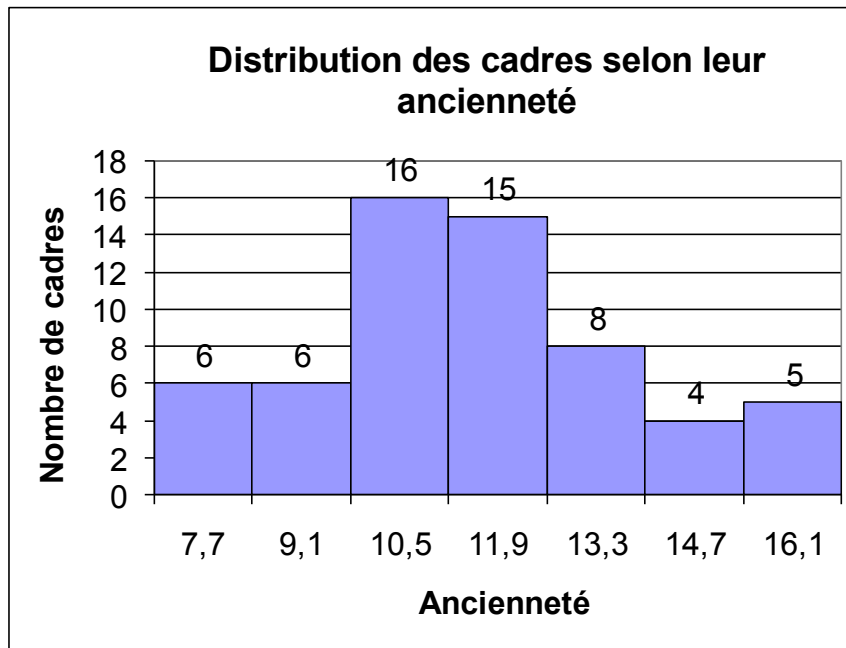


Remarque : le diagramme à secteurs circulaire convient bien au cas de l'exemple 2 où le caractère est quantitatif discret.

4.3 Histogramme

Il convient particulièrement pour présenter un caractère quantitatif continu, il est constitué par la juxtaposition de bandes rectangulaires verticales, mais **adjacentes**. De plus, chaque rectangle doit présenter une largeur équivalente à l'amplitude de la classe qu'il représente, et une hauteur telle que sa surface soit proportionnelle à l'effectif.

Exemple1 : on reprend le tableau-synthèse de la distribution de l'ancienneté de 60 cadres d'une société



Dans cet exemple, puisque l'amplitude des classes est uniforme c.à.d. la même, la hauteur des rectangles coïncide avec l'effectif de la classe correspondante.

Dans d'autres exemples, si une classe (le plus souvent la 1^{ère} ou la dernière) a une amplitude différente de celles des autres classes, alors on doit ajuster la hauteur de cette classe tel que la surface du rectangle lui correspondant soit proportionnelle à l'effectif.

Exemple2 : Les notes obtenues par un groupe d'étudiants à un examen d'informatique sont répartis comme suit :

Note sur 20	Nombre d'étudiants n_i
[00 ; 10[	135
[10 ; 12[	43
[12 ; 14[	16
[14 ; 16[	13
[16 ; 20[	09
Total	216

La 1^{ère} et la 5^{ème} classe sont d'amplitudes respectivement 10 et 4, tandis que les autres classes ont une amplitude standard égale à 2.

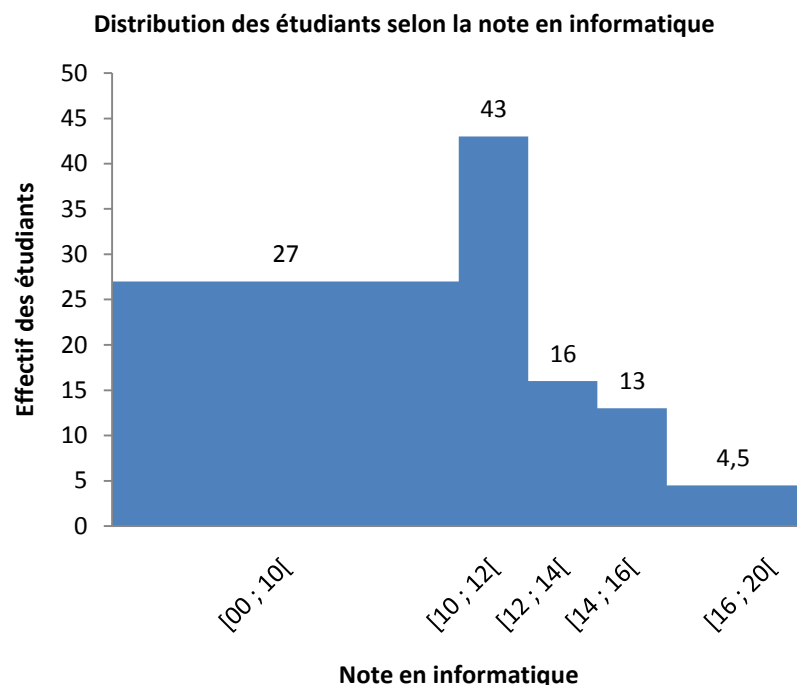
Pour la 1^{ère} classe, on doit déterminer la hauteur h_1 du rectangle de surface $h_1 * 10$, cette surface correspondra au rectangle de largeur 2 et de hauteur 135 c.à.d. qu'on doit vérifier :

$$h_1 * 10 = 270, \text{ ce qui revient à la valeur : } h_1 = 27.$$

(270 est la surface du rectangle de la 1^{ère} classe qui est calculée en supposant le rectangle est de hauteur 135 et de largeur standard 2)

Pour la 5^{ème} classe, la hauteur h_2 du rectangle lui correspondant vérifie $h_2 * 4 = 18$, c.à.d. que $h_2 = 4.5$.

Si on trace l'histogramme des effectifs on obtient le graphe suivant :



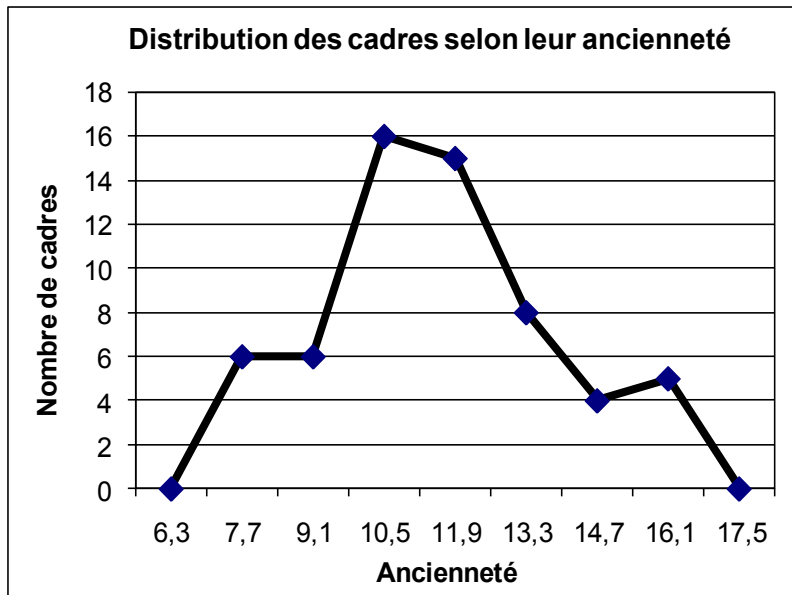
4.4 Polygone

Pour dessiner ce graphe, on place tout d'abord les centres des classes sur l'axe des abscisses et pour chaque centre on place à hauteur verticale un autre point avec une distance correspondant à l'effectif de la classe considérée. On relie les différents points obtenus par des segments de droite.

Par définition, un polygone est une figure fermée délimitée par des segments.

Pour terminer le graphe, on ferme le polygone en ajoutant deux segments à chaque extrémité, les points qu'on ajoutera seront sur l'axe des abscisses à une hauteur égal à zéro.

Exemple :



Remarque : Dans les exemples précédents, au lieu de tracer les effectifs sur l'axe vertical, on peut tracer les fréquences ou les pourcentages, sans que l'allure du graphe change.

4.5 L'ogive

Ce graphe sera utilisé pour tracer les effectifs cumulés, on place tout d'abord les points correspondants aux limites des classes sur l'axe des abscisses, et pour chaque classe on place un point au dessus de la **limite supérieure** de la classe à hauteur l'effectif cumulé de celle-ci. Puis on relie les points par des segments de droite. On ajoutera un point sur l'axe des abscisses distant de la première limite de classe par l'amplitude standard des classes puis on relie ce point au graphe.

Remarque : on suppose qu'à l'intérieure d'une classe la distribution est uniforme, ce qui nous permet de justifier l'interpolation entre les deux bornes de la même classe par un segment linéaire.

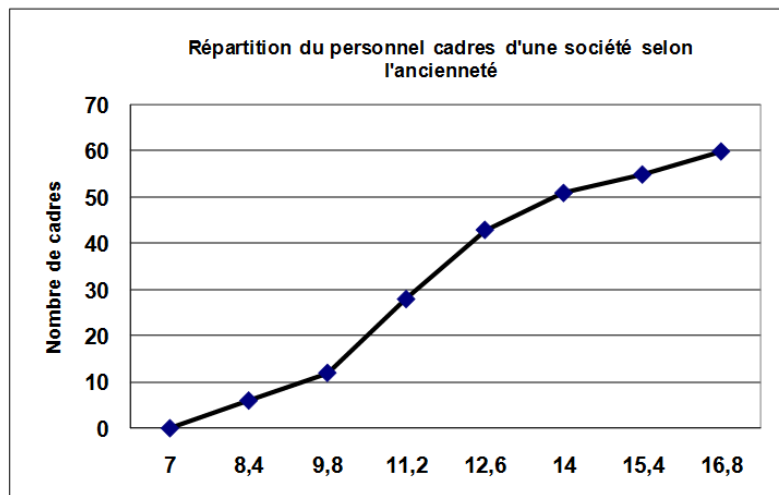
Exemple :

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Centre m_i	Effectif n_i	Effectif cumulé N_i
[7.0 ; 8.4[	7.7	6	6
[8.4 ; 9.8[	9.1	6	12
[9.8 ; 11.2[	10.5	16	28
[11.2 ; 12.6[	11.9	15	43
[12.6 ; 14[	13.3	8	51
[14 ; 15.4[	14.7	4	55
[15.4 ; 16.8[	16.1	5	60
Total		60	

Dans cet exemple, on doit tracer les points de coordonnées suivants :

(8.4 ; 6) ; (9.8 ; 12) ; (11.2 ; 28) ; (12.6 ; 43) ; (14 ; 51) ; (15.4 ; 55) ; (16.8 ; 60) puis on complète par un dernier point (7 ; 0). On relie ensuite par des segments de droite.



