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 :

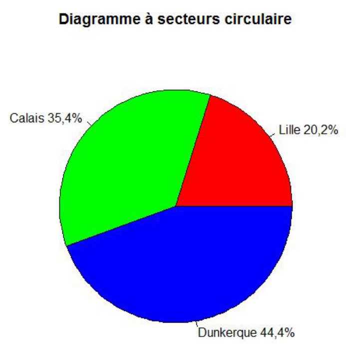
i	Région	Pourcentage	Fréquence f_i
1	Lille	20,2%	0,202
2	Calais	35,4%	0,354
3	Dunkerque	44,4%	0,444

Voici trois graphiques possibles pour cette série statistique.

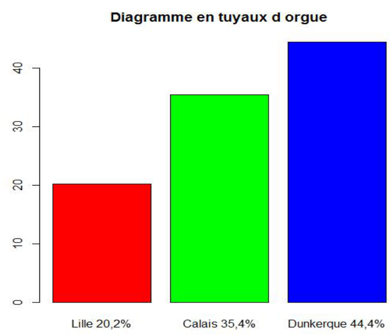
- **Diagramme à secteurs circulaire**
  Chaque classe correspond à un secteur dont l'angle est proportionnel à l'effectif donc à la fréquence de la classe. On a les angles suivants :

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

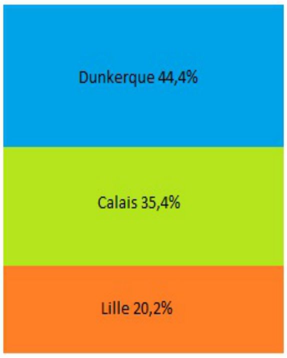
  ce qui donne le diagramme



- **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**

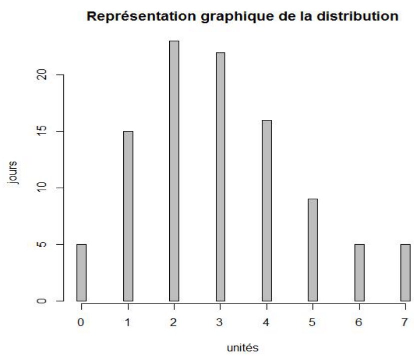


(b) Caractère quantitatif discret

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 :

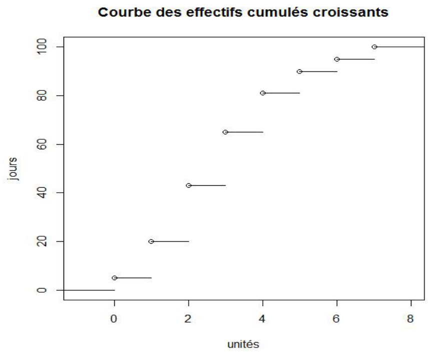
Nombre x_i d'unités demandées par jour	Nombre n_i de jours où l'on a vendu x_i	Effectifs cumulés croissants $n_i \nearrow$	Fréquence f_i	Fréquence cumulée croissante $f_i \nearrow$
0	5	5	0,05	0,05
1	15	20	0,15	0,20
2	23	43	0,23	0,43
3	22	65	0,22	0,65
4	16	81	0,16	0,81
5	9	90	0,09	0,9
6	5	95	0,05	0,95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

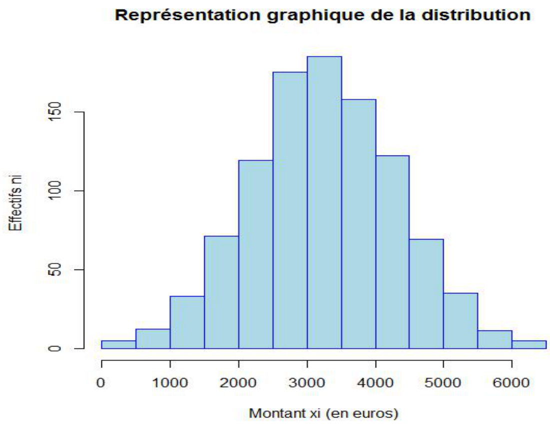
Exemple 1.2.4 On relève dans une banque à une date donnée les montants des économies de 1000 clients en euros. Les résultats obtenus sont les suivants :

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.

