ISCID-CO - PRÉPA 1ère année
STATISTIQUES ET PROBABILITÉS

Université du Littoral - Côte d'Opale

Laurent SMOCH

Janvier 2013

Laboratoire de Mathématiques Pures et Appliquées Joseph Liouville
Université du Littoral, zone universitaire de la Mi-Voix, bâtiment H. Poincaré
50, rue F. Buisson, BP 699, F-62228 Calais cedex

Table des matières

Chapitre 1

Séries statistiques à une variable

1.1 Introduction

À l'origine (sans doute en Chine, plus de 2000 ans avant Jésus-Christ et en Égypte, vers 1700 avant J.-C.), la ***statistique*** fournissait des renseignements intéressant l'État concernant la population (le nombre d'habitants d'un pays et leur répartition par sexe, par âge, par catégorie socio-professionnelle,...) et l'économie (l'évaluation des ressources de l'État, des stocks,...). Il faut préciser que le mot statistique, traduction du mot allemand "statistik" apparu au milieu du $XVIII^e$ siècle, provient du mot latin "status" qui signifie état.
Le premier bureau de statistique a été créé en France en 1800 par Napoléon. Cet organisme a pris en 1946 le nom d'Institut National de la Statistique et des Études Économiques (INSEE).

Les méthodes statistiques sont aujourd'hui employées principalement
- en médecine pour l'évaluation de l'efficacité d'un médicament, de l'état sanitaire d'une population,
- en agronomie pour la recherche d'engrais spécifiques ainsi que pour la sélection des variétés,
- en sociologie pour des enquêtes et sondages d'opinion,
- dans l'industrie pour l'organisation scientifique du travail, le contrôle de la qualité, la gestion des stocks

et dans bien d'autres domaines.

1.2 Méthodes de représentation

1.2.1 Vocabulaire

La statistique a traditionnellement un vocabulaire spécifique. Récapitulons ci-après les définitions des termes courants les plus utilisés.

Définition 1.2.1 *La **population** est l'ensemble que l'on observe et dont chaque élément est appelé **individu** ou unité statistique.*

Définition 1.2.2 *Un **échantillon** (ou lot) est une partie (ou sous-ensemble) de la population considérée.*

On étudie un échantillon de la population notamment lorsque celle-ci est impossible à étudier dans son ensemble ; c'est le cas pour les sondages d'opinion ou pour des mesures rendant inutilisables les objets étudiés, par exemple la durée de vie de piles électriques d'un certain type.

Définition 1.2.3 *Le **caractère** étudié est la propriété observée dans la population ou l'échantillon considéré.*

On peut citer par exemple la région de résidence de chaque français observé lors d'un recensement, ou le nombre d'enfants par famille observé à cette même occasion, ou encore la taille des élèves d'un lycée.

Dans ces deux derniers exemples, le caractère est dit **quantitatif** car il est mesurable : nombre d'enfants,

taille. Ça n'est pas le cas du premier exemple où le caractère est dit **qualitatif** : région.

Dans le deuxième exemple, le caractère quantitatif est **discret** car il ne peut prendre que des valeurs isolées (ici entières) alors que dans le troisième, le caractère quantitatif est **continu** car il peut prendre, au moins théoriquement, n'importe quelle valeur d'un intervalle de nombres réels.

1.2.2 Les tableaux

Dans chaque exemple, les résultats obtenus se présentent, au départ sous forme d'une liste éventuellement longue et sans autre classement que l'ordre d'arrivée des informations. Aussi, pour faciliter leur lecture, est-on amené à les présenter de manière plus synthétique sous forme de tableau ou de graphique.

Exemple 1.2.1 *(exemple de référence)* Un concessionnaire d'automobiles neuves a enregistré au cours de ses 40 premières semaines d'opération le nombre X d'automobiles qu'il a vendu hebdomadairement. Il a obtenu les résultats suivants :

5,7,2,6,3,4,8,5,4,3,9,6,5,7,6,8,3,4,4,0,8,6,7,1,5,5,4,6,6,10,9,8,1,5,5,6,7,8,5,5

Présenter de manière synthétique les résultats précédents à l'aide du tableau ci-dessous :

i	1	2	3	4	5	6	7	8	9	10	11
x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	1	3	5	9	7	4	5	2	1

Définition 1.2.4 *Une **classe** (ou modalité) est un sous-ensemble de la population correspondant à une même valeur ou à des valeurs voisines prises par le caractère.*

Ces classes peuvent donc être des valeurs ponctuelles (nombre d'enfants par famille, "2 enfants" est une classe) ou des intervalles (salaires en euros des employés d'une entreprise, $[700; 800]$ est une classe).

Définition 1.2.5 *L'**effectif** d'une classe est son nombre d'éléments.*

Ainsi, une série statistique à une variable peut être définie par un tableau de la forme :

i	1	2	...	...	p
Valeurs prises par le caractère (ou classes) x_i	x_1	x_2	...	...	x_p
Effectifs correspondants n_i	n_1	n_2	...	...	n_p

p est le nombre de classes (ou modalités) et n_i est l'effectif de la i-ème classe.

L'effectif total de l'échantillon n est tel que

$$n = n_1 + n_2 + \ldots + n_p \text{ avec } n_i \leq n, \ \forall\ 1 \leq i \leq p.$$

On peut noter également

$$\boxed{n = \sum_{i=1}^{p} n_i}$$

cette formule se lisant littéralement "somme des n_i pour i allant de 1 à p".

On considère l'exemple 1.2.1 :

- quelle est la taille (ou effectif total) de l'échantillon ?
  40
- À quoi correspond le caractère x_i ?
  Le nombre de voitures vendues la semaine i
- Le caractère x_i est-il quantitatif ou qualitatif, discret ou continu ?
  Le caractère est quantitatif (quantité de voitures) et discret (les classes sont des valeurs ponctuelles)
- Combien y a-t-il de modalités ?
  Il y a 11 modalités (x_i peut prendre les 11 valeurs $0, 1, 2, \ldots, 10$).

Définition 1.2.6 *La **fréquence** d'une classe est la proportion d'individus de la population (ou de l'échantillon) appartenant à cette classe.*

Ainsi la i-ème classe a pour fréquence

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

Considérons maintenant la somme des fréquences de toutes les classes :

$$f_1 + f_2 + \ldots + f_p = \sum_{i=1}^{p} f_i = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_p}{n} = \frac{1}{n} \sum_{i=1}^{p} n_i.$$

On sait que $\sum_{i=1}^{p} n_i = n$ donc $\sum_{i=1}^{p} f_i = \frac{n}{n} = 1$.

Propriété 1.2.1 *Les fréquences de classes d'une série statistique vérifient les propriétés suivantes :*

$$\boxed{\sum_{i=1}^{p} f_i = 1}$$

$$\boxed{0 \leq f_i \leq 1, \ \forall\ 1 \leq i \leq p.}$$

On considère l'exemple 1.2.1 : calculer les fréquences de chacune des modalités de la série statistique.

i	1	2	3	4	5	6	7	8	9	10	11	Total
x_i	0	1	2	3	4	5	6	7	8	9	10	–
n_i	1	2	1	3	5	9	7	4	5	2	1	40
f_i	0,025	0,05	0,025	0,075	0,125	0,225	0,175	0,1	0,125	0,05	0,025	1

1.2.3 Les graphiques

(a) Caractère qualitatif

Exemple 1.2.2 On s'intéresse aux étudiants d'une filière de Licence de l'ULCO et plus particulièrement à la région dont ils proviennent (qu'on symbolise respectivement par Lille, Calais Dunkerque). On obtient le tableau suivant :

Classe	Lille	Calais	Dunkerque
Pourcentage	$20,2\%$	$35,4\%$	$44,4\%$

La propriété étudiée dans la population des étudiants est la ville d'origine. C'est un caractère qualitatif qui prend trois valeurs ou modalités permettant ainsi de définir trois classes avec leur fréquence :

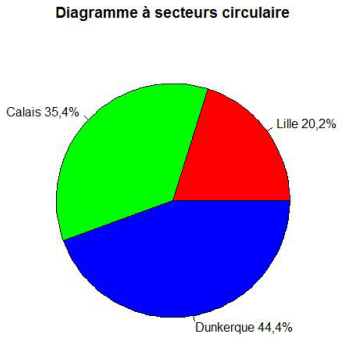
i	Région	Pourcentage	Fréquence f_i
1	Lille	20,2%	0,202
2	Calais	35,4%	0,354
3	Dunkerque	44,4%	0,444

Voici trois graphiques possibles pour cette série statistique.

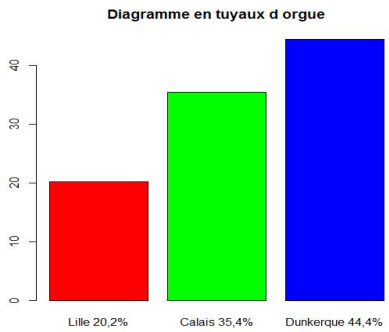
– **Diagramme à secteurs circulaire**
Chaque classe correspond à un secteur dont l'angle est proportionnel à l'effectif donc à la fréquence de la classe. On a les angles suivants :

i	Région	Pourcentage	Angle α_i
1	Lille	20,2%	72,72
2	Calais	35,4%	127,44
3	Dunkerque	44,4%	159,84

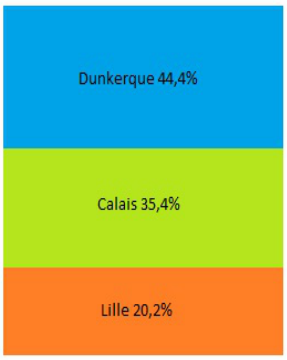
ce qui donne le diagramme



– **Diagramme en tuyaux d'orgue**
Dans ce cas et dans le suivant (diagramme en bandes), chaque classe est représentée par un rectangle de même largeur et de longueur proportionnelle à l'effectif, donc à la fréquence de la classe.