5. Calcul des mesures caractéristiques

Pour les variables quantitatives, les synthèses telles que les tableaux des effectifs, histogrammes, etc. ne sont pas suffisantes pour connaître la population ou l'échantillon. On doit apporter d'autres informations pour comprendre comment les données se comportent.

Il est très difficile d'appréhender (comprendre) l'ensemble des informations de la population ou de l'échantillon même si elles sont synthétisées.

→ Recourir à (faire appel à) des mesures unidimensionnelles pour décrire la population ou l'échantillon. Ces mesures numériques permettent entre autre la comparaison de populations ou d'échantillons ; et surtout la description des valeurs observées.

On distingue cinq types de mesures caractéristiques de la série des données :

- Les **mesures de tendance centrale** sont des valeurs autour desquelles se regroupent les différentes valeurs de la série
- Les **mesures de dispersion** informent sur la variabilité des résultats
- Les **mesures de position** renseignent sur la situation d'un résultat par rapport à l'ensemble de la distribution
- Les **mesures de forme** permettent de reconnaître la forme de la distribution de la variable étudiée et de la comparer à la distribution dite normale
- Les **mesures de concentration** s'intéressent au caractère dont une accumulation a une signification particulière (exemple accumulation des richesses)

5.1 Mesures de tendance centrale

5.1.1 Le mode

C'est la valeur la plus fréquente (qui se répète le plus souvent), on parle aussi de classe modale, c'est la classe qui correspond au maximum de l'effectif ou de fréquence.

On notera le mode par **Mo**.

Exemple1 : soit les trois séries suivantes

Série 1

1	3	0	3	2
---	---	---	---	---

Série 2

7	6	5	6	4	5	3
---	---	---	---	---	---	---

Série 3

42	35	18	33	21	23	25
----	----	----	----	----	----	----

Le mode de la série 1 est 3, ceux de la série 2 sont 6 et 5 tandis que la série 3 n'a pas de mode.

Dans le cas de la classe modale on prendra comme valeur du mode le centre de la classe.

Exemple2 : Dans l'exemple de la distribution de 60 cadres selon leur ancienneté, la classe modale est

[9.8 ; 11.2[d'effectif maximal égal à 16.

On prend comme valeur modale le centre de classe, c.à.d. la valeur $Mo = 10.5$

On dira que l'ancienneté la plus fréquente dans la société est dix ans et demi

Le mode est défini même pour les caractères qualitatifs.

Exemple3 : Soit la série représentant le caractère sexe d'un groupe de 20 étudiants,

F M F M M F M M M F F M M F M M M F M M

Le mode de cette série est masculin

5.1.2 La médiane

a. Cas de données quantitatives non groupées en classes

On va la noter **Me**. C'est la valeur qui partage la série d'observations ordonnées en deux parties de tailles égales.

D'une autre façon, si on place les données sur un axe horizontal en respectant l'ordre croissant ou décroissant des données, alors la médiane se trouvera à mi chemin entre la 1^{ère} moitié des données à gauche et la 2^{ème} moitié des données à droite

Le rang de la médiane se détermine de la manière suivante :

- Si n le nombre d'observations est impaire alors le rang de la médiane sera $\frac{n+1}{2}$

Exemple1 : soit la série des données suivantes :

1	2	4	3	2	5	2	3	1
---	---	---	---	---	---	---	---	---

La même série sera ordonnée dans le sens croissant :

1	1	2	2	2	3	3	4	5
---	---	---	---	---	---	---	---	---

Ici $n=9$ impair, donc la médiane est de rang

$$\frac{9+1}{2} = 5$$

et ça coïncide avec la valeur 2, d'où

$$Me = 2$$

- Dans le cas où n est pair, le rang de la médiane se situe entre les observations de rang

$$\frac{n}{2} \text{ et } \frac{n}{2} + 1$$

On en prend généralement la valeur intermédiaire (le milieu) comme valeur médiane

Exemple2 : soit la série des données suivantes :

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

La même série sera ordonnée dans le sens croissant :

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

Ici $n = 10$ pair, donc la médiane est la valeur entre les deux données de rangs

$$\frac{10}{2} = 5 \text{ et } \frac{10}{2} + 1 = 6$$

d'où la médiane est entre les deux valeurs 2 et 3, on prend le milieu entre les deux valeurs c.à.d. on prend

$$Me = \frac{2+3}{2} = 2.5$$

b. Cas où les données sont groupées en classes

Si les données proviennent d'un caractère quantitatif continu et si elles sont groupées en classes, on doit en premier définir la classe médiane :

→ il s'agit de la classe qui contient l'observation de rang $(n+1)/2$, c'est-à-dire la première classe dont la fréquence cumulée égale ou dépasse 50%

On donne les deux formules qui approchent la valeur médiane, l'une utilisera les effectifs et l'autre les fréquences.

Si on utilise les effectifs, on peut approcher la valeur de la médiane par la formule suivante :

$$Me \cong b_{inf} + \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right) * c$$

où : n est le nombre d'observations (taille de la série).

b_{inf} est la limite inférieure de la classe médiane.

n_i est l'effectif de la classe médiane.

N_{i-1} est l'effectif cumulé de la classe précédente la classe médiane.

c est l'amplitude de la classe médiane.

Si on utilise les fréquences la formule devient

$$Me \cong b_{inf} + \left(\frac{\frac{1}{2} - F_{i-1}}{f_i} \right) * c$$

b_{inf} est la limite inférieure de la classe médiane

f_i est la fréquence de la classe médiane

F_{i-1} est la fréquence cumulée de la classe précédente la classe médiane

c est l'amplitude de la classe médiane.

Exemple : Reprenons l'exemple de la distribution de l'ancienneté de 60 cadres d'une société

Ancienneté en année	Effectif n_i	Effectif cumulé N_i
[7.0 ; 8.4[	6	6
[8.4 ; 9.8[	6	12
[9.8 ; 11.2[	16	28
[11.2 ; 12.6[	15	43
[12.6 ; 14[	8	51
[14 ; 15.4[	4	55
[15.4 ; 16.8[	5	60
Total	60	

Calculons

$$\frac{n+1}{2} = 30.5$$

La médiane se situe dans la 4^{ème} classe, c'est la classe médiane. Elle contient aussi toutes les données (ordonnées dans le sens croissant) dont le rang est compris entre 29 et 43.

Pour le calcul de la médiane, on a les données suivantes :

$$n = 60 ; b_{inf} = 11.2 ; n_i = 15 ; N_{i-1} = 28 ; c = 1.4$$

Ainsi, on obtient :

$$Me \cong 11.2 + \left(\frac{30 - 28}{15} \right) * 1.4 = 11.387$$

Conclusion : on peut affirmer que 50% des cadres de la société ont une ancienneté inférieure à 11.39 années et que les 50% qui restent ont une ancienneté de plus de 11.39 années

Si on veut utiliser la colonne des fréquences

Ancienneté en année	Fréquence f_i	Fréquence cumulée F_i
[7.0 ; 8.4[	0.10	0.10
[8.4 ; 9.8[	0.10	0.20
[9.8 ; 11.2[	0.27	0.47
[11.2 ; 12.6[	0.25	0.72
[12.6 ; 14[	0.13	0.85
[14 ; 15.4[	0.07	0.92
[15.4 ; 16.8[	0.08	1.00
Total	1	

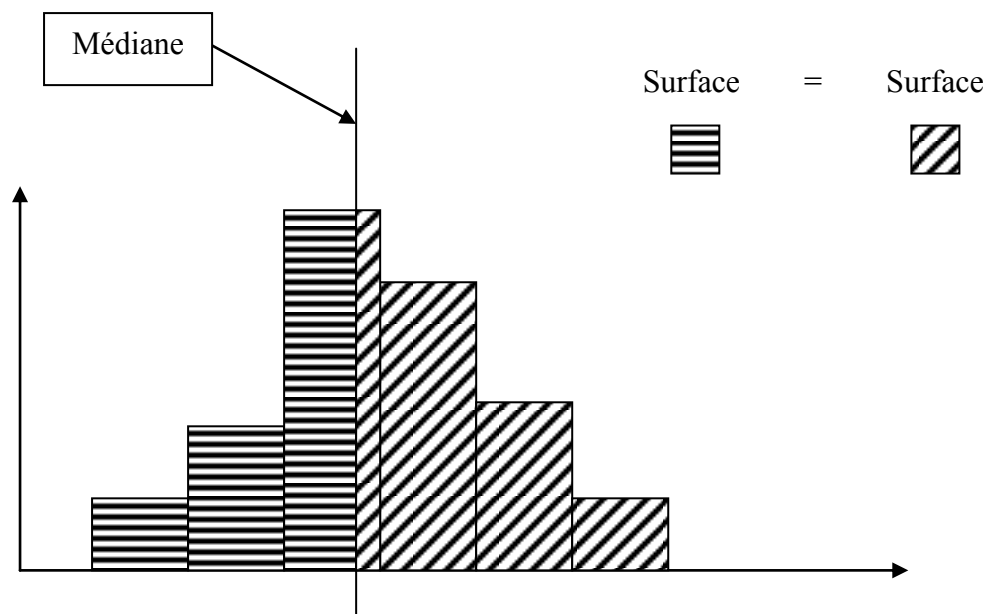
On cherchera la classe qui correspond à une accumulation de plus de 0.5=50% des données