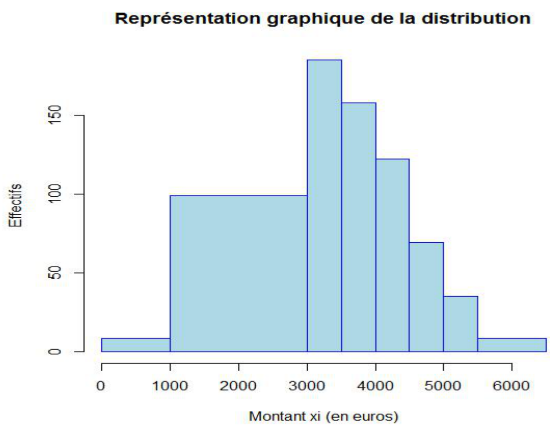
Diagramme des effectifs, histogramme.



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.



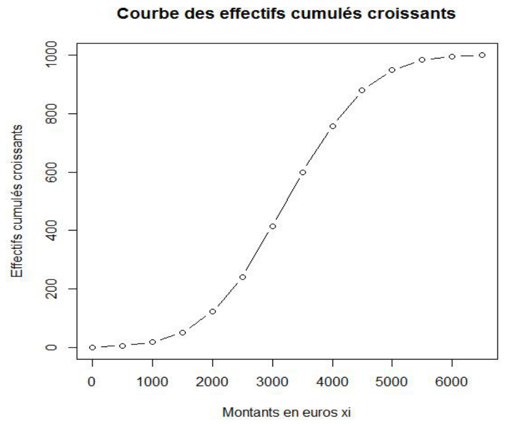
On remarque alors de manière générale que

- 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.

  $$\boxed{\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n} = \frac{1}{n}\sum_{i=1}^{p} n_ix_i}$$

  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 $c_i = \dfrac{a_i + b_i}{2}$.

  $$\boxed{\overline{x} = \frac{n_1c_1 + n_2c_2 + \ldots + n_pc_p}{n} = \frac{1}{n}\sum_{i=1}^{p} n_ic_i}$$

  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.

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

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 :

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

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.

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

Exemple 1.3.2 Soit la série

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

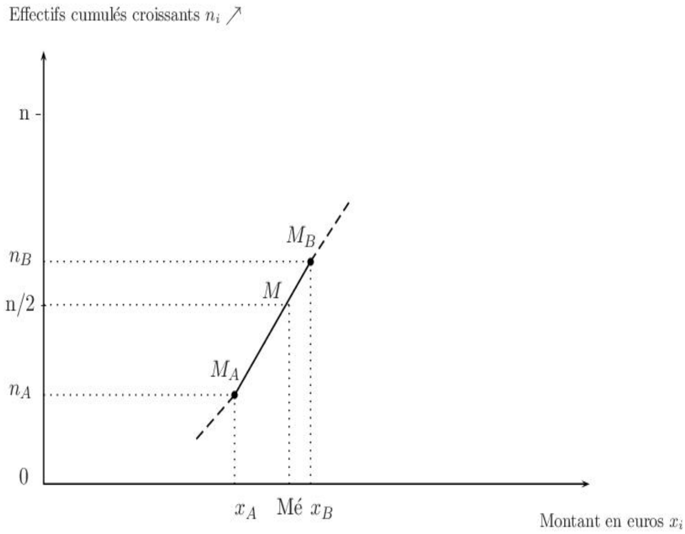
On vérifie que la série comporte $10 = 2 \times 5$ valeurs. Si la série est ordonnée, on peut affirmer que la médiane est la $1/2$ somme des 5-ième et 6-ième valeurs. Rangeons cette série en ordre de grandeur croissant :

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

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,

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

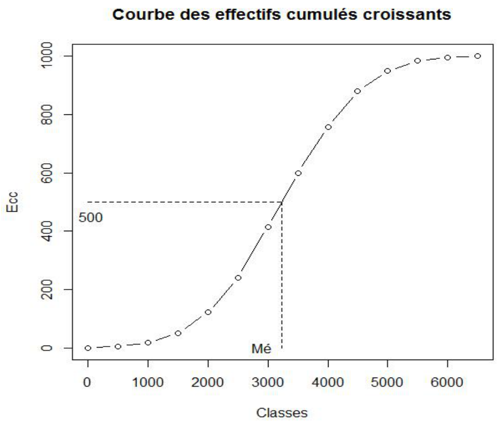
Considérons l'exemple 1.2.4.