– **Diagramme en bandes**

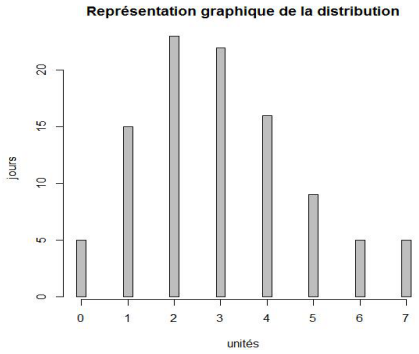


(b) Caractère quantitatif discret

Exemple 1.2.3 Le responsable des ventes d'un magasin a noté le niveau de la demande journalière pour un produit pendant cent jours ouvrables consécutifs en 2008 :

Nombre x_i d'unités demandées par jour	Nombre n_i de jours où l'on a vendu x_i	Effectifs cumulés croissants $n_i \nearrow$	Fréquence f_i	Fréquence cumulée croissante $f_i \nearrow$
0	5	5	0,05	0,05
1	15	20	0,15	0,20
2	23	43	0,23	0,43
3	22	65	0,22	0,65
4	16	81	0,16	0,81
5	9	90	0,09	0,9
6	5	95	0,05	0,95
7 et plus	5	100	0,05	1
TOTAL	100	–	1	–

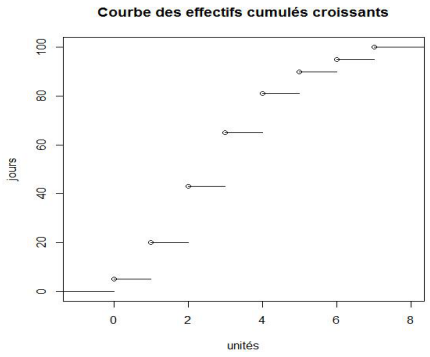
– Dans un repère orthogonal, on porte en abscisse les valeurs définissant les classes et en ordonnée les effectifs. Pour rendre le diagramme plus lisible, on trace les segments de droite correspondant aux ordonnées des points ainsi définis, et on obtient ce qu'on appelle un **diagramme des effectifs en bâtons**.



On obtient le **diagramme en bâtons des fréquences** par simple changement d'échelle sur l'axe des ordonnées. Par exemple, l'effectif "5" devient la fréquence "0,05".

– Intéressons-nous maintenant au **diagramme des effectifs cumulés croissants, en escalier** : dans un repère orthogonal, on porte en abscisse les valeurs définissant les classes et en ordonnée les effectifs

cumulés croissants.



Comme précédemment, on obtient le diagramme en bâtons des **fréquences cumulées croissantes** par simple changement d'échelle sur l'axe des ordonnées.

(c) Caractère quantitatif continu

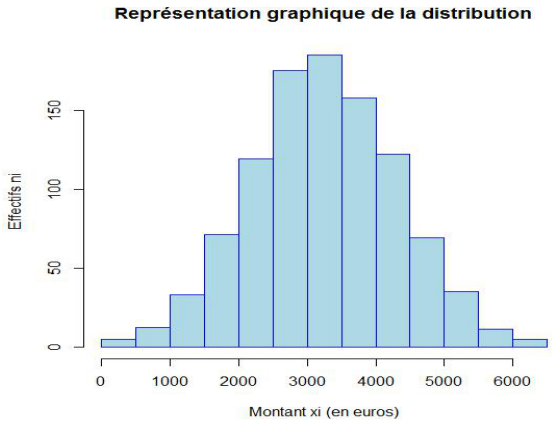
Exemple 1.2.4 On relève dans une banque à une date donnée les montants des économies de 1000 clients en euros. Les résultats obtenus sont les suivants :

Montant des économies (en euros) x_i	Nombre n_i de clients	Effectifs cumulés croissants
$[0; 500[$	5	5
$[500; 1000[$	12	17
$[1000; 1500[$	33	50
$[1500; 2000[$	71	121
$[2000; 2500[$	119	240
$[2500; 3000[$	175	415
$[3000; 3500[$	185	600
$[3500; 4000[$	158	758
$[4000; 4500[$	122	880
$[4500; 5000[$	69	949
$[5000; 5500[$	35	984
$[5500; 6000[$	11	995
6000 et plus	5	1000
TOTAL	1000	–

Toutes les autres classes ayant la même amplitude 500, on convient d'assimiler la classe "6000 et plus" à $[6000; 6500[$.

Le graphique utilisé pour représenter ce type de données est appelé **histogramme des effectifs.** Les effectifs des classes sont proportionnels aux aires des rectangles représentant les classes.

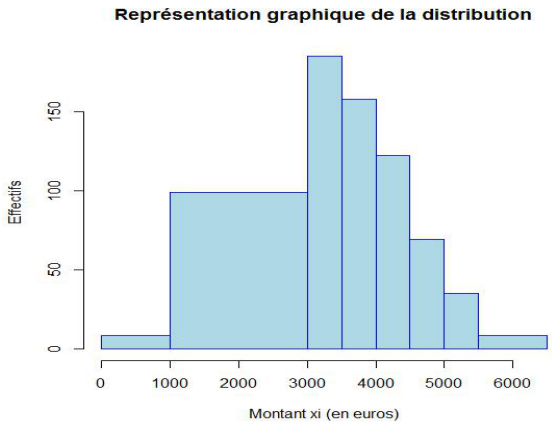
Diagramme des effectifs, histogramme.



Imaginons maintenant qu'on modifie le tableau précédent afin d'obtenir des classes d'amplitudes différentes :

Montant des économies (en euros) x_i	Nombre n_i de clients	Effectifs cumulés croissants
$[0; 1000[$	17	17
$[1000; 3000[$	398	415
$[3000; 3500[$	185	600
$[3500; 4000[$	158	758
$[4000; 4500[$	122	880
$[4500; 5000[$	69	949
$[5000; 5500[$	35	984
$[5500; 6500[$	16	1000
TOTAL	1000	–

On obtient l'histogramme des effectifs ci-dessous.



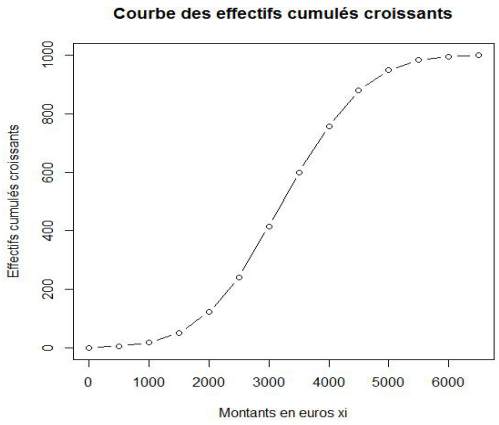
On remarque alors de manière générale que

- l'histogramme est constitué de la juxtaposition de rectangles dont les bases sont les différentes classes et dont les surfaces sont proportionnelles aux fréquences,
- si les classes sont toutes d'amplitude égale, les hauteurs sont proportionnelles aux fréquences et donc aux effectifs.

Il faut noter comme c'était le cas précédemment que l'histogramme des fréquences est obtenu par simple changement d'échelle sur l'axe des ordonnées.

Reprenons l'exemple initial et tâchons de représenter graphiquement les effectifs cumulés croissants. Ceci est réalisé à l'aide de ce qu'on nomme une **courbe polygonale**.

Diagramme des effectifs cumulés croissants, courbe polygonale



On remarquera que ce diagramme n'est pas à proprement parler une courbe mais bien une succession de segments d'où la dénomination "courbe polygonale".

Enfin, il est à noter que le diagramme des fréquences cumulées croissantes est obtenu à l'aide du diagramme précédent par simple changement d'échelle sur l'axe des ordonnées.

1.3 Caractéristiques de position

On a vu dans la première partie comment condenser les informations pour les rendre plus lisibles et utilisables. On est ainsi passé d'une liste de plusieurs dizaines, centaines, éventuellement milliers de données à un tableau ou un graphique reposant sur un regroupement de celles-ci en quelques classes.

On souhaite maintenant synthétiser davantage l'information pour les caractères quantitatifs en mettant en évidence des nombres permettant de décrire au mieux la population observée.

La première idée concerne naturellement la "tendance centrale" de la population.
Cela peut signifier :
- calculer une moyenne,
- chercher un nombre séparant la population en deux parties représentant chacune 50% de l'effectif total,
- choisir la (ou une) classe de plus grand effectif.

Ces trois points de vue présentent de l'intérêt et conduisent à définir des **caractéristiques de position** utilisées en statistique.

1.3.1 Le mode (ou dominante)

Définition 1.3.1 *On appelle* ***mode*** *d'une série statistique la ou les valeurs du caractère dont l'effectif est le plus élevé. Dans le cas d'une répartition à l'aide de classes, la classe dont l'effectif est le plus élevé est appelée* ***classe modale****, le mode étant le centre de la classe.*

Remarque 1.3.1

- Le mode correspond à un sommet sur l'histogramme ou le diagramme en bâtons.
- Il peut exister des séries unimodales ou plurimodales (dans le cas ou plusieurs classes ont le même effectif maximal).
- Le mode est un caractère peu utilisé en pratique car il ne fait pas intervenir l'ensemble des valeurs.