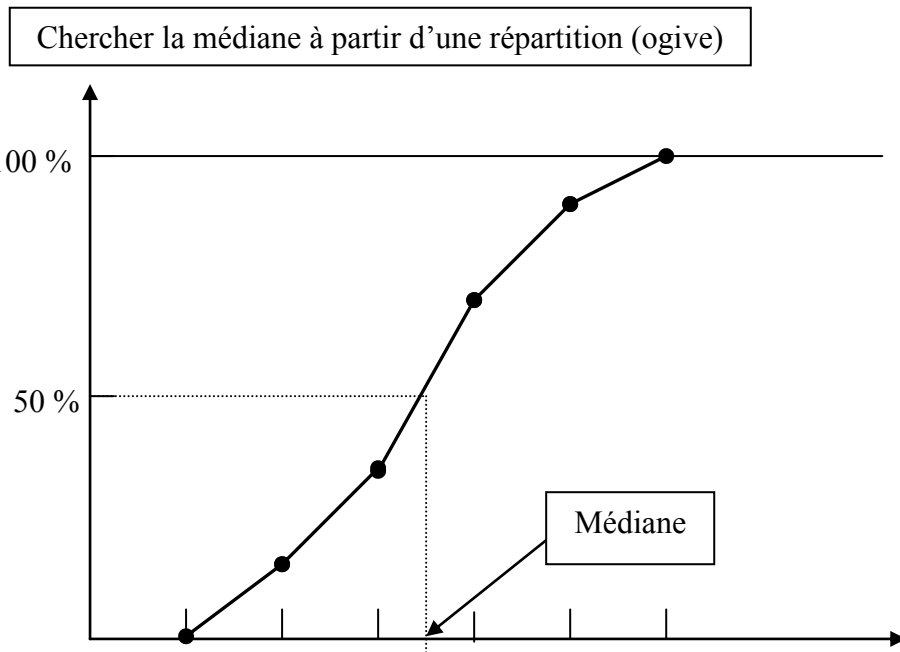
On fait le calcul : $b_{inf} = 11.2$; $f_i = 0.25$; $F_{i-1} = 0.47$; $c = 1.4$

$$Me \cong 11.2 + \left(\frac{0.5 - 0.47}{0.25} \right) * 1.4 = 11.387$$

Remarque : on peut schématiser la médiane sur un graphe de type histogramme ou ogive.

Chercher la médiane à partir d'un histogramme





5.1.3 La moyenne arithmétique

On va la noter μ pour la moyenne d'une population et $\bar{x}$ ou $\bar{x}_n$ pour la moyenne d'un échantillon (en anglais, mean ou average). C'est la somme des valeurs observées divisée par le nombre d'observations.

Elle représente le centre de gravité de l'ensemble des points représentés par les valeurs de la série, c.à.d. le point d'équilibre de toutes les données de la série en positionnant les valeurs sur une échelle graduée.

Dans le cas où toutes les données de la population sont fournies, la moyenne sera calculée sur les N valeurs observées :

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

avec x_i la $i^{\text{ème}}$ valeur observée dans la série des données.

Si le paramètre μ (mu) est inconnu, le substitue par la moyenne $\bar{x}_n$ calculée sur les n données de l'échantillon.

Exemple : reprenons l'exemple du nombre de voitures par foyer américain

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

On calcule la moyenne sur cet échantillon, on obtient :

$$\bar{x}_n = \frac{1 + 2 + 2 + 1 + 3 + \dots + 3}{72} = \frac{160}{72} = 2.222$$

Ce calcul long peut être allégé en utilisant le tableau-synthèse déjà construit. Posons pour cela $x_1^*, \dots, x_k^*$ les k valeurs distinctes des modalités de la série des données, on peut réécrire la moyenne $\bar{x}_n$:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^k n_i x_i^* = \sum_{i=1}^k \frac{n_i}{n} x_i^* = \sum_{i=1}^k f_i x_i^*$$

où n_i et f_i sont respectivement l'effectif et la fréquence de la modalité x_i^*

Reprenons l'exemple précédent, on utilisera le tableau-synthèse :

*Distribution des foyers américains selon
le nombre de voitures possédées*

Nombre de voitures x_i^*	Nombre de foyers n_i	Fréquences f_i	$f_i x_i$
0	02	0.028	0
1	19	0.264	0.264
2	25	0.347	0.694
3	16	0.222	0.667
4	07	0.097	0.389
5	03	0.042	0.208
Total	$n = 72$	1.000	$\bar{x}_n = 2.222$
La moyenne est : 2.22			

$$\bar{x}_n = \sum_{i=1}^k f_i x_i^* = 0 * 0.028 + 1 * 0.264 + 2 * 0.347 + \dots + 5 * 0.042 = 2.222$$

- **Cas où les données sont groupées en classes**

Les données considérées ici proviennent d'un caractère quantitatif continu.

On calcule approximativement la valeur de la moyenne en utilisant les centres des classes :

$$\bar{x}_n \cong \frac{1}{n} \sum_{i=1}^k n_i m_i = \sum_{i=1}^k f_i m_i ,$$

avec

k est le nombre de classes

m_i est le centre de la $i^{\text{ème}}$ classe

Exemple :

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Centre m_i	Effectif n_i	$n_i m_i$
[7.0 ; 8.4[	7.7	6	46.2
[8.4 ; 9.8[	9.1	6	54.6
[9.8 ; 11.2[	10.5	16	168.0
[11.2 ; 12.6[	11.9	15	178.5
[12.6 ; 14[	13.3	8	106.4
[14 ; 15.4[	14.7	4	58.8
[15.4 ; 16.8[	16.1	5	80.5
Total		60	693.0

D'où

$$\mu = \frac{693}{60} = 11.55$$

Remarque : les valeurs extrêmes ne modifient pas la médiane. La moyenne arithmétique par contre y est sensible, on dit qu'elle n'est pas robuste.

- **Moyenne pondérée**

Dans beaucoup de cas, on est amené à pondérer les valeurs en les multipliant par des coefficients de pondération (des poids), le but est de dire que les données recueillies n'ont pas la même importance.

Si on attribue à la $i^{\text{ème}}$ donnée x_i un poids p_i , on peut définir la moyenne pondérée par

$$\bar{x}_p = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i}$$

Dans le cas où on accorde la même importance pour toutes les données de la série, c.à.d. que $p_i = p$ pour tout i , alors la formule de la moyenne pondérée $\bar{x}_p$ coïncidera avec la moyenne arithmétique $\bar{x}$:

$$\bar{x}_p = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1}^n p_i} = \frac{\sum_{i=1}^n p x_i}{\sum_{i=1}^n p} = \frac{p \sum_{i=1}^n x_i}{n p} = \bar{x}$$

Exemple1 : Dans un groupe d'étudiants, on a calculé la moyenne de groupe des 3 matières examinées, le système de coefficients différencie chaque matière par le coefficient qui lui correspond, on voudrait calculer la moyenne générale de groupe.

Matière	Coefficient p_i	Moyenne de groupe x_i	$p_i x_i$
1	1	11.3	11.3
2	2	9.5	19
3	1.5	12.7	19.05
Total	4.5		49.35

$$\bar{x}_p = 49.35/4.5 = 10.97$$

Exemple2 : Dans une usine, on fabrique des pièces simultanément par 5 machines différentes, placées parallèlement, on veut connaître le pourcentage des pièces défectueuses résultant de cette production. On a trouvé les résultats suivants :

Distribution des pièces défectueuses sur 5 machines

Machine	Taille de l'échantillon	Pourcentage des pièces défectueuses
1	60	5.0
2	70	4.5
3	50	7
4	100	2
5	160	1

Le calcul direct de la moyenne arithmétique du pourcentage ne prend pas en compte l'importance de chaque échantillon prélevé

$$\bar{x} = \frac{5 + 4.5 + 7 + 2 + 1}{5} = 3.9$$

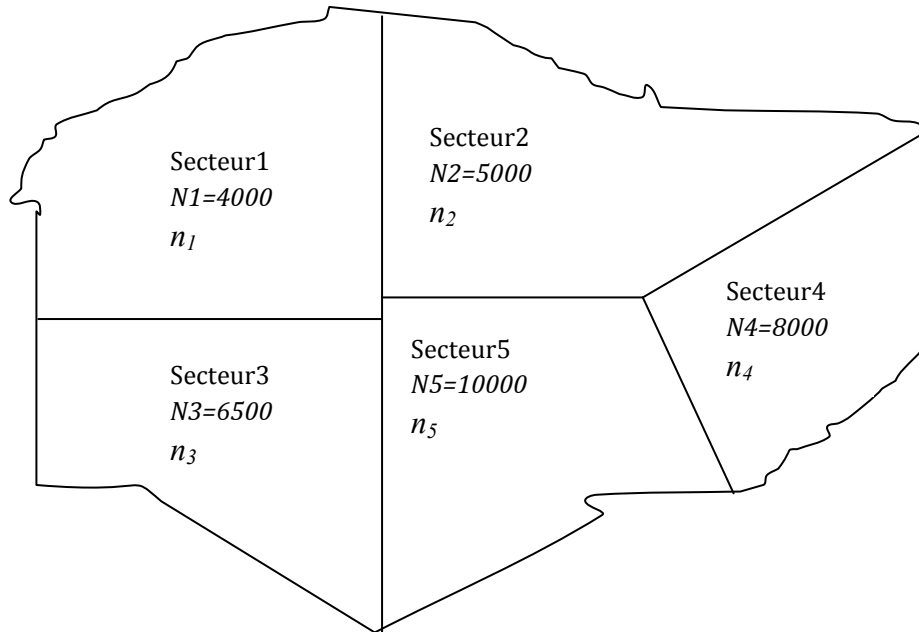
Cette valeur ne sera considérée que si on a pris la même taille de l'échantillon pour chaque machine.

On doit inclure le poids de chaque échantillon, le poids est présenté ici par la taille de l'échantillon :

$$\bar{x}_p = \frac{5 * 60 + 4.5 * 70 + 7 * 50 + 2 * 100 + 1 * 160}{60 + 70 + 50 + 100 + 160} = \frac{1325}{440} = 3.01.$$

On utilise souvent ce procédé de calcul dans les enquêtes et sondages.