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

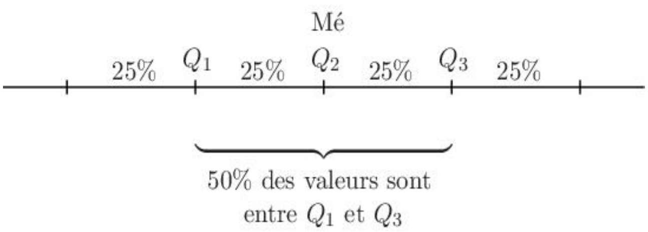
- Lorsque la population est répartie en classes $[a_i, b_i[$, la médiane peut donc être évaluée soit graphiquement soit par une interpolation affine (ou linéaire) à l'aide d'une courbe des effectifs cumulés en faisant l'hypothèse supplémentaire : **les éléments de la classe contenant la médiane sont 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).

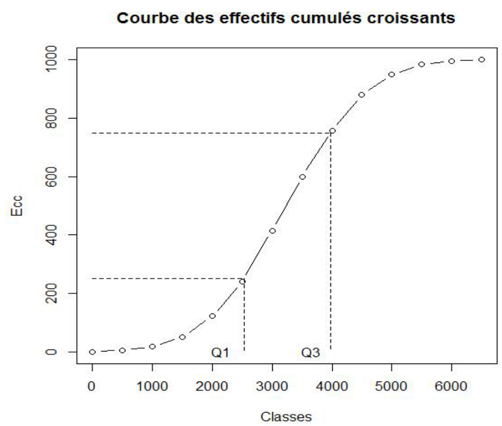
Considérons l'exemple 1.2.4.

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 $\overline{x}$*

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

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 :

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

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

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.

<u>Preuve</u> :

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

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 :

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

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

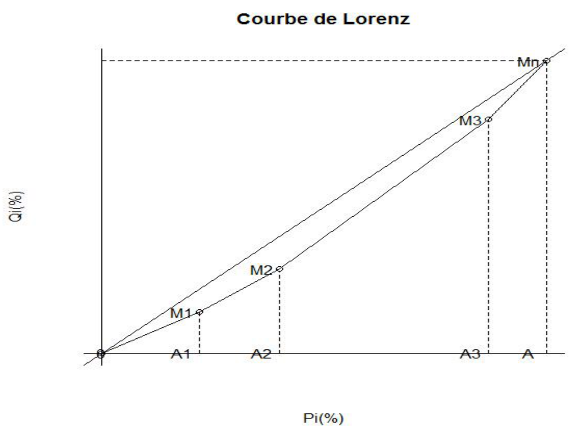
. $\overline{x} = \frac{M}{N} = \frac{550500}{100} = 5505$ qui correspond au salaire moyen dans l'entreprise.

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) :