- Considérons l'exemple 1.2.3. Le mode de la série statistique est égal à 2. En effet, c'est la valeur de la série qui admet l'effectif le plus élevé c'est-à-dire 23. La série est unimodale.
- Considérons l'exemple 1.2.4. Les éléments de la série statistique sont répartis à l'aide de classes. $[3000; 3500[$ est la classe modale puisque c'est elle qui admet l'effectif le plus élevé à savoir 185, le mode étant égal à 3250. La série est unimodale.

1.3.2 La moyenne

Pour Pythagore (V^e siècle avant J.-C.), "les nombres sont les éléments de toutes choses, tout est nombre, l'harmonie est divine, elle consiste en rapports numériques". On doit à cette école pythagoricienne plusieurs sortes de moyennes (moyenne arithmétique, moyenne géométrique, moyenne harmonique). Cette dernière fut d'ailleurs inventée par Hippase (un des premiers pythagoriciens) qui travaillait sur les différents types de liens que trois nombres peuvent entretenir entre-eux et qu'on nommait alors "médiétés".

Définition 1.3.2 *La **moyenne arithmétique** de n nombres $y_1, y_2, \ldots, y_n$ est*

$$\boxed{\overline{y} = \frac{y_1 + y_2 + \ldots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.}$$

Les séries statistiques (à une variable quantitative) peuvent se présenter directement ou indirectement sous l'une des trois formes suivantes :

- 1^{er} **cas** : on dispose de la liste des n éléments $x_1, x_2, \ldots, x_n$. La moyenne est alors obtenue à l'aide de la formule

$$\boxed{\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

  Reprenons l'exemple 1.2.1 : la série consiste en une suite de 40 éléments, sa moyenne arithmétique vaut

$$\overline{x} = \frac{5 + 7 + 2 + 6 + 3 + \ldots + 6 + 7 + 8 + 5 + 5}{40} = \frac{216}{40} = 5,4$$

- 2^e **cas** : on dispose du tableau des effectifs n_i des p classes x_i.

$$\boxed{\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{n} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i}$$

  Considérons une fois encore l'exemple 1.2.1 une fois que la synthèse est réalisée, la moyenne arithmétique vaut alors

$$\overline{x} = \frac{1 \times 0 + 2 \times 1 + 1 \times 2 + \ldots + 5 \times 8 + 2 \times 9 + 1 \times 10}{40} = \frac{216}{40} = 5,4$$

- 3^e **cas** : on dispose du tableau des effectifs n_i des p classes $[a_i; b_i[$ de centre $c_i = \dfrac{a_i + b_i}{2}$.

$$\boxed{\overline{x} = \frac{n_1 c_1 + n_2 c_2 + \ldots + n_p c_p}{n} = \frac{1}{n}\sum_{i=1}^{p} n_i c_i}$$

  On considère l'exemple 1.2.4 : on travaille dans ce cas avec un caractère quantitatif continu, on va donc considérer pour chacune des classes son centre. La moyenne arithmétique vaut alors :

$$\overline{x} = \frac{5 \times 250 + 12 \times 750 + \ldots + 11 \times 5750 + 5 \times 6250}{1000} = \frac{3243000}{1000} = 3243$$

Remarque 1.3.2

- Dans le deuxième cas, la population est donnée avec autant de précision que dans le premier. Au contraire, dans le troisième cas, nous ne connaissons pas la valeur exacte de chaque élément à l'intérieur de sa classe $[a_i, b_i]$.
- La formule donnée pour $\overline{x}$ dans le troisième cas est valable lorsque, dans chaque classe $[a_i, b_i[$, tous les éléments sont concentrés au milieu c_i de la classe mais cette hypothèse est rarement satisfaite.
  En revanche, on peut admettre plus fréquemment que, dans chaque classe $[a_i, b_i[$, les n_i éléments sont uniformément répartis et dans ce cas, la formule est correcte.

1.3.3 La médiane

En économie, la moyenne arithmétique n'est pas toujours la caractéristique de position la plus pertinente ; il en est de même des autres moyennes. C'est pour cette raison qu'on définit la médiane.

Définition 1.3.3 *Dans une série statistique rangée en ordre de grandeur croissant (ou décroissant), la* ***médiane*** *est la valeur qui occupe la position centrale.*

Cette valeur coupe donc en deux sous-ensembles égaux l'ensemble de départ. Le calcul de cette valeur va bien évidemment dépendre de la nature de la série et plus précisément de celle de la variable.

(a) Variable discrète

• 1^{er} cas : la série comporte un nombre **impair** de valeurs, soit $2k+1$ valeurs, la médiane sera la $(k+1)$-ième valeur.

$$\underbrace{x_1 \leq x_2 \leq \ldots \leq x_{k-1} \leq x_k}_{k\ valeurs} \leq x_{k+1} \leq \underbrace{x_{k+2} \leq x_{k+3} \leq \ldots \leq x_{2k} \leq x_{2k+1}}_{k\ valeurs}$$

Exemple 1.3.1 Soit la série

$$5,7,8,3,4,3,4,9,4,5,10,9,7$$

On vérifie que la série comporte $13 = 2 \times 6 + 1$ valeurs. Si la série est ordonnée, on peut affirmer que la médiane est la 7-ième valeur. Rangeons cette série en ordre de grandeur croissant :

$$3,3,4,4,4,5,\textcircled{5},7,7,8,9,9,10$$

la médiane vaut donc Mé=5.

• 2^e cas : la série comporte un nombre **pair** de valeurs, soit $2k$ valeurs, la médiane sera la $1/2$ somme des k-ième et $(k+1)$-ième valeurs.

$$\underbrace{x_1 \leq x_2 \leq \ldots x_{k-1} \leq x_k}_{k\ valeurs} \leq \underbrace{x_{k+1} \leq x_{k+2} \leq \ldots \leq x_{2k-1} \leq x_{2k}}_{k\ valeurs}$$

Exemple 1.3.2 Soit la série

$$5,5,5,7,6,5,6,4,3,7$$

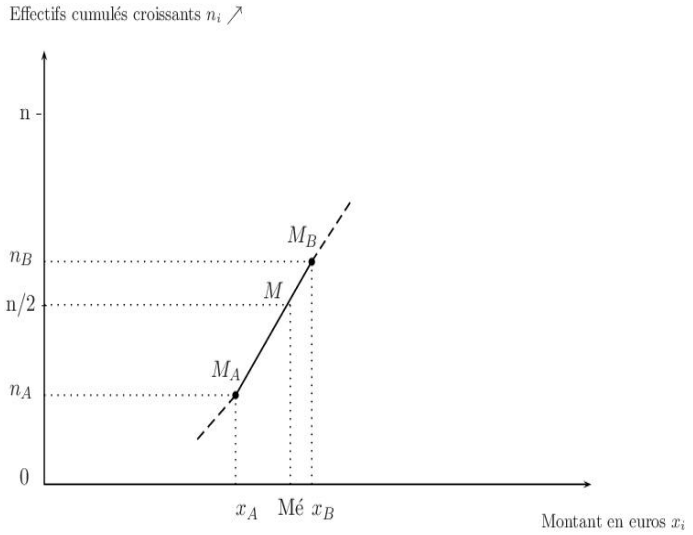
On vérifie que la série comporte $10 = 2 \times 5$ valeurs. Si la série est ordonnée, on peut affirmer que la médiane est la $1/2$ somme des 5-ième et 6-ième valeurs. Rangeons cette série en ordre de grandeur croissant :

$$3,4,5,5,\textcircled{5},\textcircled{5},6,6,7,7$$

la médiane vaut donc Mé=$\dfrac{5+5}{2}$.

(b) Variable continue

Dans ce cas, la détermination de la médiane est très différente. Il faut tout d'abord repérer la classe qui contient la médiane à l'aide de la moitié de l'effectif total $\frac{n}{2}$ soit $[x_A, x_B[$. Cette classe peut également être repérée sur le diagramme des effectifs (ou fréquences) cumulés croissants :



Il est possible de déterminer explicitement la valeur de la médiane en utilisant l'interpolation linéaire. En voici le principe.

