

STATISTIQUE DESCRIPTIVE



GHIZLAN LOUMRHARI
FÉVRIER 2025

La séance d'aujourd'hui



- ☐ Prise de contact
- ☐ Plan du cours
- ☐ Bibliographie
- ☐ Déroulement du cours
- ☐ Evaluation
- ☐ Introduction générale

Tables des matières



Partie 1. Statistiques descriptive univariée.

Chapitre 1 : Les caractéristiques de tendance centrale

Chapitre 2 : Les caractéristiques de dispersion

Chapitre 3 : Les caractéristiques de forme

Chapitre 4 : Les caractéristiques de concentration

Partie 2. Calcul des probabilités, variables aléatoires et lois usuelles

Chapitre 1 : Généralités sur le calcul des probabilités

Chapitre 2 : Les variables aléatoires

Chapitre 3 : Lois de probabilités discrètes usuelles

Chapitre 4 : Lois de probabilités continues usuelles

Chapitre 5. Théorèmes limites

Références (sélectives)



- ❑ Gérard Calot, « Cours de statistique descriptive », Dunod.
- ❑ Bernard Py , « Statistique descriptive », Economica.
- ❑ Murray R. Spiegel et Larry J. Stephens, « Statistiques », Mc Graw Hill.
- ❑ Bernard Delmas « Statistique descriptive », Nathan Université.
- ❑ Fouzi Mourji et Najib Mikou « Eléments de mathématiques et statistiques pour économistes »

Déroulement des cours



- ❑ Cours est présenté sous forme de slides assez détaillés présentant les principaux concepts, théories, définitions et des applications.
- ❑ Les étudiants doivent participer activement pendant les séances à travers des implications dans la résolutions des différentes applications.
- ❑ Consacrer une séance pour corriger une planche de TD à la fin de chaque chapitre.

Définition de la statistique



Le terme statistique admet trois acceptations possibles :

- ❑ Les statistiques désignent les chiffres et les données se rapportant à un phénomène particulier. Ainsi, on parle des statistiques du chômage pour désigner, par exemple, l'évolution du taux de chômage au Maroc sur une période de temps donnée ;
- ❑ La statistique peut désigner une caractéristique d'un échantillon ; la moyenne dans un échantillon est par exemple une statistique ;
- ❑ La statistique est « une science » qui regroupe l'ensemble des méthodes numériques et mathématiques qui permettent la description, la synthèse et l'explication des données.

Définition de la statistique



La statistique en tant que science est généralement divisée en deux grandes parties :

- ❑ la statistique descriptive dont le but est la description et la présentation simplifiée des données et,
- ❑ la statistique inférentielle, ou mathématique, dont l'objet est d'appréhender les caractéristiques d'une population en se basant sur une information incomplète *i.e.* un échantillon.

Définition de la statistique



- ❑ La statistique descriptive est une « science » qui s'attache à décrire le plus simplement et le plus clairement possible les données et les informations relatives à des grandeurs observées. Pour ce faire, elle utilise des méthodes et des outils, essentiellement mathématiques, qui constituent le corps de cette « science » et qu'on peut regrouper sous l'expression de réduction de données.
- ❑ La statistique descriptive peut s'attacher à décrire les données et les informations statistiques d'un point de vue d'une seule variable, de deux variables ou de plusieurs variables. On parle respectivement de statistique univariée, bivariée ou multivariée (analyse des données).

Définition de la statistique



- ❑ Comme toutes les sciences, la statistique utilise un vocabulaire propre qu'il est nécessaire de connaître et de pouvoir utiliser avec précision.
- ❑ C'est Roland Fisher qui, le premier à utilisé l'expression *réduction des données* pour définir la statistique descriptive. Cette expression désigne le calcul de caractéristiques, qui permettent de résumer une distribution statistique.
- ❑ En réalité, la statistique fut appliquée pour la première fois à la démographie dont elle a hérité la terminologie.

Vocabulaire et mots clés



Population : On appelle population un grand ensemble au sens mathématique (fini ou infini) auquel s'intéresse la statistique. À titre d'exemples, peuvent être considérés comme des populations, les ensembles suivants :

- ❑ Les étudiants inscrits en première année à l'ENCG au titre de l'année universitaire 2023 – 2024 ;
- ❑ Le parc automobile de la région Tanger-Tetouan pour l'année 2023.
- ❑ Le nombre d'actions échangées à la bourse de Casablanca pendant une journée donnée.

Vocabulaire et mots clés



Comme on peut le constater, les ensembles considérés comme des populations sont très variés. Le plus important est que ces ensembles doivent être définis sans ambiguïté en comportant les précisions nécessaires telles que la date et le lieu comme dans nos exemples.

Vocabulaire et mots clés



Individus : On appelle individus ou unités statistiques les éléments qui composent une population. Si nous gardons nos précédents exemples, les individus ou les unités statistiques sont : les étudiants, les voitures et les actions.

Ainsi, on appelle individus les éléments qui composent une population qu'ils soient des êtres humains des objets ou toute autre chose.

Vocabulaire et mots clés



Caractère : Les individus d'une population peuvent être identifiés en fonction de plusieurs caractéristiques.

Chaque caractéristique est appelée caractère.

Un caractère est tout ce qui peut différencier un individu d'un autre : L'âge des étudiants, le niveau de qualification des employés, la couleur des voitures, par exemples, sont des caractères.

Vocabulaire et mots clés



Modalité : La modalité est la situation ou l'état dans lequel peut se trouver un caractère. Si le caractère considéré est la couleur des voitures, par exemple, les modalités sont : noire, blanche, rouge, etc. Si le caractère étudié est l'âge, les modalités seront alors 18ans, 19 ans, etc. Enfin, le caractère sexe admet deux modalités : masculin et féminin.

Vocabulaire et mots clés



Echantillon : Ainsi nous l'avons déjà précisé, la statistique s'intéresse à l'étude de grands ensembles (parfois infinis), il est souvent difficile voire impossible, ou simplement très coûteux d'étudier toute la population. Le seul moyen dont nous disposons pour avoir des informations sur la population à laquelle on s'intéresse est d'en observer une partie. Le sous-ensemble observé est appelé échantillon.

C'est pour cette raison que les recensements des populations sont très espacés dans le temps.

Vocabulaire et mots clés



Enquête : On appelle enquête l'ensemble des moyens mis en œuvre pour recueillir des informations relatives à une population.

- Lorsque l'enquête porte sur tous les individus de la population on parle d'enquête **exhaustive ou de recensement**. C'est le cas par exemple du recensement général de la population.
- Lorsque l'enquête porte sur une partie seulement de la population, on parle **d'enquête partielle ou de sondage**. C'est le cas par exemple des sondages d'opinions.
- Parmi tous ces termes qu'on vient de présenter brièvement, le concept fondamental en statistique est celui de caractère c'est pourquoi nous allons détailler .

Typologie des caractères



Un caractère est toute caractéristique qui permet de différencier et de distinguer les individus d'une même population. Si par exemple on s'intéresse à la population des entreprises installées dans la région de Tanger, il est possible de les classer et de les différencier, entre autres, par rapport aux caractéristiques suivantes :

- Effectif (nombre de salariés employés) ;
- Catégorie socioprofessionnelle ;
- Chiffre d'affaires ;
- Niveau d'étude des employés ;
- Masse salariale ;
- Secteur d'activité ;
- Nombre de filiales ;
- Sexe du PDG.

Typologie des caractères



Ces caractéristiques sont autant de caractères. Dans ces cas les modalités sont

- 20, 50, 200, 1000, etc. ;
- Ouvrier, employé, cadre, etc. ;
- 50000, 200000, 1000000 ;
- Primaire, secondaire, supérieur ;
- 30000, 469400, 750890 ;
- Agroalimentaire, textile, Industrie chimie, services, etc. ;
- 2, 3, 4 ;
- Féminin, masculin ;

Typologie des caractères



Les caractères qualitatifs

- ❑ Un caractère est dit qualitatif s'il n'est pas mesurable, ses modalités ne sont pas des valeurs numériques mais des états, des catégories, des opinions, des couleurs, etc. exemple : niveau d'études et le secteur d'activité.
- ❑ Les caractères qualitatifs peuvent être de deux types : les variables **qualitatives nominales** et les variables qualitatives ordinales. Lorsque la moyenne est impossible à calculer et elles ne peuvent pas être ordonnées (leurs classement n'a pas de sens). C'est le cas du sexe, de la nationalité ou du secteur d'activité dans notre exemple. Par contre, lorsque la moyenne peut être calculée mais ne présente aucun sens mais que le classement des modalités est possible et admet un sens et un intérêt, le caractère est dit **qualitatif ordinal**. C'est le cas du niveau d'études et de la catégorie socioprofessionnelle par exemple.

Typologie des caractères



Les caractères quantitatifs

- ❑ Un caractère est dit quantitatif lorsque ses modalités sont mesurables : des valeurs numériques et que la moyenne de ces valeurs a un sens. Dans ce cas, on parle de variables statistiques. Ces dernières se scindent en deux catégories : **les variables statistiques discrètes et les variables statistiques continues.**
- ❑ Une variable statistique est dite **discrète** lorsque ses modalités ne peuvent pas prendre n'importe quelle valeur sur un intervalle donné mais sont un ensemble fini ou dénombrable de valeurs entières et réelles. C'est le cas de la variable "nombre d'enfants par ménage" qui admet 0, 1, 2, 3, ... comme modalités. Elle ne peut pas prendre de valeur décimale ; la modalité 2,3 enfants n'a pas de sens. Dans notre exemple l'effectif est une variable statistique discrète.