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

 - 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

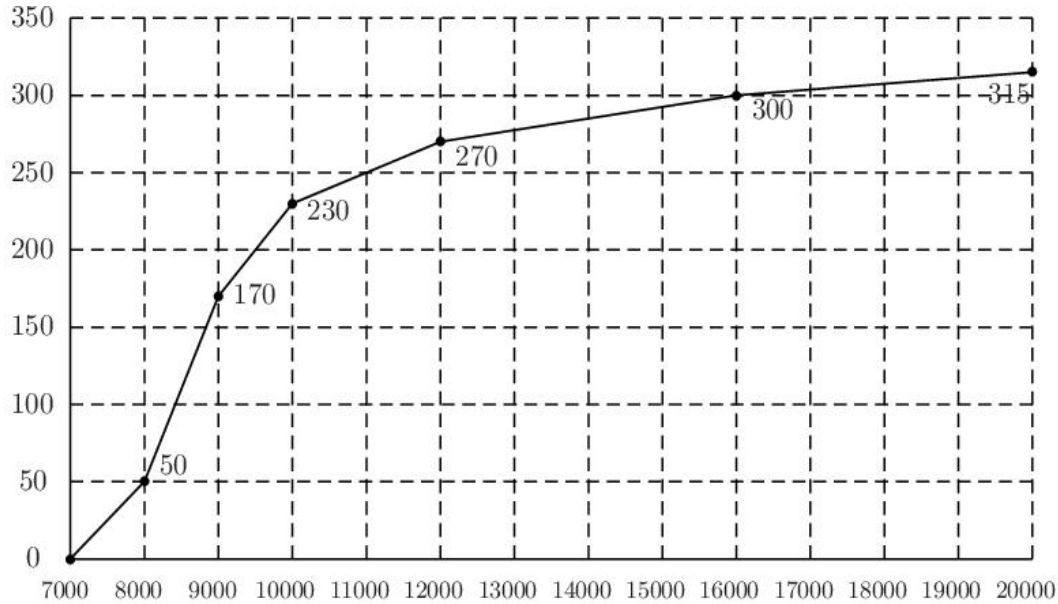
Chaque livraison est facturée 100 euros au client servi.
Le gérant de cet hypermarché, après une étude, a estimé à 5 euros le prix de revient du kilomètre par colis.
Il envisage de modifier le tarif d'une livraison afin d'équilibrer le coût de ce service.

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 ($\times 100$)	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	–		–	–		–	–

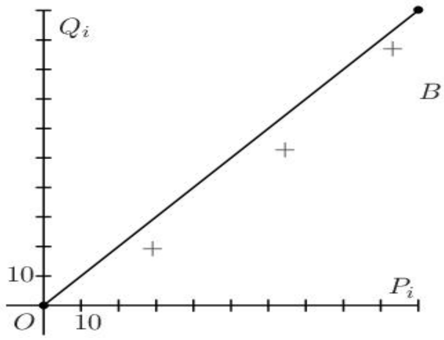
1. Compléter le tableau précédent
2. Quel pourcentage de la masse salariale totale se partagent les 50% des salariés les mieux payés de cette entreprise ?
3. Construire sur le graphique ci-joint la courbe de Lorenz associée à cette répartition de salaires (10 centimètres représentent 100%). Tracer sur 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_ix_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_ix_i	$n_ix_i \nearrow$	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.

Chapitre 2

Séries statistiques à deux variables

2.1 Introduction

Dans certains cas, il semble exister un lien entre les deux caractères d'une **série statistique à deux variables**, par exemple entre le poids et la taille d'une personne, les heures de révision pour un examen et la note obtenue, etc... Il est alors intéressant d'étudier simultanément deux caractères X et Y d'une même population E.

2.2 Tableaux de données. Nuages de points

On peut représenter les résultats sous forme de tableaux ou de graphiques.

2.2.1 Tableaux de données

On se donne plusieurs exemples ci-dessous impliquant indifféremment des caractères X et Y discrets ou continus.

Exemple 2.2.1 Au cours du troisième trimestre 2008, une marque automobile a lancé la commercialisation d'une nouvelle voiture avec deux motorisations distinctes de puissances respectives 138 chevaux DIN et 177 chevaux DIN. On dispose des quantités de voitures vendues par zones :

Zones	Nombre d'unités 138 CV vendues x_i	Nombre d'unités 177 CV vendues y_i
1	400	240
2	200	120
3	600	300
4	300	150
5	300	150
6	600	270

Exemple 2.2.2 Un même produit est vendu conditionné sous différentes formes et différents volumes. Le tableau suivant indique pour chaque type d'emballage le volume x_i et le prix y_i du produit.

x_i en cm^3	100	150	200	300	500	600	700	800	900	1000
y_i en euros	7	8	9,5	13	20	23	25	28,3	30,5	34