Considérons la droite $(\mathcal{D})$ d'équation $y = ax + b$ passant par les points $M_A \begin{pmatrix} x_A \\ n_A \end{pmatrix}$ et $M_B \begin{pmatrix} x_B \\ n_B \end{pmatrix}$. Les coordonnées de ces deux points vérifient bien-évidemment l'équation de $(\mathcal{D})$ ce qui nous donne le système linéaire suivant :

$$\begin{cases} n_A = ax_A + b & (1) \\ n_B = ax_B + b & (2) \end{cases}$$

En soustrayant (1) à (2) puis en isolant "a", on obtient l'égalité

$$a = \frac{n_B - n_A}{x_B - x_A}.$$

Or le point $M \begin{pmatrix} \text{Mé} \\ n/2 \end{pmatrix}$ vérifie également l'équation de $(\mathcal{D})$ donc on a le second système

$$\begin{cases} n_A = ax_A + b & (1) \\ \frac{n}{2} = a\text{Mé} + b & (2) \end{cases}$$

En réalisant les mêmes étapes que précédemment, on obtient

$$a = \frac{\frac{n}{2} - n_A}{\text{Mé} - x_A}.$$

ce qui permet d'affirmer que

$$\frac{n_B - n_A}{x_B - x_A} = \frac{\frac{n}{2} - n_A}{\text{Mé} - x_A} \Leftrightarrow \boxed{\frac{\text{Mé} - x_A}{x_B - x_A} = \frac{\frac{n}{2} - n_A}{n_B - n_A}}$$

Il est assez simple de retenir cette formule à l'aide des encadrements suivants :

x_A	Mé	x_B
n_A	$\frac{n}{2}$	n_B

On récupère ainsi l'expression de la médiane

$$\boxed{\text{Mé} = x_A + \frac{(\frac{n}{2} - n_A)}{(n_B - n_A)}(x_B - x_A)}$$

Il est possible de travailler avec les fréquences plutôt que les effectifs. Dans ce cas, les seules modifications à apporter concernent les effectifs n_A, n_B et $\frac{n}{2}$. Cette dernière valeur devient $0,5$ si on travaille avec des proportions, i.e.,

$$\boxed{\text{Mé} = x_A + \frac{(0,5 - f_A)}{(f_B - f_A)}(x_B - x_A)}$$

et 50 (%) si on travaille avec des pourcentages, i.e,

$$\boxed{\text{Mé} = x_A + \frac{(50 - p_A)}{(p_B - p_A)}(x_B - x_A)}$$

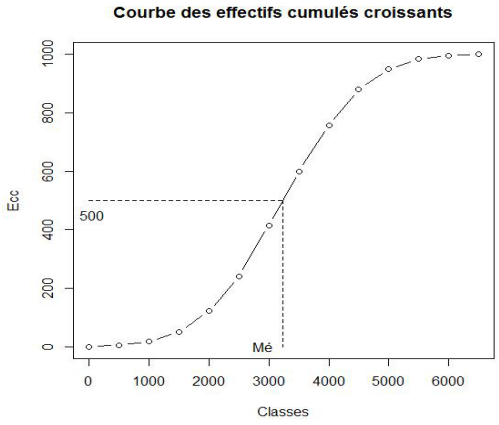
Considérons l'exemple 1.2.4.

Montant des économies **(en euros)** x_i	Nombre n_i de clients	$n_i \nearrow$	$f_i \nearrow$	$f_i \nearrow$ (%)
$[0;500[$	5	5	0,005	0,5
$[500;1000[$	12	17	0,017	1,7
$[1000;1500[$	33	50	0,05	5
$[1500;2000[$	71	121	0,121	12,1
$[2000;2500[$	119	240	0,24	24
$[2500;3000[$	175	415	0,415	41,5
$[3000;3500[$	185	600	0,6	60
$[3500;4000[$	158	758	0,758	758
$[4000;4500[$	122	880	0,88	88
$[4500;5000[$	69	949	0,949	94,9
$[5000;5500[$	35	984	0,984	98,4
$[5500;6000[$	11	995	0,995	99,5
$[6000;6500[$	5	1000	1	100
TOTAL	1000	–	1	100

Déterminons tout d'abord à l'aide du tableau la classe contenant la médiane.
On rappelle que dans le contexte de l'exemple, la médiane est le montant qui sépare l'ensemble de la population concernée en deux parties de même effectif. On considère ainsi $\frac{n}{2} = \frac{1000}{2} = 500$. Puis on se réfère à la colonne effectifs cumulés croissants ($ni \nearrow$) afin de trouver la classe correspondant à cet effectif.
On remarque tout d'abord que la valeur "500" n'apparaît pas explicitement. Il faut donc réussir à l'extraire du tableau :

- 415 personnes ont une économie comprise entre 0 et 3000 euros. On est donc certain que la médiane n'appartient pas à cet intervalle puisqu'il ne fait pas intervenir 500 personnes mais seulement 415.
- 600 personnes ont une économie comprise entre 0 et 3500 euros. On est donc certain que la médiane appartient à cet intervalle puisqu'il fait intervenir au moins 500 personnes. On en déduit que la médiane appartient à l'intervalle $[3000;3500[$.

On peut vérifier cette propriété à l'aide de la courbe des effectifs cumulés croissants.



On retrouve bien le fait, graphiquement, que Mé$\in [3000; 3500[$.

Déterminons la médiane explicitement : à l'aide du calcul précédent, on peut affirmer que

$$\begin{aligned}
\text{Mé} &= 3000 + \frac{(\frac{1000}{2} - 415)}{(600 - 415)}(3500 - 3000) && \text{(effectifs)} \\
&= 3000 + \frac{(0,5 - 0,415)}{(0,6 - 0,415)}(3500 - 3000) && \text{(fréquences)} \\
&= 3500 - \frac{(50 - 41,5)}{(60 - 41,5)}(3500 - 3000) && \text{(pourcentages)} \\
&\simeq 3229,73 \text{ à } 10^{-2} \text{ près.}
\end{aligned}$$

Remarque 1.3.3

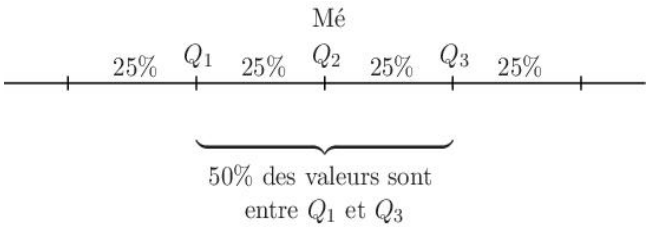
- Lorsque la population est répartie en classes $[a_i, b_i[$, la médiane peut donc être évaluée soit graphiquement soit par une interpolation affine (ou linéaire) à l'aide d'une courbe des effectifs cumulés en faisant l'hypothèse supplémentaire : **les éléments de la classe contenant la médiane sont uniformément répartis.**
- On doit distinguer la médiane Mé et la moyenne $\overline{x}$ d'une population. Le calcul de la moyenne fait intervenir toutes les données ce qui n'est pas le cas pour la détermination de la médiane. De plus, la moyenne est sensible aux variations des valeurs extrêmes de la série statistique, ce qui n'est pas le cas de la médiane.

1.3.4 Les quartiles

Le **quartile** est une extension de la médiane puisqu'il s'agit de partager l'effectif en quatre parties égales. Les quartiles sont au nombre de 3 et on les note Q_1, Q_2 et Q_3. Le quartile Q_2 correspond à la médiane Mé.

Définition 1.3.4 *Q_1 (respectivement Q_3) est la valeur telle que 25% (respectivement 75%) des valeurs de l'effectif total de la population étudiée lui sont inférieures et 75% (respectivement 25%) supérieures.*

On peut schématiser les quartiles de la manière suivante



Remarque 1.3.4

- Les quartiles Q_1 et Q_3 permettent d'apprécier l'importance de la dispersion d'une série autour de la médiane.
- $Q_3 - Q_1$ est appelée l'**amplitude** de l'intervalle interquartile $[Q_1; Q_3]$. 50% de la population de l'échantillon se retrouvent dans cet intervalle.
- On peut définir de même les déciles, les centiles dans le cas d'un effectif n très important.

On s'intéresse maintenant à la détermination pratique des quartiles dans le cas d'une variable continue. La méthode est la même que celle utilisée pour calculer la médiane :

- le premier quartile Q_1 correspond à l'abscisse du point d'ordonnée $0,25$ sur la courbe des fréquences cumulées croissantes (ou $\frac{n}{4}$ sur la courbe des effectifs cumulés croissants), sa valeur exacte est déterminée à l'aide de l'interpolation linéaire,

- le troisième quartile Q_3 correspond à la médiane,

- le troisième quartile Q_3 correspond à l'abscisse du point d'ordonnée $0,75$ sur la courbe des fréquences cumulées croissantes (ou $\frac{3n}{4}$ sur la courbe des effectifs cumulés croissants).

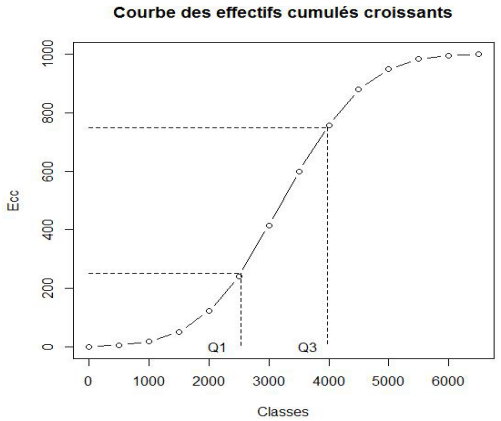
Considérons l'exemple 1.2.4.

On rappelle que la taille de l'échantillon est égale à 1000.

- Le premier quartile correspond à l'abscisse du point d'ordonnée $\frac{n}{4} = 250$. La valeur "250" n'apparaît pas explicitement dans le tableau, il faut donc réussir à l'extraire :
  . 240 personnes ont une économie comprise entre 0 et 2500 euros. On est donc certain que le premier quartile n'appartient pas à cet intervalle puisqu'il ne fait pas intervenir 250 personnes mais seulement 240.
  . 415 personnes ont une économie comprise entre 0 et 3000 euros. On est donc certain que Q_1 appartient à cet intervalle puisqu'il fait intervenir au moins 250 personnes.
  On en déduit que le premier quartile Q_1 appartient à l'intervalle $[2500; 3000[$.

- Le troisième quartile correspond à l'abscisse du point d'ordonnée $\frac{3n}{4} = 750$. La valeur "750" n'apparaît pas explicitement dans le tableau. Il faut donc réussir à l'extraire :
  . 600 personnes ont une économie comprise entre 0 et 3500 euros. On est donc certain que le troisième quartile n'appartient pas à cet intervalle puisqu'il ne fait pas intervenir 750 personnes mais seulement 600.
  . 758 personnes ont une économie comprise entre 0 et 4000 euros. On est donc certain que Q_3 appartient à cet intervalle puisqu'il fait intervenir au moins 750 personnes.
  On en déduit que le troisième quartile Q_3 appartient à l'intervalle $[3500; 4000[$.