Typologie des caractères



Les caractères quantitatifs

- ❑ Une variable statistique est dite **continue** lorsqu'elle peut prendre n'importe quelle valeur sur un intervalle donné c.a.d. lorsque le nombre de modalités est infini. La variable "poids d'un individu" peut prendre les valeurs 70 kg ou 80 kg par exemple, mais elle peut prendre n'importe quelle valeur intermédiaire. Ainsi, les modalités 70 kg 100 ou 73 kg 99 peuvent très bien exister. Dans notre exemple, le chiffre d'affaire et la masse salariale sont des variables continues.
- ❑ Les variables statistiques continues lorsqu'elles prennent un nombre très élevé de modalités peuvent être assimilées à des variables continues.

Chapitre 1. Les caractéristiques de tendance centrale



- ❑ Les caractéristiques de tendance centrale ont pour but de résumer les données statistiques du point de vue de l'ordre de grandeur (le niveau des variables).
- ❑ Il est possible de construire plusieurs de ces caractéristiques.
- ❑ Les plus importantes sont le mode, la médiane, la moyenne arithmétique et les autres moyennes.

Chapitre 1. Les caractéristiques de tendance centrale

1. Le mode



Définition. Le mode, noté M_o , se calcule pour tous les caractères. Il correspond à la modalité dont l'effectif dépasse simultanément les effectifs de la modalité qui la précède et de celle qui la suit. Le mode peut ne pas exister, et lorsqu'il existe n'est pas forcément unique.

Détermination. Dans le cas d'une série ou d'une distribution simple (variable discrète). La détermination du mode est immédiate et sans ambiguïté. Voici quelques exemples :

- Pour la série 3 ; 5 ; 7 ; 9 ; 11 le mode n'existe pas
- Pour la série 4 ; 6 ; 8 ; 10 ; 10 ; 12 $M_o = 10$
- Pour la série 3 ; 5 ; 5 ; 5 ; 7 ; 9 ; 9 ; 9 ; 11 $M_o = 5$ et $M_o = 9$
- Pour la série 4 ; 6 ; 6 ; 6 ; 8 ; 10 ; 10 ; 12 $M_o = 6$ mode global et $M_o = 10$ mode local

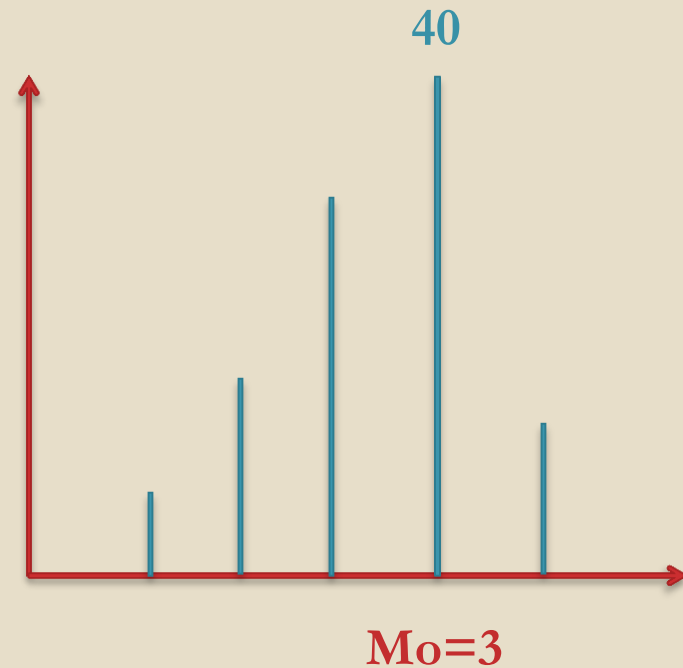
Chapitre 1. Les caractéristiques de tendance centrale

1. Le mode



Soit la distribution du nombre d'enfants par ménage suivante :

Nombre d'enfants x_i	Nombre de ménages n_i
0	5
1	15
2	30
3	40
4 et plus	10
Total	100



D'après notre définition, le nombre d'enfants par ménage modal est $M_o=3$.

Chapitre 1. Les caractéristiques de tendance centrale

1. Le mode



Dans le cas d'une variable continue (données groupées). Le mode se détermine en deux étapes :

1. On détermine d'abord une classe modale C_{M_o} de la même manière que le mode dans le cas d'une variable discrète.
2. Ensuite, on calcule le mode au sein de la classe modale par interpolation linéaire.

La méthode de l'interpolation (méthode des diagonales) rend la formule suivante :

$$M_o = b_{i-1} + a_i \left[\frac{d_i - d_{i-1}}{(d_i - d_{i-1}) + (d_i - d_{i+1})} \right]$$

Chapitre 1. Les caractéristiques de tendance centrale

1. Le mode



Exemple. Le chiffre d'affaires de 100 commerces (en milliers de DH) dans un quartier de la ville de Tanger est distribué de la manière suivante :

Chiffre d'affaires	Effectif
[0, 100[	15
[100, 300[	40
[300, 400[:	30
[400, 900[	15
Total	100

Comme les classes ne présentent pas les mêmes amplitudes il faut prendre en considération les effectifs corrigés (les densités). Les calculs sont effectués dans le tableau suivant :

	Effectif	a_i	d_i
[0, 100[	15	100	15
[100, 300[	40	200	20
[300, 400[	30	100	30
[400, 900[	15	500	3
Total	100	100	100

Chapitre 1. Les caractéristiques de tendance centrale

1. Le mode



Etape 1. La classe modale qui correspond à l'effectif le plus élevé (30) est la classe [300, 400[: le chiffre d'affaire le plus fréquemment réalisé est compris entre 300 000 DH et 400 000 DH.

Etape 2. Une approximation du chiffre d'affaire modal est donnée par la méthode des diagonales. Soit :

$$M_o = 300 + 100 \times \left[\frac{30-20}{(30-20)+(30-3)} \right] = 327,027 \equiv 327\,027\text{DH}$$

Ainsi, le chiffre d'affaires le plus fréquemment réalisé est approximativement de 327 027DH.

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Définition. Lorsque les données peuvent être ordonnées sans ambiguïté (c'est le cas des variables quantitatives ou qualitatives ordinales), on peut déterminer une caractéristique de tendance centrale appelée médiane, notée $Mé$: c'est la valeur de la modalité qui partage les observations en deux parties égales.

Détermination.

Cas d'une série ou d'une distribution discrète de taille n :

Deux cas peuvent se présenter :

1. n est impair, c'est-à-dire $n=2k+1$ alors la médiane correspond à la $(k+1)$ ème observation de la série ordonnée
2. n est pair c'est-à-dire $n=2k$ on obtient un intervalle médian et la médiane est donnée par la $\frac{2k+1}{2}$ ème observation de la série ordonnée

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Exemple : Considérons la distribution fictive suivante qui donne le nombre d'enfants par ménage

x_i	n_i	f_i	N_i^+	F_i^+
0	10	0,036	10	0,036
1	20	0,071	30	0,107
2	30	0,107	60	0,214
3	50	0,179	110	0,393
4	70	0,250	180	0,643
5	40	0,143	220	0,786
6	30	0,107	250	0,893
7	20	0,071	270	0,964
8 et plus	10	0,036	280	1,000
Total	$n = 280$	1	----	----

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Le nombre d'observations est pair (280), la médiane est la valeur de x_i qui correspond à la $\frac{2k+1}{2}$ ème observation. Comme $k = \frac{n}{2} = 140$, la médiane correspond à la 140,5ème observation. Elle est donnée par $x_i = 4$. En effet, le premier 4 est situé à la 111ème position et le dernier à la 180ème position. Ainsi, 110 couples (39,3%) ont moins de 4 enfants et 170 (60,7%) couples ont 4 enfants et plus. En fait, dans le cas discret, la médiane ne partage pas toujours l'effectif total en deux parties exactement égal.

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Cas d'une variable continue :

Dans ce cas, il est toujours possible de trouver une valeur qui partage le nombre d'observations en deux parties exactement égales. Comme pour le mode, la médiane se détermine en deux étapes :

1. On détermine la classe médiane ; de la même manière que la médiane dans le cas discret
2. On calcule la médiane par interpolation en utilisant l'histogramme ou la courbe cumulative. On peut, ce qui est équivalent utiliser une des deux formules suivantes :

$$M\acute{e} = b_{i-1} + a_i \left(\frac{\frac{n}{2} - N_{i-1}^+}{n_i} \right)$$
$$M\acute{e} = b_{i-1} + a_i \left(\frac{\frac{1}{2} - F_{i-1}^+}{f_i} \right)$$

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Exemple. Considérons la distribution du chiffre d'affaires de 650 entreprises suivante :

x_i	n_i
[250, 260[	80
[260, 270[	100
[270, 280[	160
[280, 290[	140
[290, 300[	100
[300, 310[	50
[310, 320[	20
Total	$n = 650$

On souhaite calculer le chiffre d'affaires médian

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane



Réponse. On détermine d'abord les fonctions fréquences cumulées. Les calculs sont effectués dans le tableau suivant :

x_i	n_i	f_i	N_i^+	F_i^+
$[250, 260[=650$	80	0,123	80	0,123
$[260, 270[$	100	0,154	180	0,227
$[270, 280[$	160	0,246	340	0,523
$[280, 290[$	140	0,215	480	0,738
$[290, 300[$	100	0,154	580	0,892
$[300, 310[$	50	0,077	630	0,969
$[310, 320[$	20	0,031	650	1
Total		1	----	----

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane

x_i	n_i	f_i	N_i^+	F_i^+
[250, 260[	80	0,123	80	0,123
[260, 270[	100	0,154	180	0,227
[270, 280[	160	0,246	340	0,523
[280, 290[	140	0,215	480	0,738
[290, 300[	100	0,154	580	0,892
[300, 310[	50	0,077	630	0,969
[310, 320[	20	0,031	650	1
Total		1	----	----

Etape 1. La classe médiane est donnée par la classe [270, 280[

Etape 2. La méthode de l'interpolation linéaire donne une approximation de la médiane

$$Mé = 270 + 10 \left(\frac{\frac{1}{2} - 0,227}{0,246} \right) = 279,06$$

$$Mé = 270 + 10 \left(\frac{325 - 180}{160} \right) = 279,06$$

9,06

La moitié des entreprises réalisent un chiffre d'affaires approximatif de 279 060DH.