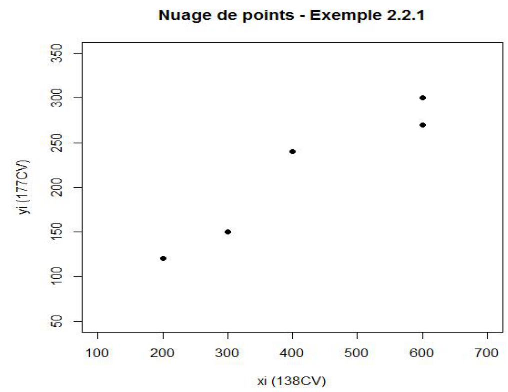
Exemple 2.2.3 Les chiffres d'affaires trimestriels d'une entreprise ont été pour les douze derniers trimestres :

Rang du trimestre x_i	1	2	3	4	5	6	7	8	9	10	11	12
Chiffre d'affaires (en milliers d'euros) y_i	300	450	130	200	280	410	200	250	320	500	210	250

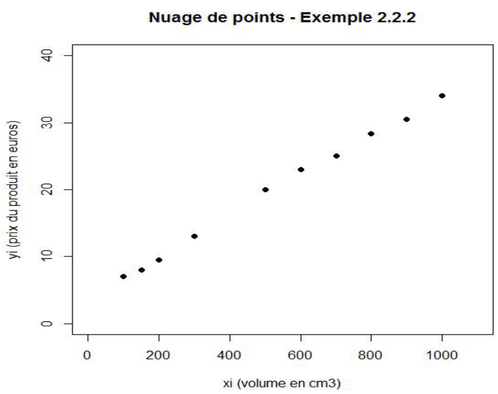
2.2.2 Nuages de points

Le plan étant muni d'un repère orthogonal, on peut associer au couple (x_i, y_i) de la série statistique double le point M_i de coordonnées x_i et y_i.
L'ensemble des points M_i obtenus constitue le **nuage de points** représentant la série statistique.

Dans l'exemple 2.2.1, on obtient le nuage ci-dessous :



Dans l'exemple 2.2.2, on obtient le nuage suivant :



Le nuage étant dessiné, on peut essayer de trouver une fonction f telle que la courbe d'équation $y = f(x)$ passe ≪ le plus près possible ≫ des points du nuage. C'est le problème de **l'ajustement**.
Lorsqu'il sera possible de tracer une droite D au voisinage des points, on parlera d'ajustement linéaire. Si l'ajustement linéaire ne convient pas, on peut penser à approcher le nuage à l'aide d'une parabole, d'une hyperbole, d'une fonction exponentielle...

2.3 Calcul des paramètres de position et de dispersion

Comme dans le chapitre sur les séries statistiques à une variable, il est possible de déterminer pour les séries statistiques à deux variables la moyenne arithmétique, la variance et tous les autres paramètres de position et de dispersion de chaque variable prise séparément. Il suffit pour cela de déterminer les distributions marginales des variables X et Y.

Exemple 2.3.1 On a réparti 1000 individus d'une population suivant

- le nombre x d'arrêts de travail suite à des accidents par an,
- l'âge y en années

de ces individus. On obtient les résultats synthétisés dans le tableau ci-dessous.

$x_i \backslash y_j$	$[20;30[$	$[30;40[$	$[40;50[$	$[50;60[$	$n_{i.}$
0	20	150	22	0	192
1	99	102	180	12	393
2	60	51	150	30	291
3	18	0	50	32	100
4	3	0	10	11	24
$n_{.j}$	200	303	412	85	1000

Ce tableau à double entrée est appelé **tableau de contingence** ou **d'effectifs** n_{ij}.

L'**effectif marginal** de la variable X est défini par :

$$\boxed{n_{i.} = \sum_{j=1}^{q} n_{ij} \text{ pour } i \in \{1, 2, \ldots, p\}}$$

où p et q sont respectivement les nombres de modalités (ou de catégories) de X et Y.

L'effectif marginal de la variable Y est défini par :

$$\boxed{n_{.j} = \sum_{i=1}^{p} n_{ij} \text{ pour } j \in \{1, 2, \ldots, q\}}$$

L'effectif total est donné par

$$\boxed{n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j}}$$

On peut définir les **fréquences marginales**

$$\boxed{f_{i.} = \frac{n_{i.}}{n} \text{ et } f_{.j} = \frac{n_{.j}}{n}}$$

ainsi que la fréquence

$$\boxed{f_{ij} = \frac{n_{ij}}{n}}$$

pour $i \in \{1, 2, \ldots, p\}$ et $j \in \{1, 2, \ldots, q\}$.