On peut vérifier ces propriétés à l'aide de la courbe des effectifs cumulés croissants.



On retrouve bien graphiquement que $Q_1 \in [2500; 3000[$ et $Q_3 \in [3500; 4000[$.
Déterminons les quartiles Q_1 et Q_3 explicitement :

- on a l'encadrement suivant pour Q_1

2500	Q_1	3000
240	250	415

On a alors l'égalité :

$$\frac{Q_1 - 2500}{3000 - 2500} = \frac{250 - 240}{415 - 240}$$
$$\Leftrightarrow Q_1 = 2500 + 500 \times \frac{10}{175}$$
$$\Leftrightarrow Q_1 \simeq 2528,57 \text{ à } 10^{-2} \text{ près.}$$

- On a l'encadrement suivant pour Q_3

3500	Q_3	4000
600	750	758

On a alors l'égalité :

$$\frac{Q_3 - 3500}{4000 - 3500} = \frac{750 - 600}{758 - 600}$$
$$\Leftrightarrow Q_3 = 3500 + 500 \times \frac{150}{158}$$
$$\Leftrightarrow Q_3 \simeq 3974,68 \text{ à } 10^{-2} \text{ près.}$$

1.4 Caractéristiques de dispersion

Exemple 1.4.1 Les élèves A et B ont obtenu dans une matière spécifique les notes ci-dessous.

7,8,11,12,13,13,13 pour A,

4,7,9,12,13,13,19 pour B.

On peut vérifier que les séries de notes de A et B ont la même médiane (12), la même moyenne (11) et le même mode (13) et pourtant, ces deux séries de notes sont différentes : les notes de B sont plus dispersées que celles de A.
Aussi, à côté des caractéristiques de position, on est amené à introduire des caractéristiques de dispersion pour décrire plus précisément une population.

1.4.1 L'étendue

Afin de mesurer l'étalement des termes d'une série, on peut tout d'abord calculer l'étendue.

Définition 1.4.1 *L'**étendue** d'une série est la différence de ses valeurs extrêmes.*

Considérons EX5, on montre aisément que les étendues des séries de A et de B valent respectivement $e_A = 13 - 7 = 6$ et $e_B = 19 - 4 = 15$. Les notes de B sont donc plus étalées que celles de A.

1.4.2 L'écart absolu moyen

Pour étudier la dispersion des valeurs d'une série, on peut également calculer la moyenne des écarts entre chaque valeur et la moyenne arithmétique. Il est nécessaire cependant que tous les termes de la somme soient positifs d'où l'utilisation de la valeur absolue.

Définition 1.4.2 *L'**écart absolu moyen** d'une série statistique est la moyenne des valeurs absolues des écarts à la moyenne arithmétique $\overline{x}$*

$$e_m = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}| = \frac{1}{n}\sum_{i=1}^{p} n_i |x_i - \overline{x}|$$

Considérons l'exemple 1.4.1, supposons que l'on veuille mesurer la dispersion des valeurs des deux séries à l'aide de l'écart moyen (non absolu). On obtient

- pour A : $\frac{1}{7}[(7-11)+(8-11)+\ldots+(13-11)]=0$
- pour B : $\frac{1}{7}[(4-11)+(7-11)+\ldots+(19-11)]=0$

En fait, ce résultat est général :

$$\boxed{\frac{1}{n}[(x_1-\overline{x})+(x_2-\overline{x})+\ldots+(x_n-\overline{x})]=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})=0}$$

Preuve : $\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})=\frac{1}{n}[\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\overline{x}]==\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\overline{x}=\overline{x}-\frac{1}{n}n\overline{x}=\overline{x}-\overline{x}=0$

Comme les écarts moyens sont nuls, on comprend l'intérêt de calculer l'écart absolu moyen.

Considérons l'exemple 1.4.1,

- pour A : $e_m=\frac{1}{7}[|7-11|+|8-11|+\ldots+|13-11|]=14$
- pour B : $e_m=\frac{1}{7}[|4-11|+|7-11|+\ldots+|19-11|]=26$

On retrouve bien le fait que les notes de B sont plus dispersées que celles de A.

1.4.3 La variance et l'écart-type

Comme cela a été montré précédemment, le calcul de la moyenne des écarts entre chaque valeur de la série et la moyenne arithmétique nécessite que les termes de la somme soient positifs. Il existe un autre procédé élémentaire qui permet de surmonter cette difficulté : les carrés.

Définition 1.4.3 *La **variance** d'une série statistique est la moyenne des carrés des écarts à la moyenne arithmétique* $\overline{x}$

$$\boxed{V(X)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}\sum_{i=1}^{p}n_i(x_i-\overline{x})^2}$$

Toutefois, ce calcul n'est pas très commode : la somme nécessite le calcul de p soustractions, p mises au carré, p multiplications et enfin $p-1$ additions.

La variance peut être donnée sous une autre forme plus pratique :

$$\boxed{V(X)=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2=\frac{1}{n}\sum_{i=1}^{p}n_ix_i^2-\overline{x}^2}$$

Cete fois-ci, la somme ne nécessite plus les soustractions. Démontrons ce résultat pour la version par regroupements.

Preuve :

$$V(X)=\frac{1}{n}\sum_{i=1}^{p}n_i(x_i-\overline{x})^2=\frac{1}{n}\sum_{i=1}^{p}n_i[x_i^2-2x_i\overline{x}+\overline{x}^2]$$

et ceci d'après l'identité remarquable $(a-b)^2=a^2-2ab+b^2$. Ainsi,

$$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \frac{2}{n}\overline{x}\sum_{i=1}^{p} n_i x_i + \frac{1}{n}\overline{x}^2 \sum_{i=1}^{p} n_i \Leftrightarrow V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - 2\overline{x}^2 + \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{x}^2$$

Définition 1.4.4 *L'**écart-type** d'une série statistique est la racine carrée de sa variance $V(X)$.*

$$\boxed{\sigma(X) = \sqrt{V(X)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2}}$$

Cette caractéristique de dispersion est la plus utilisée. L'expérience montre que dans une distribution unimodale et symétrique,

- l'intervalle $[\overline{x} - \sigma(X), \overline{x} + \sigma(X)]$ contient environ 68% des valeurs de la série,
- l'intervalle $[\overline{x} - 2\sigma(X), \overline{x} + 2\sigma(X)]$ contient environ 95% des valeurs de la série.

Dans une distribution relativement symétrique, les résultats restent voisins de ceux indiqués.

Considérons une nouvelle fois l'exemple 1.2.4. Afin de calculer la variance et l'écart-type de la série, on réalise un tableau contenant toutes les données nécessaires à leur calcul (voir page suivante).

La quatrième colonne nous permet de calculer la moyenne arithmétique de la série : sa dernière ligne nous donne le nombre $\sum_{i=1}^{p} n_i x_i = 3243000$. Donc $\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \frac{3243000}{1000} = 3243$. L'économie moyenne des 1000 clients de la banque est de 3243 euros.

La cinquième colonne nous permet de calculer la variance de la série : sa dernière ligne nous donne le nombre $\sum_{i=1}^{p} n_i x_i^2 = 11662000000$. Donc $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{x}^2 = \frac{11662000000}{1000} - (3243)^2 = 1144951$.
On en déduit l'écart-type de la série : $\sigma(X) = \sqrt{V(X)} = \sqrt{1144951} \simeq 1070,02$ à 10^{-2} près.
Ce nombre peut être interprété de la manière suivante :

- l'intervalle $[\overline{x} - \sigma(X), \overline{x} + \sigma(X)] = [3243 - 1070,02; 3243 + 1070,02] = [2172,98; 4313,02]$ contient environ 68% des valeurs de la série,

- l'intervalle $[\overline{x} - 2\sigma(X), \overline{x} + 2\sigma(X)] = [3243 - 2 \times 1070,02; 3243 + 2 \times 1070,02] = [1102,96; 5383,04]$ contient environ 95% des valeurs de la série.

Montant des économies (en euros)	n_i	Centre de classe x_i	$n_i x_i$	$n_i x_i^2$
$[0;500[$	5	250	1250	312500
$[500;1000[$	12	750	9000	6750000
$[1000;1500[$	33	1250	41250	51562500
$[1500;2000[$	71	1750	124250	217437500
$[2000;2500[$	119	2250	267750	602437500
$[2500;3000[$	175	2750	481250	1323437500
$[3000;3500[$	185	3250	601250	1954062500
$[3500;4000[$	158	3750	592500	2221875000
$[4000;4500[$	122	4250	518500	2203625000
$[4500;5000[$	69	4750	327750	1556812500
$[5000;5500[$	35	5250	183750	964687500
$[5500;6000[$	11	5750	63250	363687500
$[6000;6500[$	5	6250	31250	195312500
TOTAL	1000	–	3243000	11662000000

Remarque 1.4.1 Le calcul de la variance ne peut être réalisé avant celui de la moyenne arithmétique.

1.5 Paramètres de concentration

On se restreint au cas particulier de la masse salariale versée chaque mois par l'entreprise. La question est la suivante : "Cette masse est-elle répartie de manière égale sur l'ensemble du personnel ou bien au contraire seules quelques personnes s'en octroient-elles la plus grande partie ?"