Exemple3 : Dans un sondage fait sur une population d'une ville, on a divisé la superficie de celle-ci en 5 secteurs ou strates. On a pris au hasard dans chaque secteur un certain nombre de familles pour enfin observer le nombre d'enfants par familles. Le choix de la taille de chaque échantillon n'est pas fait au hasard mais déterminé auparavant en respectant la représentativité de chaque secteur dans l'ensemble de la population de la ville



La taille totale de la population ciblée est $N = N_1 + N_2 + N_3 + N_4 + N_5 = 33500$

Si on choisit la taille de l'échantillon global $n = 100$, il faudra déterminer les tailles n_1, n_2, n_3, n_4 , et n_5 des 5 échantillons qui seront prélevés des 5 secteurs et qui satisferont la relation suivante :

$$n_1 + n_2 + n_3 + n_4 + n_5 = n$$

Pour préserver cette représentativité de l'échantillon sectoriel par rapport à l'ensemble des données des 5 échantillons, on respecte le rapport de la division par secteur choisit au départ, et on aura :

$$\frac{N_i}{N} = \frac{n_i}{n}, i = 1, 2, 3, 4, 5.$$

Ce qui nous permet de calculer les tailles des 5 échantillons par la formule :

$$n_i = n * \frac{N_i}{N}$$

Les résultats sont les suivants :

Secteur i	Nombre de familles prélevées n_i	Moyenne du nombre d'enfants $\bar{x}_i$
1	12	1.5
2	15	3.4
3	19	3.0
4	24	2.7
5	30	2.4

$\bar{x}_i$ est la moyenne calculée sur le $i^{\text{ème}}$ échantillon, dans cet exemple le poids de chaque échantillon est sa taille. On a :

$$\bar{x}_p = \frac{\sum_{i=1}^5 n_i \bar{x}_i}{\sum_{i=1}^5 n_i}$$

Si on fait le calcul, on trouvera la valeur

$$\bar{x}_p = \frac{262.8}{100} = 2.63$$

5.1.4 Autres moyennes

- **La moyenne harmonique**

Notée **H**, elle est l'inverse de la moyenne des inverses des données x_i , c.à.d.

$$H = 1 / \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \right) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

Les valeurs x_i sont supposées non nulles.

Si les données sont regroupées en classes, alors on peut écrire :

$$H = 1 / \left(\frac{1}{n} \sum_{i=1}^k \frac{n_i}{x_i^*} \right)$$

k est le nombre de classes.

Dans le cas d'un caractère continu, la valeur x_i^* sera la moyenne de la $i^{\text{ème}}$ la classe ($x_i^* = m_i$).

La moyenne harmonique est utilisée dans le calcul des durées ou des vitesses moyennes comme elle est utilisée dans l'évaluation des taux ou des pourcentages moyens.

Exemple : Si on parcourt sur autoroute la moitié de la distance entre Oujda et Rabat avec une vitesse de 120km/h et l'autre moitié avec une vitesse de 80km/h alors la vitesse moyenne avec laquelle on parcourt le trajet Oujda-Rabat est :

$$\frac{2}{\frac{1}{120} + \frac{1}{80}} = 2 * 48 = 96 \text{ km/h}$$

- **La moyenne géométrique**

Notée **G**, elle est la racine nième du produit des valeurs x_i qui sont supposées positives :

$$G = \sqrt[n]{\prod_{i=1}^n x_i} = [x_1 x_2 \dots x_n]^{\frac{1}{n}}$$

Dans le cas de données groupées en classes, la formule de G devient :

$$G = \sqrt[n]{\prod_{i=1}^k [x_i^*]^{n_i}} = [[x_1^*]^{n_1} [x_2^*]^{n_2} \dots [x_k^*]^{n_k}]^{\frac{1}{n}}$$

on doit s'assurer que $n = \sum_{i=1}^k n_i$.

Cette moyenne est préconisée dans l'évaluation du changement des prix ou des salaires (taux d'accroissement), elle a l'avantage d'être moins sensible aux valeurs très élevées d'une série de données.

On remarquera que le logarithme de G n'est autre que la moyenne arithmétique des données $\log(x_i), i = 1, \dots, n$.

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

- **La moyenne quadratique**

Notée Q , elle est définie comme la racine carrée de la moyenne arithmétique des carrés des valeurs x_i :

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{1}{n} [x_1^2 + x_2^2 + \dots + x_n^2]}$$

Dans le cas de données groupées en classes, on aura :

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^k n_i x_i^{*2}}$$

Cette moyenne est utilisée dans le calcul de certaines mesures de dispersion.

Remarque : on peut montrer que les quatre moyennes définies ci-dessus vérifient la relation suivante :

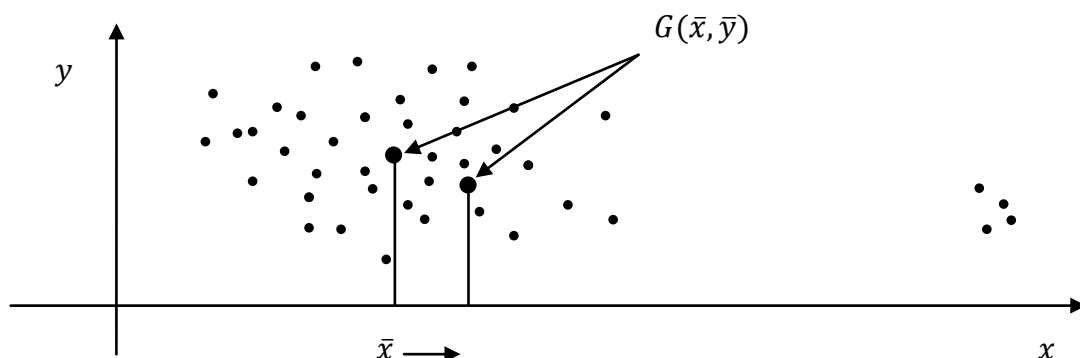
$$H \leq G \leq \bar{x} \leq Q$$

Contrainte oblige dans le calcul de H et de G , on supposera que toutes les données sont positives non nulles.

- **Choix de la mesure de tendance centrale**

La plus importante et la plus utilisée des mesures de tendance centrale est la moyenne, qui peut représenter fort bien le nuage des données dessinées sur une échelle graduée (ce nuage est constitué par les données se rapprochant d'elle même plus au moins).

Néanmoins, dans certains cas où des données extrêmes de la série se situent à l'extérieur de ce nuage et qui par leurs éloignements considérables du nuage affecte la mesure de la moyenne en ne reflétant pas la réalité des données.



Exemple : Dans une petite entreprise de 10 personnes, on a 8 ouvriers, un gestionnaire comptable et un directeur, les salaires sont les suivants (exprimées en DH) :

2400 2400 2800 2400 2600 2800 2800 2600 3200 14000

La moyenne de cette série est :

$$\bar{x} = \frac{38000}{10} = 3800$$

Tous les salaires, sauf un, sont inférieurs à cette moyenne.

Dans cet exemple, on ne peut pas se fier à cette moyenne, mais on peut utiliser d'autres mesures de tendance centrale comme la médiane :

Le nombre d'observation $n = 10$ est pair, on prend la médiane comme la moyenne de la 5^{ème} et la 6^{ème} valeurs des données ordonnées par ordre croissant :

2400 2400 2400 2600 2600 2800 2800 2800 3200 14000

$$Me = \frac{2600 + 2800}{2} = 2700$$

Cette valeur représente mieux la réalité des salaires dans cette petite entreprise.

Dans d'autres cas, on trouve que la 1^{ère} ou la dernière classe n'est pas bornée, le centre de classe n'existe pas et ne sera pas utilisé dans la formule du calcul de la moyenne, dans ce cas on utilise la médiane ou le mode.

Exemple : La distribution suivante représente la répartition du salaire hebdomadaire d'un groupe d'individus.

Salaire hebdomadaire S (en DH)	Effectif n_i	Effectif cumulé N_i
$S < 500$	52	52
$500 \leq S < 1000$	29	81
$1000 \leq S < 1500$	33	114
$1500 \leq S < 2000$	39	153
$2000 \leq S < 2500$	30	183
$2500 \leq S < 3750$	19	202
$3750 \leq S$	11	213
Total	213	

La dernière classe n'a pas de limite supérieure, il est impossible de calculer la moyenne, mais on peut procéder au calcul de la médiane :

n est impair, $n/2 = 213/2 = 106.5$; la médiane se situe dans la 3ème classe et on a :

$Me \cong 1000 + \left(\frac{n/2 - N_2}{n_3} \right) * c$, $N_2 = 81$, $n_3 = 33$ et $c = 500$, ce qui donne

$$Me \cong 1000 + \left(\frac{106.5 - 81}{33} \right) * 500 = 1386.36 DH$$

On va conclure que 50% des salariés ont des salaires hebdomadaires inférieurs ou égaux à 1386.36DH.

Pour le mode, il ne constitue pas le premier choix pour le calcul de la mesure de la tendance centrale, puisqu'il est peu stable dans certains cas où on peut changer une seule donnée pour passer d'une classe modale à une autre. Mais en général le mode se situe proche de la moyenne et on est intéressé par connaître la valeur dominante dans la série des données.

Exemple1 : Soit la série suivante : 4, 4, 3, 0, 1, 2, 5, 3, 3, 2, 0, 1, 5.

Le mode est égale à 3, mais une modification d'une seule donnée peut basculer le mode pour prendre une autre valeur.

Exemple2 : Dans une enquête, on a recensé le nombre d'individus inscrits dans un club sportif par tranche d'âge

Tranche d'âge [a_i, b_i]	Effectif n_i	$n_i / (b_i - a_i)$
[15 ; 20[	11	2.2
[20 ; 30[	23	2.3
[30 ; 40[	24	2.4
[40 ; 55[	17	1.1
[55 ; 65[	05	0.5
Total	80	

La classe modale de cette série est [30 ; 40[et le mode sera choisi comme le centre de cette classe, c.à.d. $Mo = 35 ans$.

Dans le tableau de distribution de cette série, on remarquera que si on fait basculer une seule donnée de l'intervalle [30 ; 40[vers l'intervalle [20 ; 30[, la classe modale changera et deviendra [20 ; 30[. Ainsi le mode passera de 35 ans à 25 ans.

5.2 Mesures de dispersion

Les paramètres de tendance centrale ne peuvent pas toujours résumer la série des données. Par exemple, dans une classe d'étudiants, on aura la même moyenne si tous les étudiants avait dans le premier contrôle 10 sur 20, et dans le deuxième contrôle 5 sur 20 pour 50% d'étudiants et 15 sur 20 pour les 50% qui restent. Dans les deux contrôles, on a l'impression que le résultat est le même puisque la moyenne des notes est la même, mais on remarquera que la dispersion des notes autour de la moyenne diffère entre les deux contrôles.

Une mesure de dispersion permet d'étudier la dispersion des observations, son objectif est de trouver un indicateur de cette dispersion. Il faut noter qu'un indicateur de dispersion est toujours supérieur ou égal à 0. On dira qu'il n'y a pas de variabilité dans les observations si cette dispersion est égale à 0.