Dans le cadre de l'exemple 2.3.1, on remarquera au préalable que les nombres de modalités pour X et Y sont respectivement $p = 5$ et $q = 4$. Les distributions marginales de X et Y en termes d'effectifs et de fréquences sont respectivement :

x_i	$n_{i.}$	$f_{i.}$
0	192	0,192
1	393	0,393
2	291	0,291
3	100	0,100
4	24	0,024
Total	1000	1

Classe	y_j	$n_{.j}$	$f_{.j}$
$[20; 30[$	25	200	0,200
$[30; 40[$	35	303	0,303
$[40; 50[$	45	412	0,412
$[50; 60[$	55	85	0,085
Total	–	1000	1

2.3.1 Le point moyen

Supposons qu'on souhaite calculer les moyennes arithmétiques de X et Y, il suffit pour cela d'utiliser les formules déja étudiées dans le chapitre 1 à savoir

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i} \qquad \boxed{\overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j}$$

Afin de calculer les sommes $\sum_i n_{i.}x_i$ et $\sum_j n_{.j}y_j$ on utilise le tableau des effectifs de X et Y qu'on complète avec une colonne supplémentaire, $n_{i.}x_i$ pour la variable X, $n_{.j}y_j$ pour la variable Y.
Considérons l'exemple 2.3.1 :

x_i	$n_{i.}$	$n_{i.}x_i$
0	192	0
1	393	393
2	291	582
3	100	300
4	24	96
Total	1000	1371

Classe	y_j	$n_{.j}$	$n_{.j}y_j$
$[20; 30[$	25	200	5000
$[30; 40[$	35	303	10605
$[40; 50[$	45	412	18540
$[50; 60[$	55	85	4675
Total	–	1000	38820

On en déduit que $\overline{x} = \frac{1371}{1000} = 1,371$ et $\overline{y} = \frac{38820}{1000} = 38,82$.

Lorsqu'on pense pouvoir réaliser un ajustement linéaire (ou affine) d'un nuage, il semble intéressant, avant de tracer la droite, de placer le point dont l'abscisse est la moyenne des abscisses x_i et l'ordonnée la moyenne des ordonnées y_j.

Définition 2.3.1 *On appelle **point moyen** G d'un nuage de n points M_i de coordonnées (x_i, y_i) le point de coordonnées :*

$$\boxed{(x_G, y_G) = (\overline{x}, \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i, \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j\right)}$$

On vérifie ainsi que le point moyen dans l'exemple 2.3.1 est $G(1,371; 38,82)$

2.3.2 Les variances

Comme on l'a vu précédemment, le calcul de la variance suivi de celui de l'écart-type nous permet de mesurer la dispersion des valeurs de la série statistique autour de la moyenne arithmétique. On peut calculer indépendamment les variances des deux variables X et Y à l'aide de formules déja utilisées dans le chapitre 1 dans le cadre des séries statistiques à deux variables :

$$\boxed{V(X) = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2 - \overline{x}^2} \qquad \boxed{V(Y) = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2 - \overline{y}^2}$$

où p et q sont respectivement le nombre de modalités (ou de catégories) de X et Y.
On en déduit que

$$\boxed{\sigma(X) = \sqrt{V(X)}} \qquad \boxed{\sigma(Y) = \sqrt{V(Y)}}$$

Appliquons ces formules dans le cadre de l'exemple 2.3.1. Une colonne supplémentaire est ajoutée au tableau, $n_{i.}x_i^2$ pour la variable X, $n_{.j}y_j^2$ pour la variable Y.

x_i	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$
0	192	0	0
1	393	393	393
2	291	582	1164
3	100	300	900
4	24	96	384
Total	1000	1371	2841