Chapitre 1. Les caractéristiques de tendance centrale

2. La médiane / Propriétés de la médiane



Points positifs : Elle est plus stable en termes de fluctuations d'échantillonnage que la moyenne arithmétique. Elle est préférée à la moyenne arithmétique en présence de valeurs extrêmes. Elle est préférée au mode dans le cas de caractères continus.

Points faibles : Elle ne prend pas en considération l'ensemble des observations. Généralement on lui préfère le mode ou la moyenne arithmétique lorsque le caractère est discret que la médiane est une valeur qui se répète un nombre importante de fois.

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique



Définition. Pour pouvoir donner une définition intuitive de la moyenne arithmétique considérons un exemple simple. Une entreprise qui emploie 5 salariés et qui dépense, 30 000 dirhams mensuellement, en salaires. Le directeur se verse 15 000 dirhams, le comptable reçoit 7 500 dirhams et chacun des trois ouvriers 2 500 dirhams. Si le poste occupé n'avait pas d'importance, chaque employé recevrait le même salaire. Ce dernier serait de 6 000 dirhams par mois ; c'est le salaire moyen.

Définition. La moyenne arithmétique, notée $\bar{x}$, est le rapport entre la somme de toutes les observations (masse globale) et le nombre d'observations.

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique

Détermination

Dans le cas d'une série statistique $x_1, x_2, \dots, x_n$, la moyenne arithmétique est dite simple et elle est donnée par

$$\bar{x} = \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Dans le cas d'une distribution statistique (cas discret), la moyenne arithmétique est dite pondérée et s'écrit :

$$\bar{x} = \sum_{i=1}^n n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n}$$

Chapitre 2. Les caractéristiques de tendance centrale

3. La moyenne arithmétique

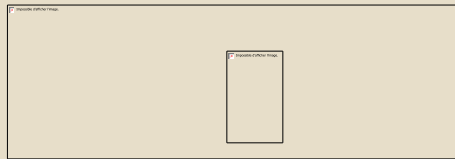
Détermination

où k est, bien entendu, le nombre de modalités.

Une formule équivalente de $\bar{x}$ est

$$\bar{x} = \sum_{i=1}^k f_i x_i = f_1 x_1 + f_2 x_2 + \cdots + f_k x_k.$$

Enfin, dans le cas continu où les données sont groupées en classes i , la moyenne arithmétique $\bar{x}$, devient



où $c_i = \frac{(b_{i-1} + b_i)}{2}$ est le centre de la classe i .

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique



Exemple. A un examen, vingt étudiants ont obtenu les notes suivantes :

11, 4, 12, 5, 10, 15, 8, 10, 17, 5, 10, 10, 5, 10, 11, 8, 11, 12, 9, 14

La note moyenne, donnée par la moyenne arithmétique simple, est donnée par

$$\begin{aligned}\bar{x} &= \frac{4 + 5 + 5 + 5 + 8 + 8 + 9 + 10 + 10 + 10 + 10 + 10 + 10 + 11 + 11 + 11 + 12 + 12 + 14 + 15 + 17}{20} \\ &= \frac{197}{20} = 9,85\end{aligned}$$

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique



Lorsqu'on examine attentivement la série, on constate que plusieurs notes se répètent. Il est alors intéressant de les regrouper dans un tableau.

Notes x_i	Effectifs n_i	Fréquences f_i	$n_i x_i$	$f_i x_i$
4	1	0,05	4	0,2
5	3	0,15	15	0,75
8	2	0,10	16	0,8
9	1	0,50	9	0,45
10	5	0,25	50	2,5
11	3	0,15	33	1,65
12	2	0,10	24	1,2
14	1	0,05	14	0,7
15	1	0,05	15	0,75
17	1	0,05	17	0,85
Total	$\sum_{i=1}^{10} n_i = 20$	$\sum_{i=1}^{10} f_i = 1$	$\sum_{i=1}^{10} n_i x_i = 197$	$\sum_{i=1}^{10} f_i x_i = 9,85$

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique



Lorsqu'on examine attentivement la série, on constate que plusieurs notes se répètent. Il est alors intéressant de les regrouper dans un tableau.

Notes x_i	Effectifs n_i	Fréquences f_i	$n_i x_i$	$f_i x_i$
4	1	0,05	4	0,2
5	3	0,15	15	0,75
8	2	0,10	16	0,8
9	1	0,05	9	0,45
10	5	0,25	50	2,5
11	3	0,15	33	1,65
12	2	0,10	24	1,2
14	1	0,05	14	0,7
15	1	0,05	15	0,75
17	1	0,05	17	0,85
Total	$\sum_{i=1}^{10} n_i = 20$	$\sum_{i=1}^{10} f_i = 1$	$\sum_{i=1}^{10} n_i x_i = 197$	$\sum_{i=1}^{10} f_i x_i = 9,85$

Comme on peut le constater, la dernière cellule du tableau donne directement la note moyenne.

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique



Admettons maintenant qu'on souhaite regrouper ces notes dans des classes d'amplitude 5, le résultat sera le suivant

Notes x_i	Centres c_i	Effectifs n_i	f_i	$n_i x_i$	$f_i x_i$
[0, 5[	2,5	1	0,05	2,5	0,125
[5, 10[	7,5	6	0,30	45	2,25
[10, 15[	12,5	11	0,55	137,5	6,87
[15, 20[	17,5	2	0,10	35	1,75
Total	-----	$\sum n_i = 20$	$\sum f_i = 1$	$\sum n_i x_i = 220$	$\sum f_i x_i = 10,99$

Là encore, la note moyenne est directement donnée par la dernière cellule du tableau. On remarque que le résultat obtenu est différent du cas précédent. En effet on peut penser que le regroupement effectué n'est pas judicieux ; plus de la moitié de l'effectif se retrouve dans une seule classe.

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique/ Propriétés

La moyenne arithmétique est la caractéristique de tendance centrale la plus utilisée en raison de ses propriétés



Se prête facilement aux manipulations algébriques

$$- \overline{x + y} = \bar{x} + \bar{y}$$

$$- \overline{ax} = a\bar{x}$$

$$- \overline{x + b} = \bar{x} + b$$

$$- \sum_{i=1}^k n_i (x_i - \bar{x}) = 0$$

- Très affectée par les fluctuations d'échantillonnage (plus que le mode et la médiane)

- On lui préfère le mode dans certains cas discrets (l'exemple typique est celui de la pointure de chaussures)

- On lui préfère la médiane dans le cas de valeurs extrêmes

Chapitre 1. Les caractéristiques de tendance centrale

3. La moyenne arithmétique/ Propriétés

$\sum_{i=1}^k n_i(x_i - \bar{x}) = 0$: Cette propriété signifie que la moyenne arithmétique est la meilleure estimation de x .

- Considérons un échantillon E qu'on peut décomposer en n sous-échantillons chacun de taille n_i . Soient $\bar{x}_1, \dots, \bar{x}_n$ les moyennes arithmétiques sur les n sous-échantillons. La moyenne arithmétique sur l'échantillon E est la moyenne arithmétique des n moyennes arithmétiques : $\bar{x} = \frac{1}{n} \sum_{i=1}^n n_i \bar{x}_i$
- La moyenne des carrés des différences de valeurs x a une constante x_0 est minimale $x_0 = \bar{x}$

Chapitre 1. Les caractéristiques de tendance centrale

4. Généralisation de la moyenne



La moyenne arithmétique n'est qu'un cas particulier d'un concept plus général appelé moyenne d'ordre r , noté $\bar{x}_r$, qui se définit comme la racine $r^{\text{ième}}$ de la moyenne des valeurs x_i à la puissance r . Soit

$$\bar{x}_r = \left[\frac{1}{n} \sum_{i=1}^k n_i x_i^r \right]^{\frac{1}{r}}$$

En pratique, on ne calcule cependant que les moyennes d'ordre $r = -1$ (moyenne harmonique), $r \rightarrow 0$ (moyenne géométrique), $r = 1$ (moyenne arithmétique) et $r = 2$ (moyenne quadratique).

4. Généralisation de la moyenne/ la moyenne harmonique



Exemple : Un automobiliste a effectué un trajet total de 600 km. Il a parcouru 300 km à une vitesse moyenne de 100 km/h et 300 km à une vitesse moyenne de 60 km/h. On souhaite déterminer la vitesse moyenne avec laquelle l'automobiliste a traversé les 600 km.