Si on résume les données x_i par leur moyenne $\bar{x}_n$. La moyenne serait la valeur qui représentera toutes les n données x_i , ainsi on commet n erreurs $x_i - \bar{x}_n$

$$\begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} \rightarrow \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \rightarrow \begin{matrix} x_1 - \bar{x}_n \\ x_2 - \bar{x}_n \\ \vdots \\ x_n - \bar{x}_n \end{matrix}$$

On doit de même résumer les n erreurs, on peut proposer

$$\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}_n|$$

Ou peut être

$$\left(\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}_n|^p \right)^{\frac{1}{p}}$$

En particulier si $p = 2$, cette mesure sera appelée l'écart-type.

5.2.1 L'étendue

C'est la 1^{ère} mesure de la variabilité des données, elle ne reflète pas la réalité des données puisqu'elle est la différence de deux valeurs, la plus petite et la plus grande, qui peuvent coïncider avec des valeurs extrêmes.

Exemple : le résultat d'un examen d'un petit groupe d'étudiants est le suivant :

$$11, 11, 10, 10, 10, 11, 01, 10, 17,$$

L'étendue est égale à : $e = 17 - 1 = 16$.

Tandis que presque toutes les données sont entre 10 et 11, on a l'impression qu'il y a une grande variabilité entre les données puisque $e = 16$.

5.2.2 La variance

Elle évalue la distance entre la moyenne et toutes les données. Dans le cas où les données de toute la population seraient connues, on définit la variance qu'on note en général par σ^2 (sigma au carré) :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

avec $\mu = \frac{1}{N} \sum_{i=1}^N x_i$ la moyenne de toute la population.

Pour le calcul on retiendra la formule suivante :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$$

La variance est toujours positive puisque c'est la somme de carrés. Dans le cas où on n'aurait pas toutes les données, sauf celles d'un échantillon, on calcule la variance (qu'on notera s^2 ou s_n^2) sur l'échantillon en remplaçant dans la formule de σ^2 , μ par $\bar{x}$ et N par n .

Exemple :

Distribution des foyers américains selon le nombre de voitures possédées

Nombre de voitures x_i^*	x_i^{*2}	Nombre de foyers n_i	$n_i x_i^*$	$n_i x_i^{*2}$
0	0	2	0	0
1	1	19	19	19
2	4	25	50	100
3	9	16	48	144
4	16	7	28	112
5	25	3	15	75
Total		n = 72	160	450

Les données sont regroupées en classes, on a :

$$s^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^k n_i x_i^{*2} - \left(\frac{\sum_{i=1}^k n_i x_i^*}{n} \right)^2$$

Si on fait le calcul, on aura la variance égal à :

$$s^2 = \frac{450}{72} - \left(\frac{160}{72} \right)^2 = 1.312$$

Cas des données groupées en classes : (cas d'un caractère continu)

On représente chaque classe par son centre et on définit une approximation de σ^2 :

$$\sigma^2 \cong \frac{1}{N} \sum_{i=1}^k n_i (m_i - \mu)^2$$

Qui peut être exprimé encore par :

$$\sigma^2 \cong \frac{1}{N} \sum_{i=1}^k n_i m_i^2 - \mu^2$$

Exemple :

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Centre m_i	m_i^2	Effectif n_i	$n_i m_i$	$n_i m_i^2$
[7.0 ; 8.4[	7.7	59.29	6	46.2	355.74
[8.4 ; 9.8[	9.1	82.81	6	54.6	496.86
[9.8 ; 11.2[	10.5	110.25	16	168	1764.00
[11.2 ; 12.6[	11.9	141.61	15	178.5	2124.15
[12.6 ; 14[	13.3	176.89	8	106.4	1415.12
[14 ; 15.4[	14.7	216.09	4	58.8	864.36
[15.4 ; 16.8[	16.1	259.21	5	80.5	1296.05
Total			60	693	8316.28

On fait le calcul pour obtenir la variance :

$$\sigma^2 \cong \frac{8316.28}{60} - \left(\frac{693}{60}\right)^2 = 5.202$$

5.2.3 L'écart type

La variance qu'on a exprimée auparavant est d'autant plus grande que les données s'éloignent de la moyenne, cette mesure est exprimée au carré de l'unité de mesures des données, c'est pour cette raison qu'on a besoin de l'exprimer dans l'unité des données en introduisant une mesure de dispersion qui est la racine carrée de la variance.

On note σ l'écart type associé à la population et s l'écart type associé à l'échantillon. On pose :

$$s = \sqrt{s^2} \text{ et } \sigma = \sqrt{\sigma^2}$$

Cette mesure, par définition, est positive.

Exemple : On calcule l'écart type dans l'exemple de la distribution des foyers américains selon le nombre de voitures possédées, on trouve la valeur $s = 1.145$.

De même pour la distribution des 60 cadres selon leur ancienneté, on trouve la valeur $\sigma = 2.28$.

5.2.4 Écart absolu moyen

L'écart entre la $i^{\text{ème}}$ observation x_i et la moyenne de la série $\bar{x}$ est

$$x_i - \bar{x}$$

On s'intéresse à la valeur absolue $|x_i - \bar{x}|$ (la mesure de l'écart doit être positive). Une mesure de dispersion de la série autour de la moyenne est la moyenne des écarts absolus ou dit *écart absolu moyen* :

$$EAM = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Dans le cas de données groupées en classes, on aura :

$$EAM = \frac{1}{n} \sum_{i=1}^K n_i |x_i^* - \bar{x}|$$

Exemple :

Distribution de 60 cadres selon leur ancienneté

Ancienneté en année	Centre m_i	Effectif n_i	$ m_i - \bar{x} $	$n_i m_i - \bar{x} $
[7.0 ; 8.4[	7.7	6	3.85	23.10
[8.4 ; 9.8[	9.1	6	2.45	14.70
[9.8 ; 11.2[	10.5	16	1.05	16.80
[11.2 ; 12.6[	11.9	15	0.35	5.25
[12.6 ; 14[	13.3	8	1.75	14.00
[14 ; 15.4[	14.7	4	3.15	12.60
[15.4 ; 16.8[	16.1	5	4.55	22.75
<i>Total</i>		60		109.20

On connaît que $\mu = 11.55$; en utilisant le calcul fait dans le tableau ci-dessus, on aura :

$$EAM = \frac{109.20}{60} = 1.82$$

Remarque : Un critère de comparaison entre les différentes mesures de tendances centrales est le calcul de la moyenne des erreurs soit par la variance ou par l'écart absolu moyen.

5.2.5 Coefficient de variation

Il est dit encore coefficient de dispersion relative, ce coefficient est par définition le rapport, en valeur absolue, de l'écart type à la moyenne, rapport exprimé en pourcentage.

Notée V , cette mesure permet de comparer la dispersion entre deux séries de données rattachée à des caractères qui peuvent être tout à fait différents.

Il permet ainsi de comparer un caractère sur des populations différentes avec des moyennes différentes ou exprimés avec des unités différentes ; par exemple comparer le prix d'un article vendu au Maroc en milliers de dirhams avec son prix vendu en Europe en centaines d'euros.

Plus grande est l'homogénéité des données d'une série par rapport à la moyenne, plus faible sera le coefficient de variation du caractère. Inversement, plus étalées sont les données par rapport à la moyenne, plus grande sera la valeur du coefficient de variation.

Dans le cas d'une population, on pose donc

$$V = \left| \frac{\sigma}{\mu} \right| * 100$$

où σ est l'écart type et μ la moyenne.

Dans le cas d'un échantillon, on a :

$$V = \left| \frac{s}{\bar{x}} \right| * 100$$

Puisqu'on évalue la valeur absolue du rapport de l'écart type à la moyenne, le coefficient de variation est une valeur qui est toujours supérieur ou égale à 0.

Exemple : comparaison de salaires

On considère deux entreprises E1 et E2 qui emploient chacune 150 salariés. Dans l'entreprise E1 le salaire moyen est égal à 5000 DH avec un écart-type de 1800 DH. Tandis que dans l'entreprise E2 le salaire moyen est égal à 8500 DH avec un écart-type de 2500 DH. On veut comparer les coefficients de variation des deux entreprises :

$$\text{Pour E1 : } V_1 = \left| \frac{\sigma_1}{\mu_1} \right| * 100 = 36\% \text{ et pour E2 : } V_2 = \left| \frac{\sigma_2}{\mu_2} \right| * 100 = 29.41\%$$

La dispersion des salaires dans l'entreprise E1 est plus élevée que dans l'entreprise E2.

5.3 Mesures de position

Les mesures de position situent les données par rapport à l'ensemble de la série ordonnée selon un classement ascendant ou descendant. On distingue trois catégories particulières de mesures de position. Ce sont les quantiles, la cote Z et les rangs. On verra ici seulement les deux premiers.

5.3.1 Les quantiles

Les quantiles correspondent aux valeurs numériques qui occupent des positions bien précises dans le classement ascendant (croissant) des données d'une série.

Définition1 : le quantile d'ordre α (alpha), noté q_α , est la valeur d'un caractère pour laquelle on retrouve une proportion α de données qui lui sont inférieures ou égales dans la série.

Remarque : la médiane correspond au quantile d'ordre 50%, on peut écrire $Me = q_{0.5} = q_{50\%}$.

Parmi les quantiles les plus utilisés, on retrouve les quartiles, les déciles et les centiles.

Définition2 : les quartiles sont les 3 valeurs qui divisent une série en quatre parties contenant chacune, autant que possible, 25% des données. Ces valeurs sont notées Q_1, Q_2 et Q_3 , ou $q_{1/4}, q_{2/4}$ et $q_{3/4}$ respectivement.

Définition3 : Les déciles sont les 9 valeurs qui divisent une série en 10 parties contenant chacune, autant que possible, 10% des données. Ces valeurs sont notées $D_1, D_2, \dots, D_9$ ou $q_{1/10}, q_{2/10}, \dots, q_{9/10}$ respectivement.

Définition4 : Les centiles sont les 99 valeurs qui divisent une série en 100 parties contenant chacune, autant que possible, 1% des données. Ces valeurs sont notées $C_1, C_2, \dots, C_{99}$ ou $q_{1/100}, q_{2/100}, \dots, q_{99/100}$ respectivement.

Remarque : de façon pratique, les quantiles prennent vraiment leurs sens quand le nombre de données d'une série s'avère assez considérable.