1.5.1 Définitions

Pour répondre à la question précédente, on considère une série statistique de forme générale $(x_i, n_i)_{1\leq i\leq p}$. On définit

- la fréquence cumulée croissante $\boxed{p_i = \sum_{j\leq i} f_j}$ et
- le coefficient q_i comme étant le rapport entre la masse salariale cumulée divisée par la masse salariale totale $M = \sum_{j\leq p} n_j x_j$: $\boxed{q_i = \dfrac{\sum_{j\leq i} n_j x_j}{M}}$.

Illustrons ces formules dans le cadre de l'exemple suivant.

Exemple 1.5.1 Les salaires (en euros) des employés d'une entreprise sont répartis de la manière suivante.

Classes	x_i	n_i	$\sum n_j$	p_i	$n_i x_i$	$\sum n_j x_j$	q_i
$[3000;4000[$	3500	22	22	0,22	77000	77000	0,140
$[4000;5000[$	4500	18	40	0,40	81000	158000	0,287
$[5000;7000[$	6000	47	87	0,87	282000	440000	0,799
$[7000;10000[$	8500	13	100	1	110500	550500	1
Total	–	100	–	–	550500	–	–

On utilise les notations suivantes :

. $N = \sum_{i=1}^{4} n_i = 100$,

. $M = \sum_{i=1}^{4} n_i x_i = 550500$,

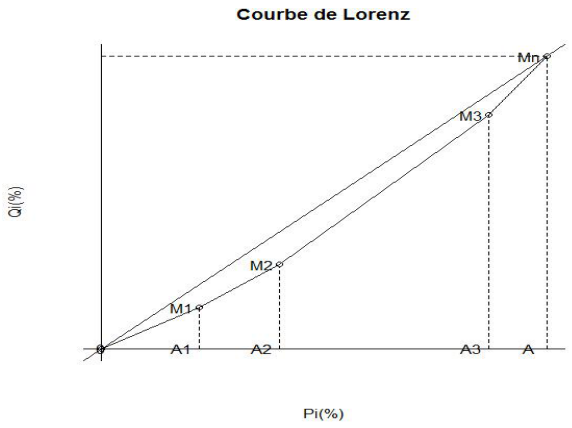
. $\overline{x} = \frac{M}{N} = \frac{550500}{100} = 5505$ qui correspond au salaire moyen dans l'entreprise.

Interprétation de la colonne des q_i : si on considère sa deuxième ligne, $28,7\%$ de la masse salariale est distribuée à 40% des employés, ceux gagnant moins de 5000 (euros).

1.5.2 La courbe de Gini ou de Lorenz

La courbe de Gini ou de Lorenz va nous permettre de mesurer la concentration de la masse salariale graphiquement. Elle est représentée à l'aide des points (p_i, q_i) avec $p_0 = q_0 = 0$.

Considérons l'exemple 1.5.1, la courbe de Gini associée est la suivante



La courbe polygonale $(OM_1M_2M_3M_n)$ représente la courbe de Gini. Afin de mesurer la concentration de cette courbe, il faut la comparer à une courbe de référence. La droite en pointillés appelée **bissectrice** représente une concentration salariale nulle c'est-à-dire que les salaires sont répartis de manière égale sur l'ensemble du personnel.

Plus la courbe de Gini sera proche de la bissectrice, plus la concentration sera faible. Plus elle en sera éloignée, plus la concentration sera forte. Néanmoins, cette mesure graphique n'est pas suffisante pour quantifier précisément le niveau de concentration de la masse salariale. Pour pallier ce manque de précision on calcule l'indice de Gini.

1.5.3 L'indice de la concentration ou indice de Gini

Définition 1.5.1 *On appelle* ***indice de concentration*** *ou* ***indice de Gini*** *le rapport i de la surface S comprise entre la courbe de concentration et la bissectrice du repère avec la surface S_0 du triangle OAM_n. Il apparaît que*

$$\boxed{i = \frac{S}{S_0}}$$

vérifie l'inégalité $0 \leq i \leq 1$

Interprétation :

– lorsque i est faible (ou proche de 0), la courbe est proche de la bissectrice, la série est peu concentrée, la répartition de la masse salariale est assez homogène,
– lorsque i est proche de 1, la courbe de Gini reste longtemps proche de l'axe des abscisses, la série est très concentrée, il y a une répartition très inégalitaire de la masse salariale.

1.5.4 Calcul du coefficient de Gini

Selon qu'on travaille avec des pourcentages ou des fréquences, les valeurs de S et S_0 s'expriment différemment.

- Calcul de S_0 : par définition,

$$\boxed{\text{Aire}_{\text{ triangle}} = \frac{\text{base b} \times \text{hauteur h}}{2}}$$

donc $S_0 = \dfrac{1 \times 1}{2} = \dfrac{1}{2}$ en termes de fréquences et $S_0 = \dfrac{100 \times 100}{2} = 5000$ en termes de pourcentages.

- Calcul de S (en termes de pourcentages) :

$$\begin{array}{rcl} S & = & (\text{Aire du triangle } OAM_n) - (\sum \text{aires des trapèzes } A_iA_{i+1}M_{i+1}M_i) \\ & = & 5000 - (\sum \text{aires des trapèzes } A_iA_{i+1}M_{i+1}M_i). \end{array}$$

– Le premier trapèze est en fait un triangle. On peut calculer son aire :
$S_1 = \text{aire}(OA_1M_1) = \dfrac{p_1 \times q_1}{2} = 154$
– On rappelle que pour les trapèzes, l'aire est donnée par la formule

$$\boxed{\text{Aire}_{\text{ trapèze}} = \frac{[(\text{petite base b}) + (\text{grande base B})] \times (\text{hauteur h})}{2}}$$

Ainsi
$S_2 = \text{aire}(A_1A_2M_2M_1) = \dfrac{1}{2}(p_2 - p_1)(q_1 + q_2) = \dfrac{1}{2}(18 \times 42,7) = 384,3$
– puis
$S_3 = \text{aire}(A_2A_3M_3M_2) = \dfrac{1}{2}(p_3 - p_2)(q_2 + q_3) = \dfrac{1}{2}(47 \times 108,6) = 2552,1$
– et
$S_4 = \text{aire}(A_3AM_nM_3) = \dfrac{1}{2}(100 - p_3)(q_3 + 100) = \dfrac{1}{2}(13 \times 179,9) = 1169,35$

Finalement $S = 5000 - (154 + 384,3 + 2552,1 + 1169,35) = 740,25$.

On en déduit que $i = \dfrac{740.25}{5000} = 0,14805$ et on peut affirmer ainsi que la concentration est faible puisque l'indice de Gini est proche de 0. La masse salariale est répartie de manière égalitaire.

1.5.5 La médiale

Il existe un autre moyen de mesurer la concentration de la masse salariale et qui se nomme la **médiale**.

Définition 1.5.2 *On appelle **médiale** de la série statistique, notée M_l, la première valeur du caractère à partir de laquelle la moitié de la masse salariale a été étudiée, c'est-à-dire la valeur du caractère qui partage la masse salariale en deux parties égales. Dans une répartition par classes, on procède par interpolation linéaire.*

Il existe deux possibilités pour déterminer la médiale
– soit déterminer la valeur du caractère correspondant à $\dfrac{M}{2}$ dans la colonne "masse salariale cumulée"
– soit déterminer la valeur du caractère qui correspond à $\dfrac{1}{2}$ dans la colonne « q_i ».

Considérons l'exemple 1.5.1. Le calcul de la médiale se fait de la même manière que celui de la médiane. On repère tout d'abord $\frac{1}{2}$ dans la colonne q_i. Si $\frac{1}{2}$ n'apparaît pas explicitement dans le tableau, on utilise les valeurs qui l'encadrent :

5000	M_l	7000
0,287	$\frac{1}{2}$	0,799

Il suit par interpolation linéaire que

$$M_l = 5000 + 2000 \times \frac{0,213}{0,512} = 5832,03$$

Ainsi, 50% de la masse salariale est distribuée aux salariés gagnant moins de 5832,03 euros et 50% à ceux gagnant plus de 5832,03 euros.

1.6 Exercices

Exercice 1 Dans le fichier du service ORL du centre hospitalier de Dunkerque, on trouve pour chaque patient les informations suivantes :

- sexe,
- âge,
- profession,
- poids,
- taille,
- groupe sanguin.

1. Quelle est la population étudiée ? Quels sont les individus ?
2. Donner le type de chacune des variables statistiques ci-dessus, en précisant éventuellement leurs modalités.

Exercice 2 Pour chaque commune française de plus de 20000 habitants, on note

- le département auquel elle appartient,
- le nombre de ses habitants,
- le nombre de ses établissements d'enseignement secondaire.

Reprendre les questions de l'exercice précédent.

Exercice 3 D'après l'INSEE, la structure sociale de la population active du Littoral et de la région Nord-Pas de Calais était en 1990 la suivante :

CSP	Littoral	Région
Agriculteurs	3,2	2
Artisans, Commerçants Chefs d'entreprise	5,7	5,3
Professions libérales Cadres supérieurs	5,6	5,6
Professions intermédiaires	13,8	14,4
Employés	17,7	17,8
Ouvriers	26,1	27
Autres (*)	27,9	26,6

(*) Autres : retraités, inactifs, chômeurs ...

1. Quelle est la variable statistique étudiée ? Quel est son type ?
2. Quelles représentations graphiques peut-on envisager ?
3. Représenter sur un même graphique ces deux distributions.