Comme on cherche une vitesse moyenne, on est tenté d'utiliser la moyenne arithmétique. Dans ce cas, la vitesse moyenne est donnée par

$$V_m = \bar{x} = \frac{100+60}{2} = 80 \text{ km/h}$$

4. Généralisation de la moyenne/ la moyenne harmonique



Ainsi, les 600 km auraient été parcourus à la vitesse moyenne de 80 km/h ce qui est bien entendu faux. En effet, d'après cette valeur moyenne l'automobiliste aurait consacré

$$: \frac{600 \text{ km}}{80 \text{ km/h}} = 7,5 \text{ heures}$$

pour le trajet. Cependant, les vitesses moyennes sur les deux tranches donne une durée de

$$\frac{300 \text{ km}}{100 \text{ km/h}} + \frac{300 \text{ km}}{60 \text{ km/h}} = 8 \text{ heures.}$$

4. Généralisation de la moyenne/ la moyenne harmonique



Lorsqu'on calcule la vraie vitesse moyenne en utilisant sa formule de définition on trouve 75 km/h

$$V_m = \frac{\text{Distance}}{\text{Durée}} = \frac{600}{\frac{300}{100} + \frac{300}{60}} = \frac{2}{\frac{1}{100} + \frac{1}{60}}$$

C'est la formule correcte de la moyenne dans cet exemple. Comme on peut le constater elle se présente comme l'inverse de la moyenne arithmétique des inverses des valeurs. Elle est appelée moyenne harmonique.

4. Généralisation de la moyenne/ la moyenne harmonique



Définition : La moyenne harmonique, notée H , est la moyenne d'ordre - 1, elle se définit comme l'inverse de la moyenne arithmétique des inverses des valeurs. Soit

$$H = \left[\frac{1}{n} \sum_{i=1}^k n_i x_i^{-1} \right]^{-1} = \left[\frac{1}{n} \sum_{i=1}^k \frac{n_i}{x_i} \right]^{-1} = \frac{n}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

Bien entendu, dans le cas d'une série statistique ou de données groupées en classes, la formule précédente est valable avec néanmoins, $n_i = 1$ dans le premier cas et $x_i = c_i$ dans le second.

Pour garder le même principe de présentation d'une moyenne *i.e.* rapport entre la masse globale et l'effectif, la moyenne harmonique est souvent exprimée par son inverse. Soit :

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

4. Généralisation de la moyenne/ la moyenne harmonique



La moyenne harmonique s'utilise à chaque fois que la grandeur varie selon une raison inverse d'une autre valeur. Il s'agit essentiellement de variables qui se présentent sous forme de rapports ou de pourcentages. On peut penser à des vitesses, taux de change, prix par article, lits par patients, instituteurs par habitants etc.

.

4. Généralisation de la moyenne/ la moyenne géométrique

Exemple. On a placé 10 000 dirhams pendant trois années. L'obligation rapporte 5% la première année, 8% la deuxième année et 11% la dernière année. Calculons le taux d'intérêt moyen sur les 3 années.

Là encore, on est tenté d'utiliser la moyenne arithmétique. Cependant il est erroné d'assimiler un taux d'intérêt moyen à une telle moyenne. En effet, la moyenne arithmétique montre que l'obligation est placée à un taux d'intérêt annuel moyen de 8% puisque

$$\bar{x} = \frac{5+8+11}{3} = 8\%.$$

4. Généralisation de la moyenne/ la moyenne géométrique



Aussi, on s'attendra à la fin des trois années à récupérer 12 597,12 dirhams ($10\,000 \times 1,08^3$). Alors que la vraie somme que la banque acceptera de nous donner est de seulement 12 587,40 dirhams puisque

$$10\,000 \times 1,05 \times 1,08 \times 1,11 = 12587,4$$

Le taux d'intérêt moyen, τ_m , se calcule alors de la manière suivante

$$10\,000 \times g \times g \times g = 12587,4$$

où $g = 1 + \tau_m$ est le facteur d'intérêt

En égalisant les deux dernières équations il vient :

$$10\,000 \times g \times g \times g = 10\,000 \times 1,05 \times 1,08 \times 1,11$$

4. Généralisation de la moyenne / la moyenne géométrique



Ou encore

$$g = (1,05 \times 1,08 \times 1,11)^{\frac{1}{3}} = 1,07972$$

Soit un taux d'intérêt moyen de 7,972%.

Comme on peut le constater, le facteur d'intérêt moyen g est la racine troisième du produit de trois facteurs d'intérêt ; c'est la moyenne géométrique.

4. Généralisation de la moyenne / la moyenne géométrique

Définition. La moyenne géométrique, notée G , est la moyenne d'ordre $r \rightarrow 0$. Elle se définit comme la racine $n^{\text{ième}}$ du produit de n valeurs positives.

$$G = (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}},$$

Où

$$G = (x_{11}^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k})^{\frac{1}{n}} = \left(\prod_{i=1}^k x_i^{n_i} \right)^{\frac{1}{n}}$$

si certaines valeurs se répètent plusieurs fois.

Là encore, pour garder le même principe de construction de la moyenne arithmétique, la moyenne géométrique est généralement exprimée par son logarithme.

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

4. Généralisation de la moyenne/ la moyenne géométrique



Utilisations. La moyenne géométrique s'utilise pour le calcul de la moyenne de valeurs se présentant sous forme de taux de croissance tels que les taux d'intérêt et divers taux de croissance.

4. Généralisation de la moyenne/ la moyenne quadratique



Définition : La moyenne quadratique, notée Q , est la moyenne d'ordre $r=2$. Elle s'écrit

$$Q = \left[\frac{1}{n} \sum_{i=1}^k n_i x_i^2 \right]^{1/2} = \left[\sum_{i=1}^k f_i x_i^2 \right]^{1/2}$$

Utilisation : Elle s'utilise pour le calcul de la moyenne de valeurs qui présentent des signes différents comme des écarts.

4. Généralisation de la moyenne/ la moyenne quadratique



Exemple : 5 étudiants ont obtenu les notes suivantes :

7 ; 9 ; 10 ; 11 ; 13

On souhaite avoir une idée sur les différences de niveau entre les étudiants. Pour ce faire il faut calculer l'écart moyen à la moyenne.

La moyenne arithmétique est donnée par

$$\bar{x} = \frac{7 + 9 + 10 + 11 + 13}{5} = 10$$

L'écart moyen par rapport à la moyenne arithmétique est donné par

$$\bar{E} = \frac{(7 - 10) + (9 - 10) + (10 - 10) + (11 - 10) + (13 - 10)}{5}$$

Chapitre 1. Les caractéristiques de tendance centrale

4. Généralisation de la moyenne/ la moyenne quadratique

Cela signifie que les étudiants présentent le même niveau. Or il est clair que ce n'est pas le cas. L'utilisation de la moyenne arithmétique n'est pas appropriée.

La moyenne appropriée est la moyenne quadratique :

$$\bar{E} = \left[\frac{(7 - 10)^2 + (9 - 10)^2 + (10 - 10)^2 + (11 - 10)^2 + (13 - 10)^2}{5} \right]^{1/2} = 2$$

Ainsi, l'écart moyen entre les étudiants de 2 points.

Chapitre 2. Les caractéristiques de dispersion



1. LES CARACTÉRISTIQUES DE DISPERSION ABSOLUES
2. LES CARACTÉRISTIQUES DE DISPERSION RELATIVES

Chapitre 2

. Les caractéristiques de dispersion



1. Les caractéristiques de dispersion absolue

- L'étendue
- Les intervalles interquantiles
- L'écart absolu moyen
- La variance et l'écart type

2. Les caractéristiques de dispersion relative

- Le Minmax
- Les écarts interquantiles
- Le coefficient de variation
- Les moments

Chapitre 2. Les caractéristiques de dispersion



Certaines distributions, tout en présentant les mêmes caractéristiques de tendance centrales et de position, peuvent être très différentes les unes des autres. Considérons deux pays A et B . Les revenus moyens dans les pays A et B sont identiques. Mais, dans le pays A il y a une proportion importante d'individus qui gagnent bien plus que la moyenne et une autre proportion qui gagne bien moins que la moyenne. Dans le pays B , en revanche, les revenus sont concentrés autour de la moyenne. Les distributions des revenus dans les deux pays sont données par la figure suivante.

Chapitre 2. Les caractéristiques de dispersion



Dans le cas du pays B la moyenne arithmétique ne présente qu'un intérêt limité car elle n'est pas très représentative des valeurs de la distribution qui sont très dispersées autour d'elle. Le résumé est complété par d'autres caractéristiques qui donnent une idée sur l'éloignement et la dispersion des valeurs autour de la tendance centrale. Il s'agit des caractéristiques de dispersion. Comme pour les caractéristiques de tendance centrale, on peut imaginer plusieurs caractéristiques de dispersion : l'étendue, les écarts interquantiles, l'écart absolu moyen et la variance ou l'écart-type

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue



Ces caractéristiques s'expriment dans l'unité même du caractère pour lequel elles sont censées mesurer la dispersion. Elles ne sont pas très indiquées pour comparer la dispersion de plusieurs distributions qui présentent des unités différentes ou des dimensions très différentes.

:

:

.

.