5.3.2 Intervalle et écart interquartile

L'intervalle interquartile est l'intervalle entre le premier et le troisième quartile Q_1 et Q_3 . Il contient 50% des observations. L'écart interquartile est la différence entre Q_1 et Q_3 .

Exemple : on a calculé les quartiles d'une série de 50 données

$$Q_1 = q_{1/4} \quad 51.79$$

$$Q_2 = q_{2/4} \quad 59.65$$

$$Q_3 = q_{3/4} \quad 65.02$$

L'écart interquartile est égal à : $Q_3 - Q_1 = 65.02 - 51.79 = 13.23$

L'intervalle interquartile $]Q_1; Q_3] =]51.79; 65.02]$ contient 50% des observations.

Définition d'une valeur aberrante : une valeur aberrante de la série est toute donnée qui s'écarte d'une façon remarquable de l'ensemble des données.

On donne une règle pratique et qui est dû à Tukey, pour décider si une donnée est une valeur aberrante ou non.

Règle de Tukey : une donnée peut être appelé valeur aberrante si elle s'écarte d'une distance d'au moins $1.5 * (Q_3 - Q_1)$ au-dessus de Q_3 ou en-dessous de Q_1 .

Calcul d'un quantile pour des données non groupées en classes

Préalablement à tout calcul, il faut réaliser le classement ascendant des n données de la série statistique. Par rapport à ce classement, la position du quantile d'ordre alpha est déterminée par

$$p = \alpha n$$

Si p est un nombre entier, on aura :

$$q_\alpha = \frac{x_p + x_{p+1}}{2}$$

Dans le cas où p n'est pas un nombre entier, on aura

$$q_\alpha = x[p]$$

où $[p]$ est le plus petit entier supérieur à p .

Exemple : on donne la série des données suivante,

11	2	0	9	12	12	11	14	20	10
----	---	---	---	----	----	----	----	----	----

On doit classer les données dans un ordre croissant :

0	2	9	10	11	11	12	12	14	20
---	---	---	----	----	----	----	----	----	----

$$Q_1 = q_{1/4} = x_3 = 9$$

$$Q_2 = q_{1/2} = \frac{x_5 + x_6}{2} = 11$$

$$Q_3 = q_{3/4} = x_8 = 12$$

Y-a-t-il des données aberrantes dans la série ?

On calcule,

$$1.5 (Q_3 - Q_1) = 1.5 (12 - 9) = 4.5$$

Toute valeur au-dessous de $Q_1 - 1.5(Q_3 - Q_1) = 9 - 4.5 = 4.5$ ou au-dessus de $Q_3 + 1.5(Q_3 - Q_1) = 12 + 4.5 = 16.5$ est une valeur aberrante, on décide que les trois valeurs 0, 2 et 20 sont aberrantes.

Si on veut étudier une série de données, il est préférable d'éliminer l'effet des valeurs aberrantes. On donne l'exemple du calcul de la moyenne et de la variance, avant et après l'élimination des trois valeurs aberrantes :

Les dix données ont une moyenne de 10.10 avec une variance de 29.09. Si on reprend le calcul avec les sept valeurs non aberrantes, on aura une moyenne de 11.29 avec une variance de seulement 2.20.

Il y a moins de dispersion dans la série quand on a éliminé les valeurs aberrantes.

Calcul d'un quantile pour des données groupées en classes

Sur la base d'une distribution d'effectifs ou de fréquences, l'approximation du quantile d'ordre alpha q_α est donnée par la formule

$$q_\alpha = L_i + \left[\frac{\alpha n - N_{i-1}}{n_i} \right] * c$$

où n la taille de l'échantillon,

L_i la limite inférieure de la classe contenant la valeur du quantile cherché,

c l'amplitude de la même classe,

n_i l'effectif de la même classe,

N_{i-1} est l'effectif cumulé de la classe précédente la classe contenant la valeur du quantile cherché.

On peut utiliser la formule des fréquences au lieu des effectifs

$$q_\alpha = L_i + \left[\frac{\alpha - F_{i-1}}{f_i} \right] * c$$

f_i la fréquence de la même classe contenant la valeur du quantile cherché,

F_{i-1} est la fréquence cumulée de la classe précédente la classe contenant la valeur du quantile cherché.

Remarque : il est facile de voir que certains quantiles se coïncident. Par exemple :

$$q_{1/2} = Q_2 = D_5 = C_{50} = Me$$

$$q_{1/10} = D_1 = C_{10}$$

ou

$$q_{3/4} = Q_3 = C_{75}$$

Exemple : on reprend l'exemple de la série de notes d'informatique (page 20)

Note sur 20	Nombre d'étudiants n_i	Effectif cumulé N_i	Fréquence f_i	Fréquence cumulée F_i
[00 ; 10[	135	135	0.625	0.625
[10 ; 12[	43	178	0.199	0.824
[12 ; 14[	16	194	0.074	0.898
[14 ; 16[	13	207	0.060	0.958
[16 ; 20[	09	216	0.042	1.000
Total	216		1.000	

$$Q_1 = q_{1/4} = 0 + \left[\frac{54 - 0}{135} \right] * 10 = 4$$

$$Q_2 = q_{1/2} = 0 + \left[\frac{108 - 0}{135} \right] * 10 = 8$$

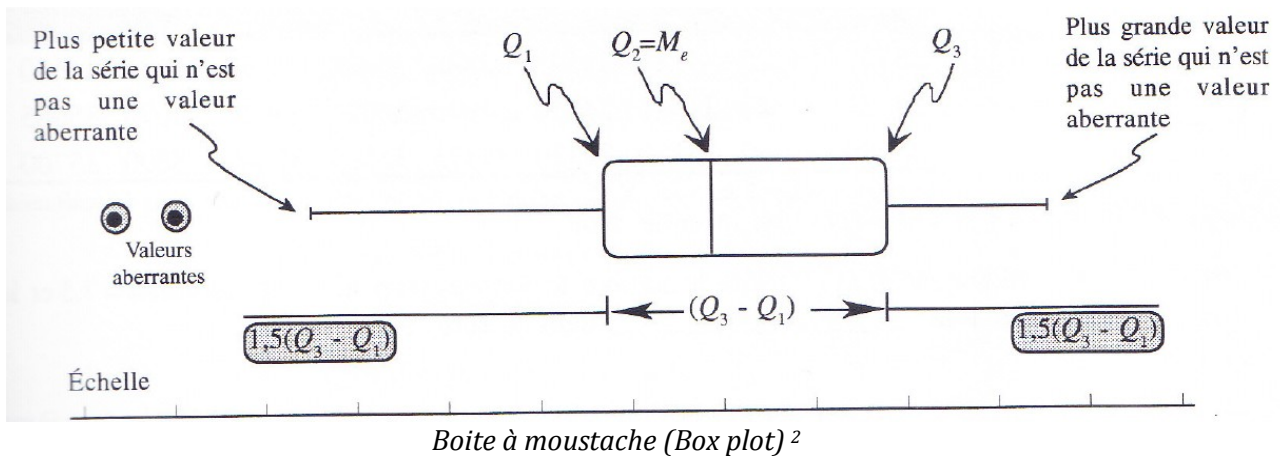
$$Q_3 = q_{3/4} = 10 + \left[\frac{162 - 135}{43} \right] * 2 = 11.256$$

$$1.5(Q_3 - Q_1) = 1.5 (11.256 - 4) = 10.884$$

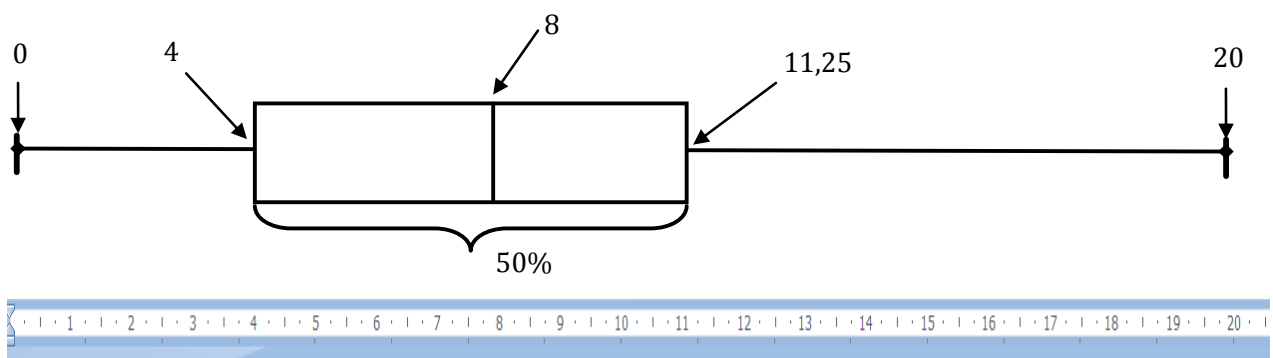
On cherche des valeurs aberrantes au-dessous de $Q_1 - 1.5(Q_3 - Q_1) = 4 - 10.88 = -6.88$ ou au-dessus de $Q_3 + 1.5(Q_3 - Q_1) = 11.256 + 10.88 = 22.136$, donc aucune valeur n'est aberrante.

5.3.3 Construction d'un diagramme en boîte à moustache (Box plot)

En calculant les trois quartiles d'une série, on peut faire une représentation des données sous une forme de boîte avec deux moustaches (voir diagramme suivant). La représentation est faite sur l'axe des observations, en rapportant dans l'ordre croissant, la plus petite valeur non aberrante, puis le premier, le deuxième et le troisième quartile et en dernier on ajoute la plus grande valeur non aberrante :



Exemple : dans le dernier exemple, on a pas trouvé de valeurs aberrantes, on va tracer une boite à moustaches des notes d'informatique



5.3.4 La cote Z

Egalement appelée cote standard, la cote Z situe essentiellement une valeur par rapport à la moyenne de la série d'où elle provient.

Définition : la cote Z d'une valeur représente le nombre d'écarts types séparant cette valeur de la moyenne de la série.

Ainsi, pour une valeur quelconque x_i , on a

$$\text{Cote } Z = \frac{x_i - \bar{x}}{s},$$

dans le cas des mesures échantillonnages, et

$$\text{Cote } Z = \frac{x_i - \mu}{\sigma},$$

dans le cas où les données sont celles de toute une population.