Exercice 4 On donne dans le tableau ci-dessous la répartition des étudiants inscrits à l'Université du Littoral 1998/1999 par secteurs disciplinaires :

Lettres	5%
Langues	13%
Sciences humaines et sociales	10%
Science de la nature et de la vie	7%
Science et structure de la matière	13%
Sciences et technologies	8%
Sport	6%
Droit	11%
Sciences économiques et gestion	10%
SESA	17%

Faire le diagramme à secteurs représentant cette distribution.

Exercice 5 Le maire d'une commune rurale située dans une zone d'élevage et de polyculture a fait relever la superficie des 70 exploitations agricoles de la commune. On obtient la distribution statistique suivante :

Superficie (en ha)	Nombre d'exploitations
0 à moins de 20	7
20 à moins de 40	20
40 à moins de 50	18
50 à moins de 60	10
60 à moins de 80	15

1. Quelle est la variable étudiée ? Quel est son type ?
2. Représenter graphiquement cette distribution.

Exercice 6 Un syndicat de salariés a réalisé une enquête sur les salaires du personnel ouvrier d'un groupe industriel. Il a obtenu, pour les personnes ayant travaillé toute l'année à temps complet, la distribution de salaires annuels suivante :

Salaire annuel	Nombre d'ouvriers
moins de 5000 euros	3145
5000 à moins de 5800 euros	2465
5800 à moins de 6600 euros	4675
6600 à moins de 8200 euros	11220
8200 à moins de 9800 euros	9180
9800 à moins de 13000 euros	8160
13000 euros et plus	3655
Total	42500

La masse des salaires correspondant à la première classe (moins de 5000 euros) s'élève à $10,693$ millions d'euros tandis que celle correspondant à la dernière classe s'élève à $53,363$ millions d'euros.
Reprendre les questions de l'exercice précédent.

Exercice 7 En réponse à une offre d'emploi visant à recruter une secrétaire sténodactylo, 7 candidates se sont présentées. Le test qui leur est proposé consiste à dactylographier un texte préalablement noté en sténo. Le tableau suivant donne le nombre d'erreurs commises par chaque candidate.

Candidate	1	2	3	4	5	6	7
Nombre d'erreurs	1	5	4	3	7	6	10

1. Calculer la moyenne et déterminer la médiane de cette distribution.
2. Une huitième candidate arrive en retard et est admise à passer le test. Elle fait 9 erreurs.
   Calculer la moyenne et déterminer la médiane de cette nouvelle distribution.

Exercice 8 La distribution selon le nombre d'enfants des 110 familles inscrites sur la liste d'attente d'un office de HLM est la suivante :

Nombre d'enfants	Nombre de familles
0	18
1	27
2	27
3	18
4	15
5	5

1. Représenter graphiquement cette distribution.
2. Calculer le nombre moyen d'enfants de ces familles.
3. Déterminer la médiane et le mode de cette distribution et calculer son écart-type.
4. Quelle est la proportion de familles comptant au plus 3 enfants ?
5. Quelle est la proportion de familles comptant moins de 3 enfants ?

Exercice 9 Calculer la moyenne, la médiane, les quartiles, l'écart-type et donner la classe modale de la variable statistique de l'exercice 5.
Calculer la proportion d'exploitations agricoles dont la superficie est inférieure à 45 hectares.

Exercice 10 Calculer la moyenne, la médiane, l'écart-type et donner la classe modale de la variable statistique de l'exercice 6.

Exercice 11 Le tableau suivant donne la répartition des entreprises d'au moins 20 salariés dans le secteur de l'industrie, en France en 1994, ainsi que les parts respectives du chiffre d'affaires total de ce secteur.

Nombre de salariés / part (en %)	$[20, 50[$	$[50, 100[$	$[100, 200[$	$[200, 500[$	500 et plus
Nombre	$59,7\%$	$18,5\%$	$10,9\%$	7%	$3,9\%$
Chiffre d'affaires	$8,3\%$	$6,2\%$	$8,1\%$	$13,4\%$	64%

1. Tracer la courbe de Lorenz associée à cette distribution.
2. Sur la courbe de Lorenz, lire la part des entreprises (ayant le moins de salariés) qui réalisent 50% du chiffre d'affaires total de ce secteur.
3. Calculer l'indice de Gini associé à cette distribution.
4. Calculer la médiane de cette distribution.

Exercice 12 Les notes de 1000 étudiants lors d'une épreuve de statistiques se répartissent de la manière suivante :

Notes x_i	6	7	8	9	10	11	12	13	14	15	16	17	18
Effectifs n_i	10	23	45	78	116	147	162	148	117	77	46	20	11

1. Présenter un tableau où figureront entre-autres :
   – les fréquences,
   – les effectifs cumulés croissants,
   – les fréquences cumulées croissantes.
2. Effectuer un diagramme en bâtons des fréquences de la série statistique.
3. Tracer la courbe des fréquences cumulées croissantes.

Exercice 13 La répartition des salaires des employés d'une entreprise est donnée dans le tableau suivant :

Salaires en euros	Centres	Effectifs	Fréquences (%)	Effectifs cumulés croissants
$[900; 1100[$			5	
$[1100; 1200[$			$8,5$	
$[1200; 1300[$			$23,75$	
$[1300;\quad[$			25	
$[\quad;\quad[$	1500			
$[1600; 1700[$			$6,75$	400

1. Compléter le tableau.
2. Combien d'employés ont un salaire inférieur à 1300 euros ?
3. Réaliser un histogramme des effectifs de la série statistique.

Exercice 14 Le personnel d'une entreprise se répartit ainsi :

Fonction	Nombre
Manœuvres	96
Ouvriers professionnels	288
Ouvriers qualifiés	184
Employé(s)s	40
Cadres et direction	32

1. Calculer le pourcentage que représente chaque catégorie.
2. Représenter ces pourcentages par un diagramme à secteurs circulaire. (Mesure des secteurs au degré le plus proche)

Exercice 15 On considère la série statistique suivante :

3, 4, 9, 10, 7, 6, 5, 4, 4, 3, 7, 9, 8

On notera X la variable statistique prenant ces valeurs.

1. Déterminer la médiane pour la variable X.
2. Extraire de ces données un tableau de distribution contenant entre-autres les effectifs et les effectifs cumulés.
3. Retrouver la médiane pour X à l'aide du tableau.
4. Représenter graphiquement à l'aide d'un diagramme en bâtons la distribution des fréquences.

Exercice 16 On considère le nombre de pièces A sortant d'un entrepôt pendant 40 jours consécutifs :

10, 12, 17, 6, 13, 20, 18, 18, 16, 15, 15, 14, 7, 8, 9, 11, 11, 12, 9, 9,
12, 14, 15, 15, 10, 12, 7, 7, 13, 14, 14, 14, 16, 16, 15, 18, 8, 9, 8, 9.

1. Regrouper par modalités cette série statistique en complétant, après l'avoir reproduit, le tableau suivant.

Nombre de pièces A	6	7	...
Effectif	1	3	...

2. Déterminer la médiane de cette série statistique.
3. Déterminer une valeur approchée à 10^{-2} près de la moyenne $\overline{x}$ de cette série statistique.
4. Calculer la variance et l'écart-type de cette série statistique à 10^{-2} près. Interpréter la valeur de l'écart-type obtenue en termes de dispersion.

Exercice 17 Avant d'accepter un contrat de livraison de véhicules, une société d'équipements automobiles établit une statistique de production journalière sur 100 jours.
Le nombre de véhicules équipés journellement se répartit comme suit :

Production journalière de véhicules équipés	Nombre de jours
95	1
96	3
97	6
98	8
99	10
100	13
101	18
102	14
103	9
104	8
105	6
106	2
107	2
Total	100

Déterminer la valeur moyenne de la production journalière et une valeur approchée à 10^{-2} près de l'écart-type de cette production.

Exercice 18 Un nouveau responsable de magasin a enregistré au cours de ses 40 premières semaines d'activité le nombre X de tonnes de marchandises qu'il a stocké hebdomadairement. Il a obtenu les résultats suivants :

$$5, 7, 2, 6, 3, 4, 8, 5, 4, 3, 9, 6, 5, 7, 6, 8, 3, 4, 4, 0, 8, 6, 7, 1, 5, 5, 4, 6, 6, 10, 9, 8, 1, 5, 5, 6, 7, 8, 5, 5$$

1. Déterminer la distribution de fréquences (n_i) et la distribution de fréquences cumulées $\left(\sum_{j=1}^{i} n_j\right)$ de cette variable X et représenter graphiquement ces deux distributions à l'aide respectivement d'un diagramme en bâtons et d'un graphique en escaliers.

2. Calculer la médiane pour la variable X.

Exercice 19 On se donne le tableau de données suivantes :

i	Classes	Centres x_i	Effectifs n_i	Effectifs cumulés croissants
1	$500 \leq X < 1500$		31	
2	$1500 \leq X < 2500$		46	
3	$2500 \leq X < 3500$		86	
4	$3500 \leq X < 4500$		151	
5	$4500 \leq X < 5500$		197	
6	$5500 \leq X < 6500$		167	
7	$6500 \leq X < 7500$		107	
8	$7500 \leq X < 8500$		65	
9	$8500 \leq X < 9500$		32	
10	$9500 \leq X < 10500$		18	

1. Remplir le tableau de distribution.
2. Tracer la courbe des effectifs cumulés croissants et en déduire approximativement la valeur de la médiane.
3. Déterminer une valeur de la médiane à 10^{-3} près.

Exercice 20 Un hypermarché assure les livraisons à domicile. Le tableau suivant donne le nombre de livraisons effectuées dans un trimestre, selon la distance du magasin au point de livraison.