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / L'étendue

L'étendue, notée e , représente la longueur de l'intervalle de variation de la série. C'est la différence entre la plus grande et la plus petite valeur.

$$e = x_{\max} - x_{\min}$$

L'étendue est très rarement utilisée seule. Certes elle est très facile à calculer mais il ne prend pas en considération toutes les observations et elle est très sensible aux valeurs extrêmes.

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / Les intervalles interquartiles

Si les quantiles sont des mesures de tendance centrale, les écarts entre quantiles de même type peuvent être utilisés pour apprécier la dispersion. Généralement, on définit les écarts interquartiles ($q_3 - q_1$ qui contient 50% des observations) et l'écart inter-déciles ($d_9 - d_1$ qui contient 80% des observations).

Bien entendu, lorsqu'on compare plusieurs distributions différentes on préférera utiliser des écarts relatifs. Les écarts interquartiles et interdéciles deviennent respectivement

$$(q_3 - q_1) / q_2 \quad \text{et} \quad (d_9 - d_1) / d_5.$$

Le plus grand reproche qui est fait à ces caractéristiques et qu'elles ne se prêtent pas facilement aux manipulations algébriques.

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / L'écart absolu moyen

L'écart absolu moyen, noté $\bar{e}_m$, est la moyenne arithmétique de la valeur absolue des écarts par rapport à la moyenne arithmétique

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

L'écart absolu médian se définit de la même manière

$$\bar{e}_m = \frac{1}{n} \sum_{i=1}^k n_i |x_i - M\acute{e}| = \bar{e}_m = \sum_{i=1}^k f_i |x_i - M\acute{e}|$$

Comme pour la caractéristique précédente, le seul inconvénient de l'écart absolu moyen ou médian est la présence de la fonction valeur absolue qui la rend très difficile à traiter algébriquement.

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / La variance et l'écart type

Définition : La variance, notée $V(x)$ ou σ^2 , se définit comme la moyenne arithmétique des carrés des écarts des valeurs à la moyenne.

$$V(x) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

Cette formule est dite formule de définition, elle conduit à quelques approximations dans les calculs. En pratique, on utilise une autre formule de la variance obtenue par décomposition de la formule précédente et appelée formule développée. Il s'agit de

$$V(x) = \frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2 = \sum_{i=1}^k f_i x_i^2 - \bar{x}^2$$

La variance peut ainsi être définie comme *la moyenne des carrés moins le carré de la moyenne.*

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / La variance et l'écart

type

Propriétés de la variance :

(i) La variance, à la différence des autres caractéristiques de dispersion se prête aisément aux calculs.

(ii) $V(x) \geq 0$ avec $V(x) = 0$ pour x une constante.

(iii) $V(ax + b) = a^2 V(x)$.

(iv) Si V_1 et V_2 sont les variances de deux distributions de tailles respectives n_1 et n_2 alors la variance totale $V(x)$ est donnée par

$$V(x) = \frac{n_1 V(x)_1 + n_2 V(x)_2}{n_1 + n_2}$$

si les deux distributions présentent la même moyenne arithmétique.

(v) La variance est très sensible aux valeurs aberrantes

(v) La variance s'exprime dans le double de l'unité de la valeur de la variable x elle n'a pas de pouvoir descriptif.

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / La variance et l'écart



La limite présentée dans la dernière propriété nous conduit à introduire une nouvelle caractéristique appelée écart-type, notée σ , définie comme la racine carrée de la variance. $\sigma = \sqrt{V(x)}$. Bien entendu, l'écart-type s'exprime dans la même unité que la variable x .

La variance, et donc l'écart-type, sous-estiment l'écart des valeurs par rapport à la moyenne notamment pour n petit.

Dans ce cas on peut utiliser une variance corrigée appelée quasi-variance. Elle est donnée par

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

Chapitre 2. Les caractéristiques de dispersion

1. les caractéristiques de dispersion absolue / La variance et l'écart



L'écart type est la caractéristique de dispersion la plus utilisée. Son seul inconvénient est qu'elle est une caractéristique absolue. Elle ne peut être utilisée pour comparer la dispersion des distributions très différentes (qui ont des moyennes différentes).

- :
-

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion relative

Toutes les caractéristiques précédentes sont utilisées pour décrire une seule distribution. Pour comparer plusieurs distributions entre elles d'un point de vue de la dispersion, on doit utiliser des caractéristiques relatives pour lesquelles les échelles et les unités ne présentent pas d'importance. Généralement elles se présentent sous la forme d'une caractéristique de dispersion absolue et une caractéristique de tendance centrale.

.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion relative / Le minmax

Le Minmax se définit comme le rapport entre la valeur la plus élevée est la valeur la plus faible de la distribution

$$\text{Min max} = \frac{x_{\max}}{x_{\min}}$$

:

Comme pour l'étendue cette caractéristique n'est pas très pertinente en présence de valeurs extrêmes.

.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion relative / Le écarts interquantiles

On retiendra deux écarts. Les écarts interquartiles et interdéciles. Soient

$$IQ = (q_3 - q_1)/q_2$$

$$IK = (d_9 - d_1)/d_5$$

L'inconvénient de ce type de caractéristiques est qu'elles ne présentent pas de limites supérieures.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Le coefficient de variation

Ce coefficient de variation est la caractéristique de dispersion relative la plus utilisée. Elle est donnée par

$$CV = \frac{\sigma}{\bar{x}} \quad \text{ou} \quad CV\% = \frac{\sigma}{\bar{x}} \times 100$$

Sa seule limite est qu'elle est très sensible à la moyenne arithmétique et ne peut pas être utilisée lorsque celle-ci est proche de zéro.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Le coefficient de variation

Exemple. Une entreprise emploie des hommes et des femmes. Le salaire moyen des hommes est de 6000 MAD et celui des femmes est de 3000 MAD. L'écart-type des salaires pour les hommes est de 500 MAD et pour les femmes de 300 MAD. Malgré ça, il serait faux d'affirmer que les salaires des hommes présente une plus forte dispersion que celui des femmes. En effet, comme les deux distributions sont assez différentes (les salaires moyens sont très différents) il serait plus approprié de comparer les dispersions en utilisant une caractéristique de dispersion relative *i.e* le coefficient de variation. Celui des hommes est de 8,3% et ce lui des femmes est de 10%. Aussi, la dispersion des salaires plus importante chez les femmes que chez les hommes.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion relatives / Les moments

Les moments sont une généralisation des concepts de moyenne et de variance.

Définition : On appelle moment général d'ordre r par rapport à une origine a la quantité donnée par :

$$M_r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - a)^r$$

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Les moments

Généralement on définit deux types de moments :

Les moments simples : On appelle moment simple d'ordre r , noté m_r , le moment d'origine 0. Les moments simples sont une généralisation de la moyenne

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

Ainsi, $m_1 = \bar{x}$ et $m_2 = \frac{1}{n} \sum_{i=1}^k n_i x_i^2$

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Les moments

Les moments centrés : On appelle moment centré d'ordre r , noté μ_r , le moment d'origine $\bar{x}$. Les moments centrés sont une généralisation de la variance.

$$\mu_r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^r = \sum_{i=1}^k f_i (x_i - \bar{x})^r$$

Pour $\mu_2 = \frac{1}{n} \sum n_i (x_i - \bar{x})^2 = \sigma^2$

Comme $\sigma^2 = \frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2$ il vient que $\mu_2 = m_2 - m_1^2$

- On peut mettre ainsi, tous les moments centrés en fonctions des moments simples d'ordre inférieur.
- Les moments centrés d'ordre 3 et 4 sont très utilisés. Ils permettent de caractériser la forme d'une distribution. Ils font partie des caractéristiques de forme.

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Les moments

Propriétés :

$$\mu_2 = m_2 - m_1^2$$

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3$$

:

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4$$

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Les moments

Exemple. Calculez les moments centrés d'ordre 2, 3 et 4 de la série statistique suivante (salaires des 200 individus d'une entreprise)

x_i	n_i
[3000-4000[	42
[4000-5000[	53
[5000-6000[	85
[6000-7000[	12
[7000-8000[	8
Total	200

Chapitre 2. Les caractéristiques de dispersion

2. les caractéristiques de dispersion absolue / Les moments

x_i	n_i	c_i	$n_i \times c_i$	$(c_i - \bar{x})^2$	$n_i \times (c_i - \bar{x})^2$	$n_i c_i^2$	$n_i c_i^3$	$n_i \times (c_i - \bar{x})^3$	C_i^2	C_i^3	$n_i c_i^4$
[3-4[	42	3,5	147	2,11	88,91	514,5	1800,75	-129,37	12,25	42,88	6302,62
[4-5[	53	4,5	238,5	0,20	10,97	1073,25	4829,62	-4,99	20,25	91,13	21733,31
[5-6[	85	5,5	467,5	0,29	25,24	2571,25	14141,87	13,75	30,25	166,38	77780,31
[6-7[	12	6,5	78	2,38	28,64	507	3295,5	44,25	42,25	274,63	21420,75
[7-8[	8	7,5	60	6,47	51,81	450	3375	131,87	56,25	421,88	25312,5
Total	200	27,5	991	11,485125	205,595	5 116,00	27 442,75	55,52355	161,25	996,88	152 549,50

:

.

.