Remarque :

- Une cote Z égale à zéro indique un résultat équivalent à la moyenne
- Une cote Z négative signifie que le résultat est inférieur à la moyenne

² G. Baillargeon (2004), *Outils statistiques et analyse de données*, les éditions SMG, page 101.

- Une cote Z positive signifie que le résultat est supérieur à la moyenne
- Une cote Z est rarement inférieur à -3 ou supérieur à 3 .
- La cote Z permet de comparer la position relative de données appartenant à des distributions différentes, tel le *rendement d'élèves* issues de groupes ou de collèges différents.

Exemple : Dans un concours de sélection pour un master, on veut classer les étudiants suivants par ordre de mérite.

Nom	Note Moyenne en Licence	Moyenne de la promotion	Écart-type de la promotion
Karim	11.5	10.1	2.5
Asmae	12.3	11.5	3
Ilham	10.8	9.3	1.5
Ali	13	11.5	3

En regardant les notes moyennes en Licence, on peut observer que la note la plus élevée est obtenu par Ali et la plus petite est obtenue par Ilham, ce qui donne le classement suivant :

Classement des étudiants par ordre de mérite

Nom	Note Moyenne en Licence
Ali	13
Asmae	12.3
Karim	11.5
Ilham	10.8

En réalité dans ce classement on ne prend pas considération que les étudiants viennent de différents établissements où on peut remarquer des différences dans les régimes d'évaluations des étudiants.

Pour comparer les étudiants indépendamment de leurs provenances on doit comparer les cotes des étudiants. Cela permettra de réaliser un peu plus d'équité entre les différents étudiants.

Nom	Note Moyenne en Licence	Moyenne de la promotion	Écart-type de la promotion	Cote de l'étudiant
Karim	11.5	10.1	2.5	$\frac{11.5 - 10.1}{2.5} = \frac{1.4}{2.5} = 0.56$
Asmae	12.3	11.5	3	0.26
Ilham	10.8	9.3	1.5	1
Ali	13	11.5	3	0.5

Le nouveau classement met en tête de liste Ilham et Asmae en dernier.

Classement des étudiants par ordre de mérite

Nom de l'étudiant	Cote de l'étudiant
Ilham	1
Karim	0.56
Ali	0.5
Asmae	0.26

5.4 Mesures de forme

5.4.1 Représentation des données selon une courbe normale

C'est la distribution selon la loi normale $N(\mu, \sigma)$ (on lui donne d'autres noms comme la loi gaussienne ou la loi de Laplace-gauss), μ et σ représentent respectivement la moyenne et l'écart type de la population, ces deux paramètres sont en général inconnus. La courbe qui représente cette distribution est donnée par la fonction suivante :

$$x \mapsto f_N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

avec $\pi = 3.14159 \dots$

Cette distribution a la forme d'une cloche symétrique par rapport à l'axe $x = \mu$, ses extrémités sont tangentes à l'axe des abscisses.

Les observations sont groupées autour de la moyenne, et on montre que :

50% de ces données sont dans l'intervalle $\left[\mu - \frac{2}{3}\sigma; \mu + \frac{2}{3}\sigma\right]$

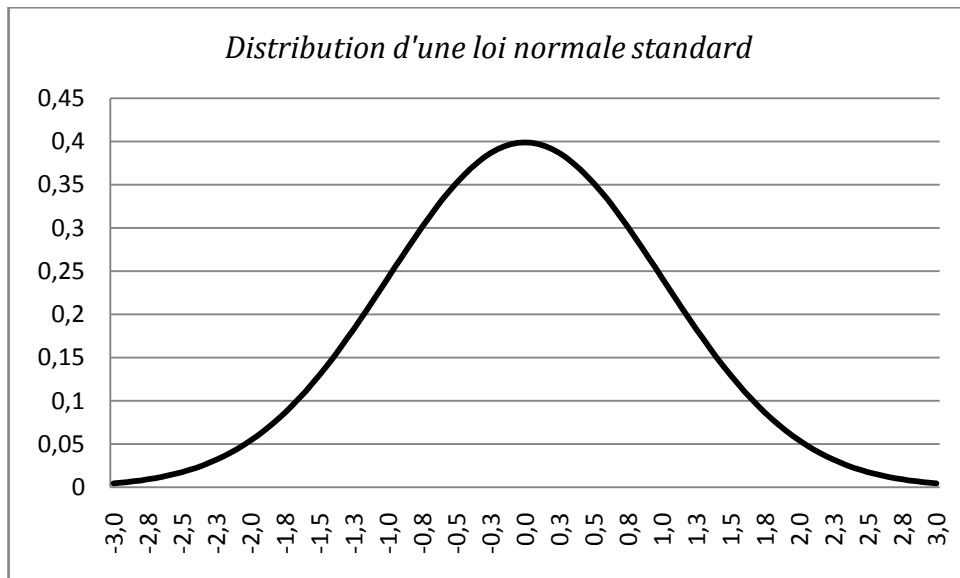
68% sont dans l'intervalle $[\mu - \sigma; \mu + \sigma]$

95% sont dans l'intervalle $[\mu - 2\sigma; \mu + 2\sigma]$

99.7% sont dans l'intervalle $[\mu - 3\sigma; \mu + 3\sigma]$

La quasi-totalité des données sont rassemblées dans un rayon de 3σ . Plus l'écart type est grand, plus les données sont dispersées autour de la moyenne.

La loi normale la plus utilisée en pratique est la loi $N(0,1)$ dite loi normale standard, toutes les autres lois normales se ramènent à la loi normale standard.



Le but est de comparer des séries de moyennes et/ou de dispersions différentes :

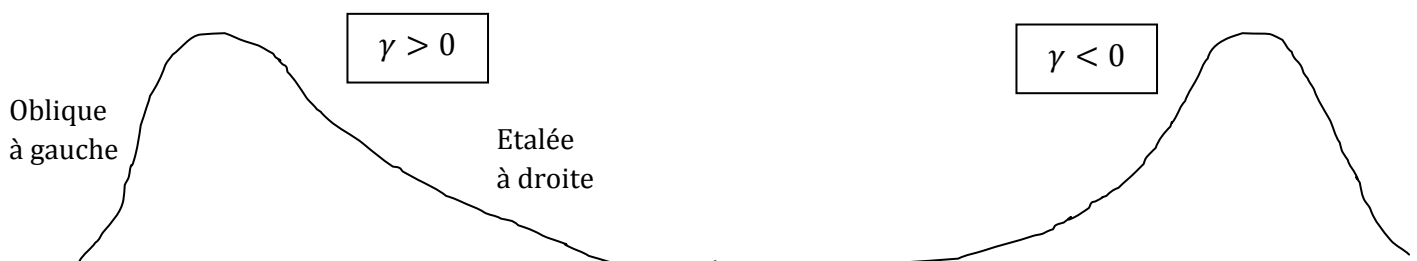
5.4.2 Coefficients d'asymétrie et d'aplatissement

- **Coefficient d'asymétrie de Fisher (ou skewness)**

$$\gamma = \frac{M_3^*}{\sigma^3}$$

où $M_3^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$ est le moment centré d'ordre 3.

- $\gamma = 0$ si la distribution est symétrique par rapport à la moyenne $\bar{x}$;
- $\gamma > 0$ si la distribution est plus étalée à droite qu'à gauche de la moyenne ;
- $\gamma < 0$ si la distribution est plus étalée à gauche qu'à droite de la moyenne.



Remarque : si la distribution est totalement symétrique, il faudra remarquer que la médiane n'est autre que la moyenne.

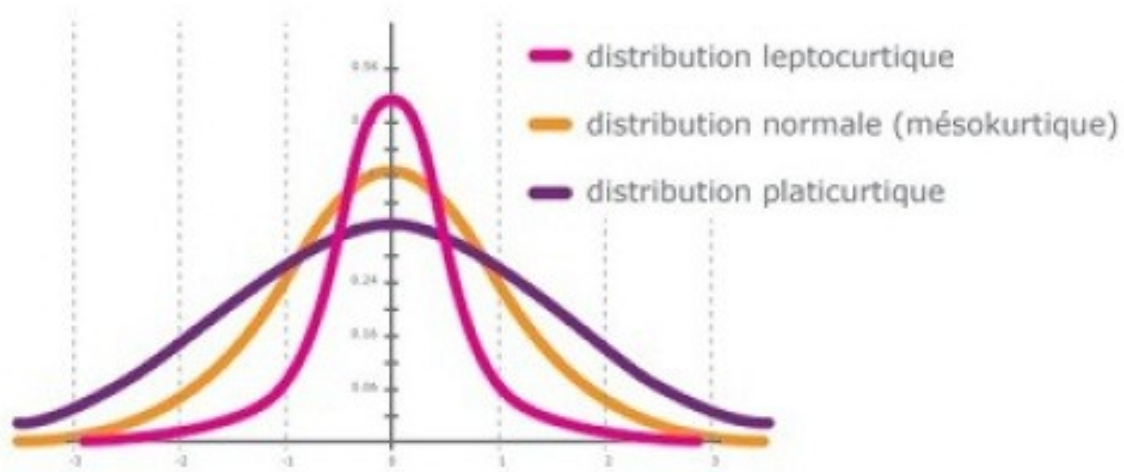
- **Coefficient d'aplatissement de Fisher (ou Kurtosis)**

$$\Gamma = \frac{M_4^*}{\sigma^4} - 3$$

48

Où $M_4^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$ est le moment centré d'ordre 4.

- $\Gamma = 0$ si l'aplatissement est le même que celui d'une loi normale standard ; on dit que la courbe est mesokurtique ;
- $\Gamma > 0$ si la distribution est moins aplatie ; on dit que la courbe est leptokurtique ;
- $-2 \leq \Gamma < 0$ si la distribution est plus aplatie ; on dit que la courbe est platikurtique .



Exemple : On reprend l'exemple de l'étude de l'ancienneté des cadres ; on a les informations suivantes : $\sigma = 2.281$ d'où $\sigma^3 = 11.865$ et $\sigma^4 = 27.06$. Et les moments d'ordre 3 et 4 calculés sont $M_3^* = 3.17$ et $M_4^* = 69.43$; ce qui donne

$$\gamma = \frac{3.17}{12.167} = 0.26 > 0$$

La distribution n'est pas symétrique, elle est plus étalée à droite, on peut bien le remarquer sur le polygone de fréquences ou l'histogramme.