Distance (en km)	$[0;5[$	$[5;10[$	$[10;15[$	$[15;20[$	$[20;25[$	$[25;30[$	$[30;35[$	$[35;40[$	Total
Effectifs	50	250	500	800	700	650	320	230	3500

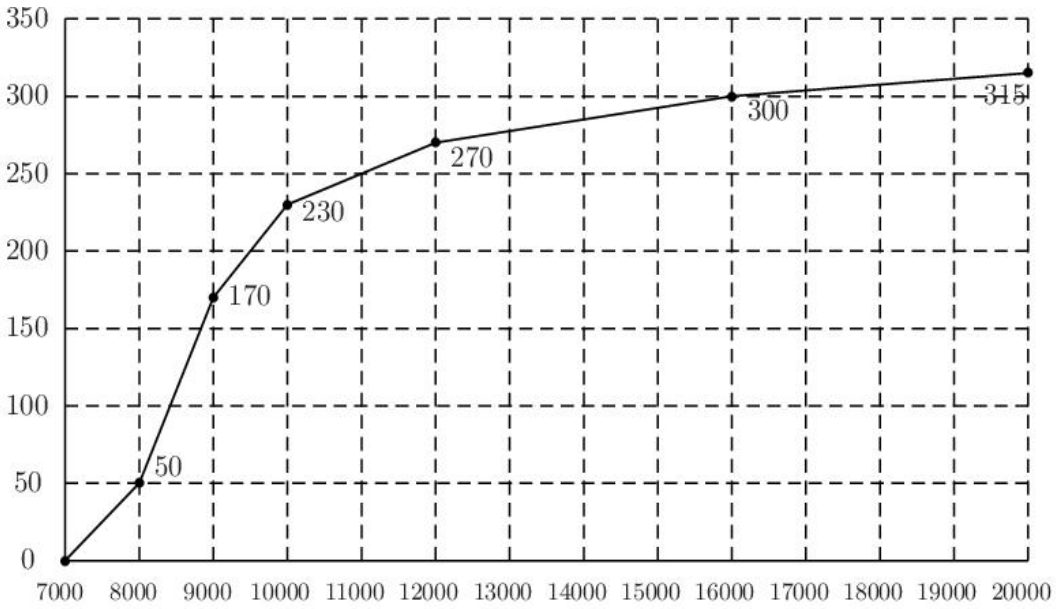
Chaque livraison est facturée 100 euros au client servi.
Le gérant de cet hypermarché, après une étude, a estimé à 5 euros le prix de revient du kilomètre par colis. Il envisage de modifier le tarif d'une livraison afin d'équilibrer le coût de ce service.

1. (a) Calculer pour chaque classe, en remplaçant celle-ci par son milieu, le nombre de kilomètres parcourus.
   (b) Vérifier que ce service est déficitaire. Quelle est approximativement la perte ?
2. (a) Dresser le tableau des fréquences cumulées croissantes.
   (b) Tracer la courbe des fréquences cumulées croissantes. Notons M_1 le point de la courbe dont l'ordonnée est 25%, M_2 celui dont l'ordonnée est 75%. L'abscisse Q_1 de M_1 est appelée le premier quartile et l'abscisse Q_3 de M_3 est appelée le troisième quartile. Trouver Q_1 et Q_3.
   (c) Calculer l'intervalle interquartile $[Q_1, Q_3]$ et l'interquartile $Q_3 - Q_1$.
3. Pour la suite, on prend $Q_1 \simeq 15$ km et $Q_3 \simeq 30$ km.
   (a) Recopier et compléter le tableau suivant, dans lequel on a réduit la série à trois classes $[0; Q_1[$, $[Q_1; Q_3[$, $[Q_3; 40[$.

Classes	Effectifs	kilomètres parcourus	Coût des livraisons	Coût moyen d'une livraison
$[0;15[$	800			
$[15;30[$	2150			
$[30;40[$	550			

(b) Quel conseil peut-on donner au gérant de cet hypermarché ?

Exercice 21 Les salaires nets des employés d'un garage automobile ont permis d'établir la courbe des effectifs cumulés croissants suivante :



1. Présenter le tableau où figureront entre autres
   - les classes
   - les effectifs
   - les effectifs cumulés croissants
   - les effectifs cumulés décroissants
2. Présenter l'histogramme de cette série.
3. Préciser l'étendue de cette série statistique.
4. Préciser la classe modale.
5. Calculer les trois quartiles.
6. Déterminer la moyenne arithmétique, puis l'écart-type de cette série :
   - directement,
   - par le changement de variable $u_i = \dfrac{x_i - 9500}{500}$.

Exercice 22 Un entrepôt stockant un certain produit envisage de modifier ses infra-structures. Le gérant de l'entrepôt effectue auparavant des observations sur les flux d'entrée et de sortie de ce produit pendant les deux dernières années.

Unités (×100)	Nombre de semaines
0 à moins de 10	1
10 à moins de 20	2
20 à moins de 30	3
30 à moins de 40	8
40 à moins de 50	25
50 à moins de 60	27
60 à moins de 70	20
70 à moins de 80	12
80 à moins de 90	6
Total	104

1. Dresser le tableau des effectifs cumulés croissants.
2. Tracer la courbe des effectifs cumulés croissants.
3. On suppose que les données sont régulièrement réparties. Trouver une valeur approchée de la médiane m par une lecture graphique puis par le calcul.
4. Déterminer les quartiles de la même façon.

Exercice 23 Une entreprise de conditionnement lance une étude sur la quantité de polystyrène utilisée pendant 200 jours pour emballer des matières fragiles. Les résultats obtenus sont les suivants

Classes (en kg)	Effectifs
$[55; 65[$	10
$[65; 75[$	23
$[75; 77[$	21
$[77; 79[$	28
$[79; 81[$	34
$[81; 83[$	28
$[83; 85[$	25
$[85; 95[$	23
$[95; 105[$	8

1. Déterminer la classe modale.
2. Calculer les trois quartiles et retrouver graphiquement, à l'aide de la courbe des effectifs cumulés croissants, ces valeurs.
3. On appelle x_i les centres de classes. On précise que
$$\sum_i n_i x_i = 15976 \text{ et } \sum_i n_i x_i^2 = 1288920$$
Calculer la valeur moyenne et l'écart-type à $0,01$ près de cette série statistique.

Exercice 24 Un chef d'entreprise s'intéresse à la répartition des salaires annuels exprimés en milliers d'euros de 100 employés d'une de ses succursales. Il récupère les données suivantes.

Classes des salaires	x_i	n_i	$n_i \nearrow$	$n_i \nearrow$ (%)	$n_i x_i$	$n_i x_i \nearrow$	$n_i x_i \nearrow$ (%)
$[10; 12[$	11	40					
$[12; 14[$	13	30					
$[14; 16[$	15	20					
$[16; 18[$	17	10					
Total	–		–	–		–	–

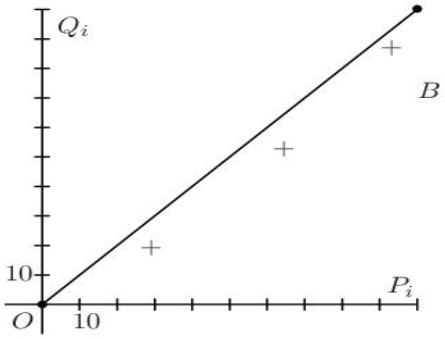
1. Compléter le tableau précédent
2. Quel pourcentage de la masse salariale totale se partagent les 50% des salariés les mieux payés de cette entreprise ?
3. Construire sur le graphique ci-joint la courbe de Lorenz associée à cette répartition de salaires (10 centimètres représentent 100%). Tracer sur ce même graphique la droite $\mathcal{D}$ d'équation $y = x$.

4. Calculer le coefficient de Gini sachant que l'aire comprise entre la courbe de Gini et $\mathcal{D}$ vaut $415,2$. Comment peut-on interpréter cette valeur ?

Exercice 25 On considère le salaire mensuel X de 200 salariés d'une petite entreprise, exprimé en milliers d'euros :

Classes	n_i	$n_i \nearrow$	$P_i\%$	$n_i x_i$	$Q_i\%$
$[1;1,5[$	52			65	17,45
$[1,5;2[$	71			124,25	50,8
$[2;2,5[$	57				85,23
$[2,5;3[$	20				100
Total		—	—		—

1. Compléter le tableau précédent.
2. Calculer le salaire médian et interpréter le résultat.
3. Expliquer comment on obtient le résultat $50,8$ dans la dernière colonne et interpréter ce résultat.
4. Calculer la médiale et interpréter le résultat.
5. À l'aide d'une calculatrice graphique, on a obtenu le graphique ci-dessous représentant les points de coordonnées (P_i, Q_i).



(a) Donner le nom de la courbe obtenue en reliant les différents points.

(b) Pourquoi peut-on dire qu'il y a une faible concentration ? Que peut-on dire alors des salaires ?

Exercice 26 On s'intéresse à la distribution des salaires mensuels dans une entreprise de confection. Les salariés de cette entreprise sont au nombre de 150. Les résultats obtenus sont les suivants :

Salaires en euros	Centres de classes x_i	n_i	$n_i \nearrow$	p_i (%)	$n_i x_i$	$n_i x_i \nearrow$	q_i (%)
$[1000;1200[$		80					
$[1200;1400[$		43					
$[1400;1600[$		17					
$[1600;1800[$		10					
Total	–		–	–		–	–

1. Compléter le tableau.
2. Construire la courbe de Lorenz.
3. Analyser la concentration de la masse salariale à l'aide du graphique.