Chapitre 3. Les caractéristiques de forme



1. L'ASYMÉTRIE
2. L'APLATISSEMENT

Chapitre 3. Les caractéristiques de forme



1. L'asymétrie

- Coefficient de Pearson
- Coefficient de Yule
- Coefficient de Fischer

2. L'aplatissement

- Coefficient de Pearson
- Coefficient de Fischer

Chapitre 3. Les caractéristiques de forme



- Deux distributions qui présentent les mêmes caractéristiques de tendance centrale et de dispersion peuvent avoir des formes très différentes. Ces caractéristiques servent essentiellement à comparer plusieurs distributions.
- La forme d'une distribution peut être appréhendé par deux caractéristiques : l'asymétrie (*Skewness*) et l'aplatissement (*Kurtosis*).

Chapitre 3. Les caractéristiques de forme

1. L'asymétrie (Skewness)



- L'asymétrie d'une distribution est mesurée par rapport à une distribution parfaitement symétrique. Cette dernière est tel que les valeurs sont distribuées de la même manière de part et d'autre des trois caractéristiques de tendance centrale $\bar{x} = M_o = M_e$.
- Plusieurs moyens sont utilisés pour mesurer l'asymétrie d'une distribution.

Chapitre 3. Les caractéristiques de forme

1. L'asymétrie / Coefficient de Pearson



K. Pearson a défini une relation empirique entre les trois caractéristiques de tendance centrale valable pour les distributions unimodales et peu asymétriques. Elle s'écrit :

$$(\bar{x} - M_o) = 3(\bar{x} - M_e)$$

Ainsi, la distance entre ces caractéristiques peut-être utilisée pour mesurer l'asymétrie. K. Pearson a défini le coefficient suivant :

$$CP = \frac{\bar{x} - M_o}{\sigma} = \frac{3(\bar{x} - M_e)}{\sigma}$$

Si $CP = 0$, la distribution est dite symétrique

Si $CP > 0$, la distribution présente une oblicité à gauche

Si $CP < 0$, la distribution présente une oblicité à droite

Chapitre 3. Les caractéristiques de forme

1. L'asymétrie / Coefficient de Yule



Yule à utilisé les écarts entre quantiles pour construire un coefficient d'asymétrie. Dans le cas des intervalles interquartiles, le coefficient de Yule peut s'écrire comme suit :

$$CY = \frac{[(q_3 - q_2) - (q_2 - q_1)]}{(q_3 - q_1)}$$

- Si $CY = 0$, la distribution est symétrique
- Si $CY > 0$, la distribution présente une oblicité à gauche
- Si $CY < 0$, la distribution présente une oblicité à droite

Bien entendu, on peut utiliser les écarts entre déciles pour construire un coefficient équivalent.

$$CY = \frac{[(d_9 - d_5) - (d_5 - d_1)]}{(d_9 - d_1)}$$

Chapitre 3. Les caractéristiques de forme

1. L'asymétrie / Coefficient de Fischer



Fisher s'est basé sur la propriété des moments centrés. En effet, les moments centrés d'ordre impair sont nulles pour les distributions symétriques. Le coefficient de Fisher s'écrit :

$$CF = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

Là encore, on peut utiliser des moments centrés impairs d'ordre supérieur.

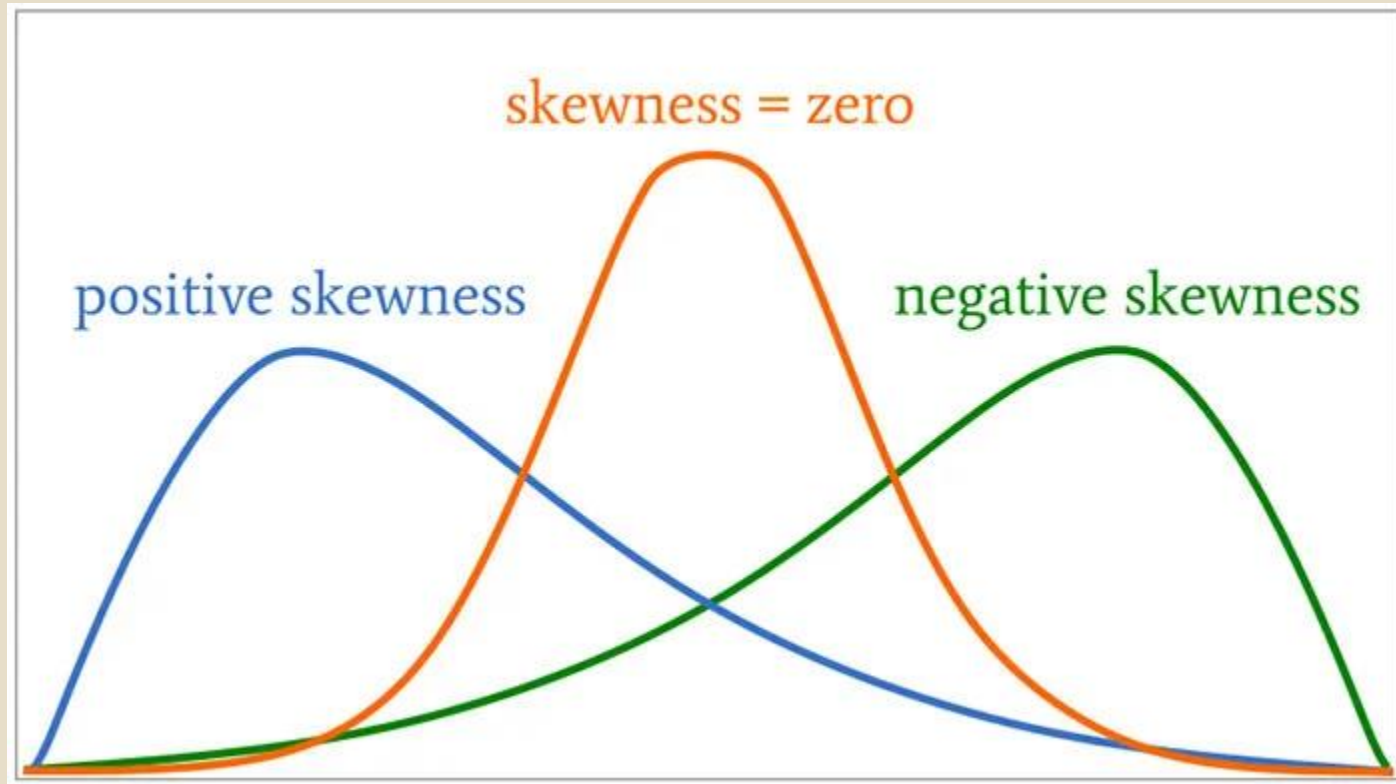
Si $CF = 0$, la distribution est symétrique

Si $CF > 0$, la distribution présente une oblicité à gauche

Si $CF < 0$, la distribution présente une oblicité à droite

Chapitre 3. Les caractéristiques de forme

1. L'asymétrie / Coefficient de Fischer



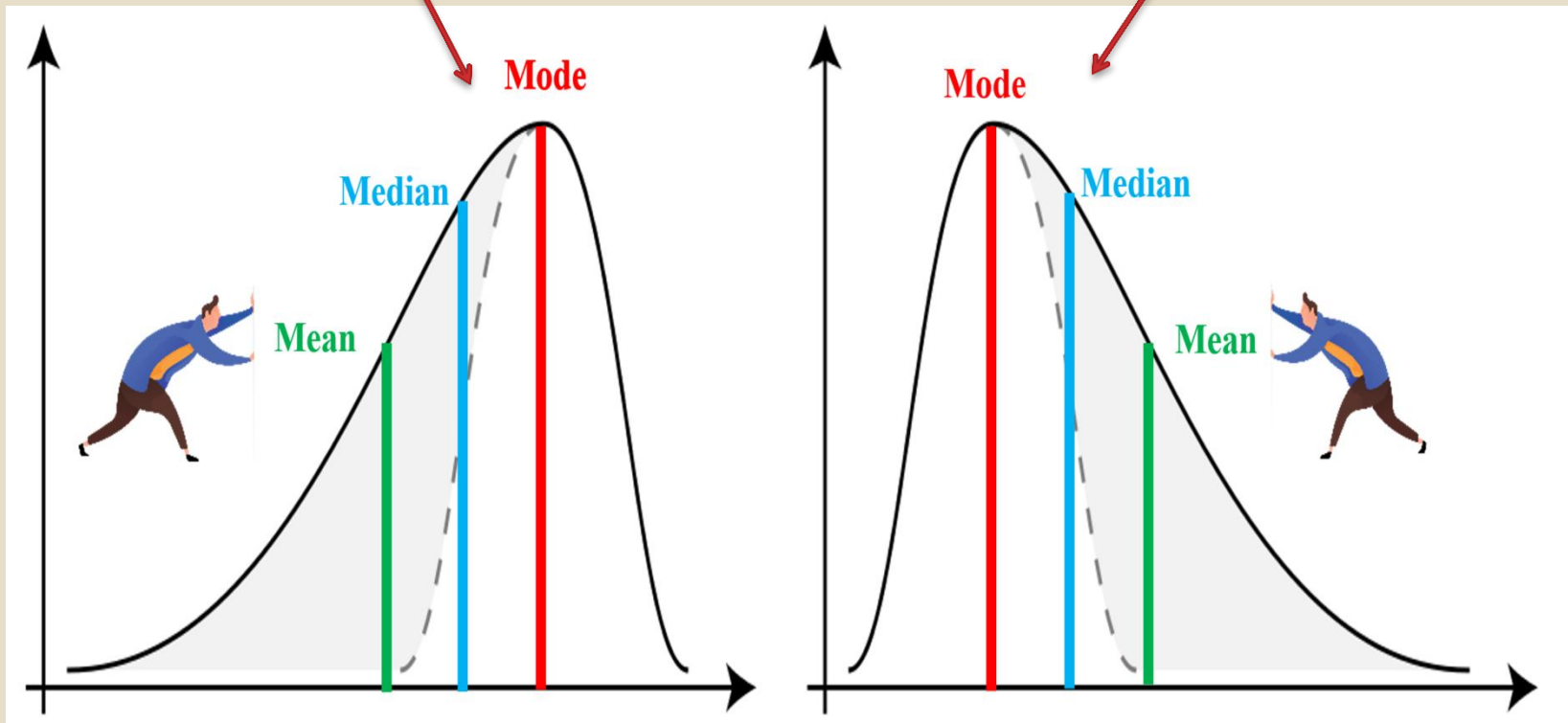
Chapitre 3. Les caractéristiques de forme

1. L'asymétrie / Coefficient de Fischer



Oblicité à droite

Oblicité à gauche



Asymétrie à gauche

Asymétrie à droite

Chapitre 3. Les caractéristiques de forme

2. L'aplatissement (Kurtosis)



L'aplatissement est mesuré par rapport à une distribution théorique dite (Gauss-Laplace ou normale). Plus précisément, l'aplatissement mesure la forme aiguë ou plate (ou de manière équivalente l'épaisseur des queues de la distribution) d'une distribution par rapport à la distribution normal.