Le coefficient d'aplatissement

$$\Gamma = \frac{69.43}{27.984} - 3 = -0.52 < 0$$

c.à.d. que la courbe est platikurtique : la distribution est plus aplatie que celle de la loi normale standard.

5.5 Mesures de concentration

Les mesures de dispersion (variance, écart-type ou coefficient de variation) nous informent sur la concentration des données autour de leur moyenne.

Quand il s'agit d'un caractère quantitatif dont l'accumulation a une signification, on cherche alors à comparer sa distribution avec la distribution uniforme, on parle alors d'équilibre parfait

(distribution uniforme) lorsque dans des intervalles de même amplitude, on a les mêmes proportions ou fréquences des individus.

On donne l'exemple de la distribution des revenus. C'est une distribution caractérisée par une concentration plus grande d'individus de revenus bas et une concentration plus faible pour les revenus plus élevés.

5.5.1 Courbe de Lorenz (1876-1959)

Lorenz a proposé en 1905 un graphique pour représenter la concentration des revenus.

En abscisse, on met le pourcentage de la population des salariés et en ordonnée le pourcentage de revenus total appartenant à cette population.

Prenons un exemple très simple pour illustrer la construction de la courbe de Lorenz. Le revenu journalier de 10 salariés est le suivant :

Nombre de salariés	Revenu journalier (DH)	Revenu total (DH)
5	80	400
3	266.66	800
2	400	800
10		2000

Le revenu moyen des 10 salariés est le suivant :

$$\bar{x} = \frac{2000}{10} = 200 \text{ DH}$$

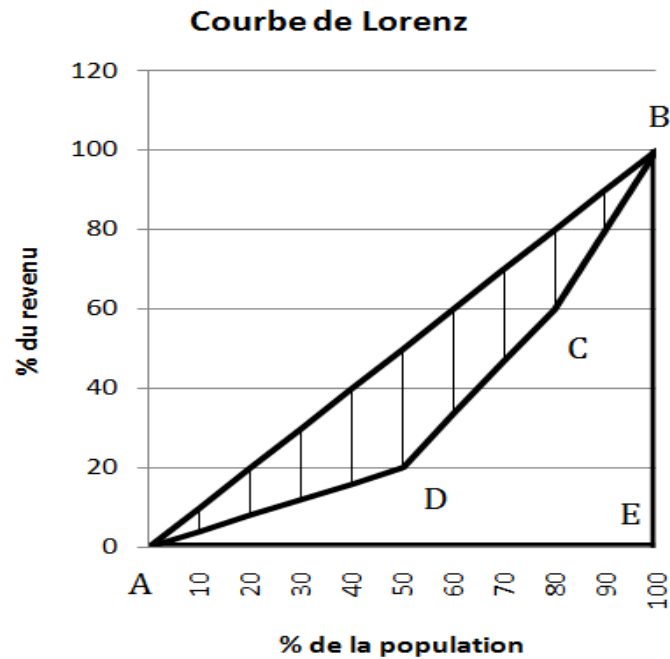
Cette distribution n'est pas symétrique, on voit bien que 50% de salariés touchent 80 DH par jour, loin des 200DH de moyenne. Il semblerait qu'il y a une concentration de la masse des revenus dans les deux autres classes qui représentent 50% des salariés.

Si on compare les pourcentages cumulés, on obtient le tableau suivant :

% salariés n_i	% cumulés N_i	% revenu r_i	% cumulés R_i
50%	50%	20%	20%
30%	80%	40%	60%
20%	100%	40%	100%
100%		100%	

On lit bien que 50% de salariés touchent 20% de la masse totale des revenus, que 80% touchent 60% de la masse totale et qu'à la fin 100% des salariés (toute la population) touchent la totalité du salaire globale.

On présente la courbe de Lorenz comme suit :



5.5.2 Indice de Gini

Corrado Gini (1884-1965) a proposé un indice pour mesurer l'inégalité. Il est calculé en prenant le rapport entre la surface ABCD et le triangle ABE :

$$I_G = \frac{ABCD}{ABE}$$

Si la distribution est uniforme, la courbe de Lorenz suit la diagonale et alors le coefficient de Gini est égal à 0.

Par contre, l'inégalité la plus forte est obtenue lorsque la courbe de Lorenz est AEB. Dans ce cas l'indice de Gini est égal à 1.

Calcul du coefficient de Gini dans l'exemple précédent :

La surface du triangle ABE est la moitié de celle du carré de surface 1 ; la surface de ABCD est celle de ABE en retranchant les surfaces d'un triangle et de deux trapèzes,

$$\begin{aligned}
 ABCD &= 0.5 - \frac{50\% * 20\%}{2} + \frac{(80\% - 50\%) * (20\% + 60\%)}{2} + \frac{(100\% - 80\%) * (60\% + 100\%)}{2} \\
 &= 0.5 - (0.05 + 0.12 + 0.16) = 0.5 - 0.33 = 0.17 \\
 I_G &= \frac{0.17}{0.5} = 0.34
 \end{aligned}$$

Remarque : un indice de Gini proche de 0 est un indicateur d'une parfaite égalité des revenus dans la population, un coefficient qui s'approche de 1 reflète un déséquilibre dans la répartition des revenus dans la population.

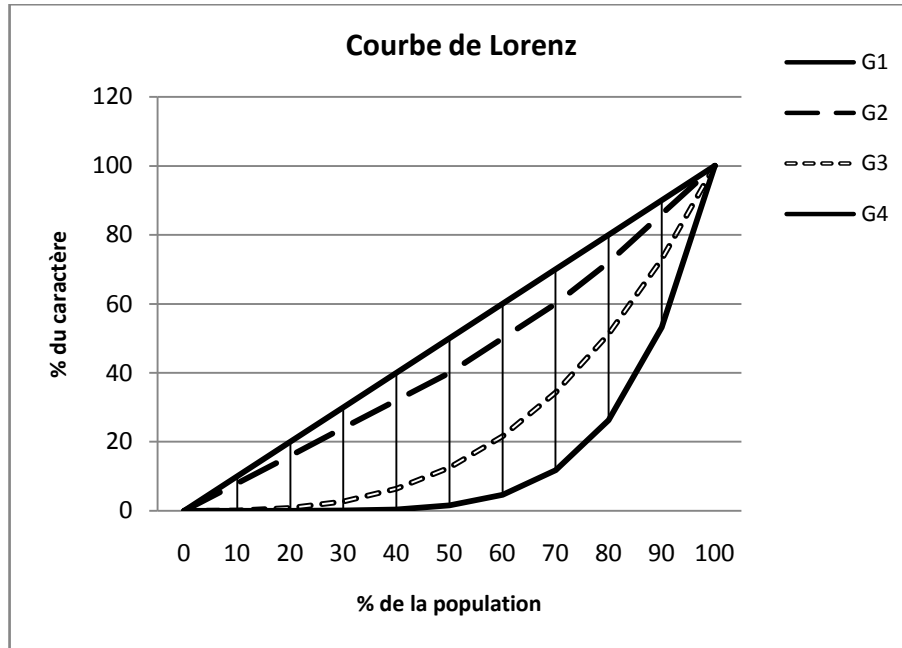
On donne dans l'exemple ci-dessous quatre courbes de Lorenz avec différents coefficients de Gini :

G1
$I_G = 0$

G2
$I_G = 12.40\%$

G3
$I_G = 49.50\%$

G4
$I_G = 70.43\%$



• Calcul de l'indice de Gini

Cas de données non groupées en classes :

Si on classe les revenus du plus petit au plus grand c.à.d. qu'on a : $y_1 \leq y_2 \leq \dots \leq y_n$ alors on peut calculer le coefficient de Gini par la formule suivante :

$$I_G = \frac{2 \sum_{i=1}^n i y_i}{n \sum_{i=1}^n y_i} - \left(\frac{n+1}{n} \right)$$

n = nombre de personnes

y_i = le $i^{\text{ème}}$ revenu (du plus petit au plus grand)

Exemple :

On revient à l'exemple précédent :

$$I_G = \frac{2[80(1+2+3+4+5) + 266.66(6+7+8) + 400(9+10)]}{10 * 2000} - \frac{11}{10} = \frac{28799.72}{20000} - 1.1$$

$$I_G = 0.34$$

Cas de données groupées en classes

Si les données sont réunies en k classes, on supposera que tous les individus de la même classe ont même revenu, celui de la moyenne de la classe.

L'indice de Gini, dans ce cas, sera approché par la formule suivante :

$$I_G = \frac{1}{\sum_{j=1}^n y_j} \sum_{i=1}^k f_i m_i \left[2 \sum_{j=1}^i n_j - n_i \right] - 1 = \left[\frac{1}{n \mu} \sum_{i=1}^k f_i m_i \{2N_i - n_i\} \right] - 1$$

- m_i est le centre de la $i^{\text{ème}}$ classe
- f_i est la fréquence de la $i^{\text{ème}}$ classe

Exemple : On veut calculer l'indice de Gini d'une entreprise de 110 salariés, la distribution des salaires horaires est donnée dans le tableau suivant,

Salaire en DH/H	m_i	n_i	f_i	N_i	$f_i m_i$	$f_i m_i \{2N_i - n_i\}$
[10 , 12[	11	15	0.14	15	1.50	22.50
[12 , 15[	13.5	35	0.32	50	4.30	279.20
[15 , 20[	17.5	45	0.41	95	7.16	1038.07
[20 , 25[	22.5	10	0.09	105	2.05	409.09
[25 , 30[	27.5	5	0.05	110	1.25	268.75
<i>Total</i>		110	1.00		16.25	2017.61
$\sum_{j=1}^n y_j = 1787.50$						

$$I_G = \frac{2017.61}{1787.5} - 1 = 12.87\%$$

BIBLIOGRAPHIE

1. D.R. Anderson, D.J. Sweeney, T.A. Williams, (2010) Statistiques pour l'économie et la gestion, De Boeck.
2. D. Audet, C. Boucher, A. Caumartin, C. Skeene, (1993) Probabilités et statistique, gaëtan morin.
3. G. Baillargeon (2004), Outils statistiques et analyse de données, les éditions SMG.
4. J.-M. Martel (1972) Décision et inférence statistique en affaires, Les Presses de l'université Laval.
5. P.-C. Pupion (2008) Statistiques pour la gestion, applications avec Excel et SPSS, Gestion Sup, Dunod