Chapitre 3. Les caractéristiques de forme

2. L'aplatissement / Coefficient de Pearson



Pearson propose le coefficient suivant basé sur le moment centré d'ordre 4

$$CP = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$

:

Si $CP = 3$, normale ou *mésokurtique*

Si $CP > 3$, la distribution est *leptokurtique*

Si $CP < 3$, la distribution est *platykurtique*

Chapitre 3. Les caractéristiques de forme

2. L'aplatissement / Coefficient de Fischer



Fisher propose un coefficient équivalent :

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

:

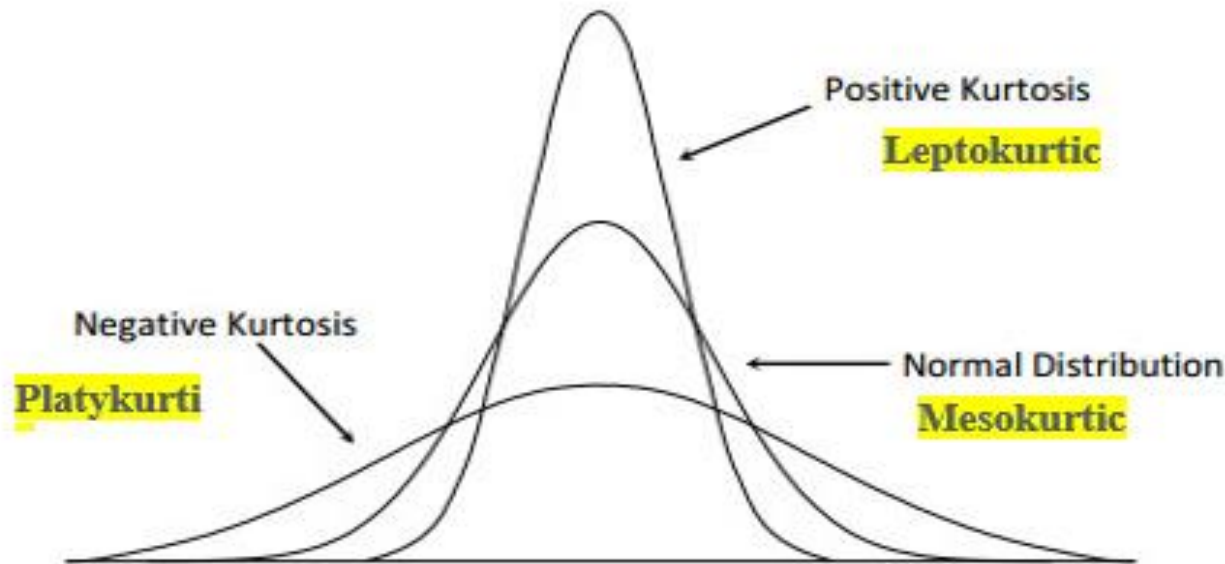
Si $CF = 0$, la distribution *mésokurtique*

Si $CF > 0$, la distribution est *leptokurtique*

Si $CF < 0$, la distribution est *platykurtique*

Chapitre 3. Les caractéristiques de forme

2. L'aplatissement / Coefficient de Fischer



Si $CF = 0$, la distribution *mésokurtique*

Si $CF > 0$, la distribution est *leptokurtique*

Si $CF < 0$, la distribution est *platykurtique*

Chapitre 4. Les caractéristiques de concentration



1. L'INDICE DE CONCENTRATION
2. LA COURBE DE LORENZE ET INDICE DE GINI

Chapitre 4. Les caractéristiques de concentration



1. L'indice de concentration

- La médiale
- L'indice de concentration

2. La courbe de Lorenz et indice de Gini

- La courbe de Lorenz
- L'indice de GINI

Chapitre 4. Les caractéristiques de concentration



La notion de concentration est très proche de la notion de dispersion. Elle permet d'examiner comment une quantité appelée masse est distribuée entre les individus d'une population. Les caractéristiques de concentration n'ont de sens que pour les variables dont la somme admet une signification concrètes : salaires, revenus, parts de marchés, etc.

Généralement deux outils sont utilisés pour apprécier la concentration :

- L'approche algébrique qui débouche sur le calcul de l'indice de concentration
- L'approche graphique à travers la courbe de Lorenz et qui débouche sur le calcul de l'indice de GINI

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration



La logique de la construction de l'indice de concentration consiste à comparer deux caractéristiques de tendance centrales spécifiques : la médiane et la médiale.

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / La médiale



La médiale :

- La médiale est une notion proche de la médiane.
- Si la médiane est la valeur de la variable qui permet de scinder l'effectif en deux parties égales, la médiale est la valeur de la variable qui scinde la masse (somme des valeurs) en deux parties égales.
- En termes techniques : si la médiane est la caractéristique de tendance centrale de la distribution (x_i, n_i) , la médiale est la même caractéristique de tendance centrale de la distribution $(x_i, n_i x_i)$.
- Elle se calcule de la même manière que la médiane.

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / La médiale

- Dans le cas d'une distribution en données groupées, elle se calcule en deux étapes.
 - *Etape 1.* On détermine la classe médiale.
 - *Etape 2.* On calcule par interpolation linéaire une approximation de la médiale.

$$Ml = b_{i-1} + a_i \left(\frac{1/2 - F_{i-1}^m}{f_i^m} \right)$$

Dans cette équation, $[b_{i-1}, b_i[$ est la classe médiale, f_i^m est la fréquence relative de la masse et F_{i-1}^m la fréquence cumulée de la masse.

Interprétation : Si le caractère étudié est le salaire par exemple alors la médiale s'interpréterait comme le salaire tel que les individus qui gagnent un salaire inférieur au salaire médial se partagent la moitié de la masse salariale

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / La médiale



Exemple. Considérons la distribution des salaires (en milliers dirhams) suivante :

x_i	n_i
[1, 2[	20
[2, 3[	40
[3, 4[	60
[4, 5[	20
[5, 6[	10
Total	150

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / La médiale

Exemple. On construit la distribution $(x_i, n_i x_i)$

x_i	n_i	$n_i x_i$	f_i^m	$F_i^{+,m}$
[1, 2[	20	30	0,06	0,06
[2, 3[	40	100	0,21	0,27
[3, 4[	60	210	0,43	0,70
[4, 5[	20	90	0,19	0,89
[5, 6[	10	55	0,11	1
Total	150	485	1	----

Etape 1. La classe médiale est la classe [3, 4[

Etape 2. Calcul de la médiale par interpolation linéaire

$$Ml = b_{i-1} + a_i \left(\frac{1/2 - F_{i-1}^m}{f_i^m} \right) = 3 + 1 \left(\frac{0,5 - 0,27}{0,43} \right) = 3,5358 \equiv 3535,8\text{DH}$$

Ainsi, les salariés qui gagnent un salaire inférieur à 3535,8DH se partagent la moitié de la masse salariale (c.a.d. 485/2).

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / L'indice de concentration

- Le principe de l'indice de concentration est de considérer une distribution parfaitement égalitaire (absence de concentration)
- La moitié des salariés se partagent exactement la moitié la masse salariale.
- Lorsque la médiane est juste égale à la médiale. Pour obtenir un indice c'est-à-dire une quantité sans unité ni dimension et compris entre 0 et 1 il suffit de diviser cet écart sur l'étendue.

Ainsi, l'indice de concentration est donné par : $IC = \frac{Ml - Mé}{e}$

- Si $IC = 0$ cela signifie qu'il n'y a pas de concentration
- Si $IC = 1$ cela signifie que la concentration est maximale.

Chapitre 4. Les caractéristiques de concentration

1. L'indice de concentration / L'indice de concentration

Reprenons notre exemple :

- Calculons d'abord la médiane. La classe médiane est la classe [3, 4[. La méthode de l'interpolation linéaire donne une approximation de la médiane comme suit :

$$Mé = b_{i-1} + a_i \left(\frac{1/2 - F_{i-1}}{f_i} \right) = 3 + 1 \left(\frac{0,5 - 0,4}{0,4} \right) = 3,25 \equiv 3250\text{DH}$$

L'étendue est donnée par 5000 DH (6000 – 1000).

Ainsi, l'indice de concentration est donné par :

$$IC = \frac{Ml - Mé}{e} = \frac{3535,8 - 3250}{5000} = 0,057 \equiv 5,7\%$$

Il est donc possible de considérer que la distribution des salaires présente une faible concentration.

Chapitre 4. Les caractéristiques de concentration

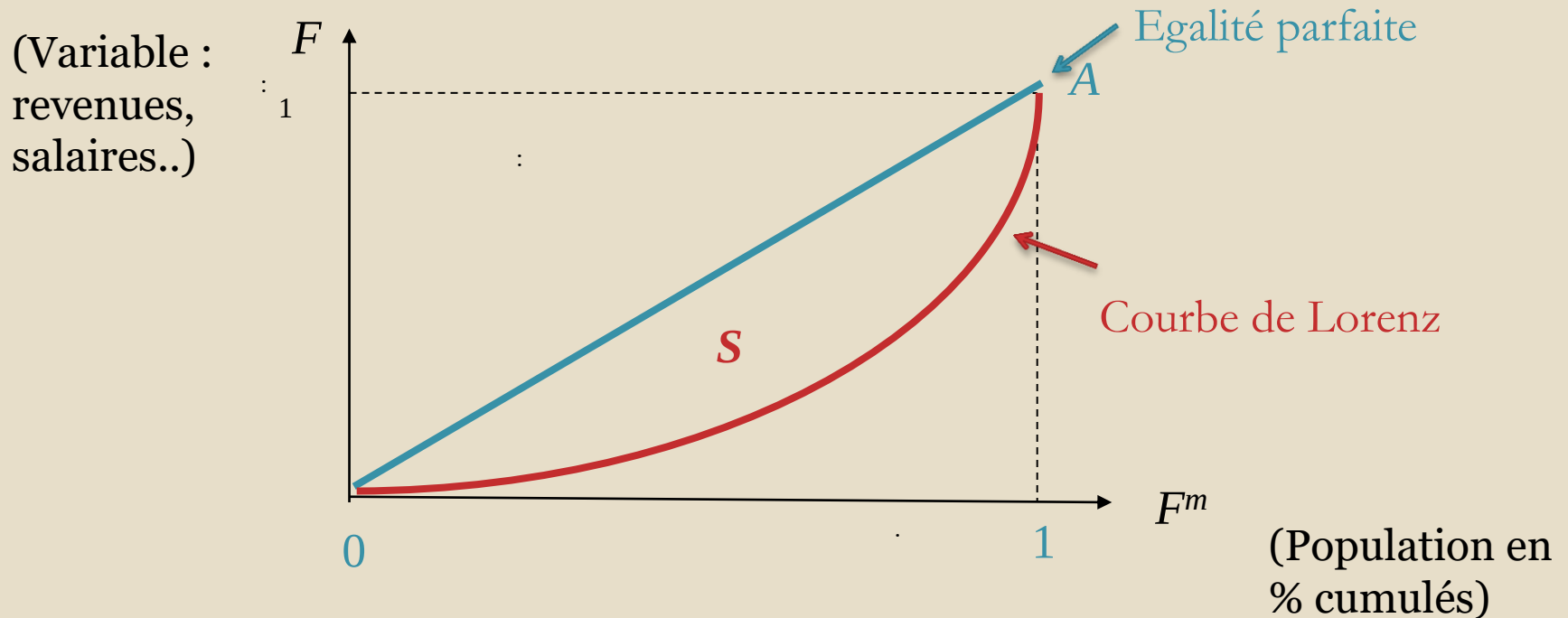
2. la courbe de Lorenz et indice de GINI/Courbe de Lorenz

- Dans l'exemple précédent on examinait combien se partagent la moitié des salariés. C'est-à-dire est ce que 50% des salariés se partagent 50% de la masse salariale.
- Dans le cas de l'approche graphique on se pose la question si la part des salariés qui gagnent moins d'un certain salaires se partagent la même part de la masse salariale. C'est-à-dire si 10%, 20%, etc. des salariés se partagent 10%, 20%, etc. de la masse salariale.
- Ainsi, la courbe de *Lorenz* ou courbe de concentration est donnée par l'ensemble des combinaisons F_i et F_i^m .

Chapitre 4. Les caractéristiques de concentration

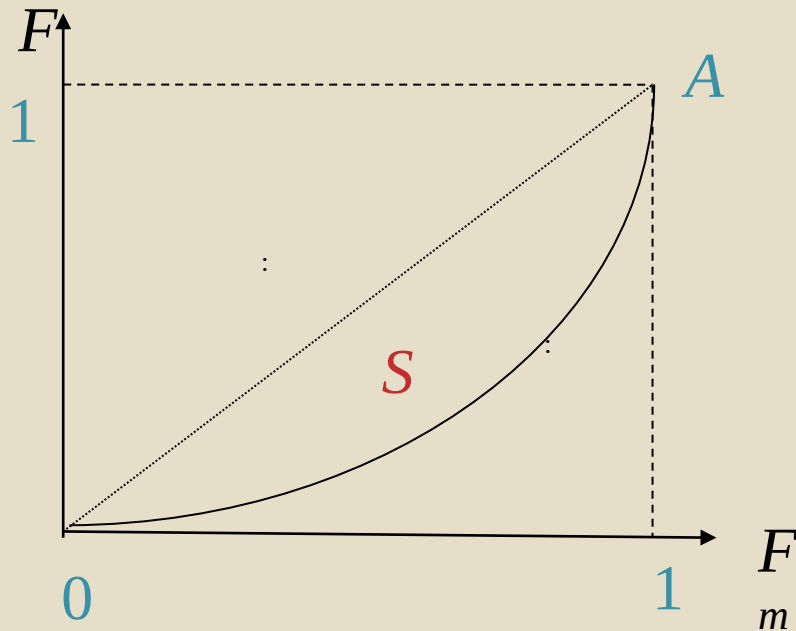
2. la courbe de Lorenz et indice de GINI/Courbe de Lorenz

Si on utilise les fonctions fréquences cumulées croissantes, la forme générale de la courbe de concentration est donnée par le graphique suivant :



Chapitre 4. Les caractéristiques de concentration

2. la courbe de Lorenz et indice de GINI/Courbe de Lorenz



- La première bissectrice représente la courbe de Lorenz lorsque la concentration est nulle.
- Plus la courbe de Lorenz s'éloigne de la première bissectrice plus la concentration est forte.
- La surface entre la première bissectrice et la courbe de Lorenz, notée S , est appelé surface de concentration. Elle mesure le degré de concentration.
- Il est possible de calculer une approximation de cette surface, il s'agit de l'indice de GINI

Chapitre 4. Les caractéristiques de concentration

2. la courbe de Lorenz et indice de GINI/L'indice de GINI

- L'indice de GINI est un moyen plus objectif de lire la courbe de concentration.
- Pour obtenir une quantité sans dimension ni unité et de surcroît compris entre 0 et 1, il est défini comme le rapport entre la surface de concentration S et la surface du triangle $0A1$ qui est égale à $1/2$.

Soit :

$$IG = \frac{S}{1/2} = 2S$$

- Une méthode très utilisée pour calculer une approximation de l'indice de GINI consiste à calculer d'abord la surface de la surface complémentaire SC en utilisant la méthode des trapèzes et déduire ensuite S et IG . IL vient :

$$IG = 2S = 2(1/2 - SC)$$

Chapitre 4. Les caractéristiques de concentration

2. la courbe de Lorenz et indice de GINI/L'indice de GINI

$$IG = 2S = 2(1/2 - SC)$$

L'idée de cette méthode est d'approcher la surface SC par un ensemble de trapèzes. :

$$SC = \sum_{i=1}^k SC_i$$

avec $SC_i = \frac{(F_i + F_{i-1})(f_i - f_{i-1})}{2}$ la surface du trapèze relatif à la classe i .

Chapitre 4. Les caractéristiques de concentration

2. la courbe de Lorenz et indice de GINI/L'indice de GINI

Exemple. Reprenons notre exemple relatif à la distribution des salaires. La courbe de Lorenz est construite à partir des données suivantes :

F_i^m	F_i
0,06	0,13
0,27	0,40
0,70	0,80
0,89	0,93
1,0	1,0

Calculons d'abord les surfaces trapèzes. Il y'en a 5 (le nombre de modalités) :

$$SC_1 = [(0 + 0,06) \times 0,13]/2 = 0,039$$

$$SC_2 = [(0,06 + 0,27) \times 0,27]/2 = 0,0445$$

$$SC_3 = [(0,27 + 0,7) \times 0,4]/2 = 0,194$$

$$SC_4 = [(0,7 + 0,89) \times 0,13]/2 = 0,103$$

$$SC_5 = [(0,89 + 1) \times 0,07]/2 = 0,066$$

Chapitre 4. Les caractéristiques de concentration

2. la courbe de Lorenz et indice de GINI/L'indice de GINI

La surface complémentaire est donnée par :

$$SC = 0,039 + \dots + 0,066 = 0,4114$$

:

On déduit la surface de concentration :

$$S = 0,5 - 0,4114 = 0,0886$$

Et l'indice de *GINI* :

$$IG = 2 \times 0,0886 = 0,1772 \equiv 17,72$$

On notera que l'indice de GINI est bien entendu par construction toujours plus élevé que l'indice de concentration