Classe	y_j	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
$[20;30[$	25	200	5000	125000
$[30;40[$	35	303	10605	371175
$[40;50[$	45	412	18540	834300
$[50;60[$	55	85	4675	257125
Total	–	1000	38820	1587600

Par conséquent,

. $V(X) = \dfrac{2841}{1000} - (1,371)^2 \simeq 0,961$ et $\sigma(X) = \sqrt{0,961} \simeq 0,98$

. $V(Y) = \dfrac{1587600}{1000} - (38,82)^2 \simeq 80,61$ et $\sigma(Y) = \sqrt{80,61} \simeq 8,98$

2.4 Vocabulaire, définitions

2.4.1 La covariance

Il est possible comme lors de l'étude sur les séries à une variable de définir une variance sur les deux variables simultanément, c'est la covariance.

Définition 2.4.1 *La **covariance** d'une série statistique à deux variables X et Y est donnée par la formule*

$$\boxed{Cov(X,Y) = \sigma_{XY} = \frac{\sum_{i,j} n_{ij}(x_i - \overline{x})(y_j - \overline{y})}{n}}$$

Comme pour la variance, il existe une écriture de la covariance plus adaptée au calcul, obtenue simplement en développant la formule précédente.

Propriété 2.4.1

$$\boxed{Cov(X,Y) = \sigma_{XY} = \frac{1}{n}\sum_{i,j} n_{ij}x_iy_j - \overline{x}.\overline{y}}$$

<u>Preuve</u> :

$$\begin{aligned}
Cov(X,Y) &= \frac{\sum_{i,j} n_{ij}(x_i-\overline{x})(y_j-\overline{y})}{n} = \frac{1}{n}\sum_{i,j} n_{ij}(x_iy_j - \overline{x}y_j - \overline{y}x_i + \overline{x}.\overline{y}) \\
&= \frac{1}{n}\left(\sum_{i,j} n_{ij}x_iy_j - \overline{x}\sum_{i,j} n_{ij}y_j - \overline{y}\sum_{i,j} n_{ij}x_i + \overline{x}.\overline{y}\sum_{i,j} n_{ij}\right) \\
&= \frac{1}{n}\sum_{i,j} n_{ij}x_iy_j - 2\overline{x}.\overline{y} + \overline{x}.\overline{y} = \frac{1}{n}\sum_{i,j} n_{ij}x_iy_j - \overline{x}.\overline{y}.
\end{aligned}$$

Remarque 2.4.1

- $Cov(X,Y) = \overline{xy} - \overline{x}.\overline{y}$
- $Cov(X,X) = V(X)$ (la covariance est en quelque sorte le « dédoublement » de la variance)

Considérons l'exemple 2.3.1. On doit tout d'abord déterminer $\sum_{i,j} n_{ij}x_iy_j$, tâche qui peut être réalisée en utilisant le tableau de contingence :

x_i \ y_j	$[20;30[$	$[30;40[$	$[40;50[$	$[50;60[$	$n_{i.}$
0	20 (0)	150 (0)	22 (0)	0 (0)	192
1	99 (2475)	102 (3570)	180 (8100)	12 (660)	393
2	60 (3000)	51 (3570)	150 (13500)	30 (3300)	291
3	18 (1350)	0 (0)	50 (6750)	32 (5280)	100
4	3 (300)	0 (0)	10 (1800)	11 (2420)	24
$n_{.j}$	200	303	412	85	1000

Par exemple, $n_{11}x_1y_1 = 20 \times 0 \times 25 = 0$, $n_{23}x_2y_3 = 180 \times 1 \times 45 = 8100$,...
On obtient $\sum_{i,j} n_{ij}x_iy_j = 56075$ ce qui permet de déterminer la covariance entre X et Y :

$$Cov(X,Y) = \frac{56075}{1000} - (1,371 \times 38,82) \simeq 2,853$$

2.4.2 Le coefficient de corrélation linéaire

Pour mesurer l'intensité de la relation linéaire entre X et Y (autrement que par interprétation graphique du nuage de points), on définit le coefficient de corrélation linéaire $r(X, Y)$.

Définition 2.4.2 *Le coefficient de corrélation linéaire d'une série statistique double de variables X et Y est le nombre r défini par*

$$\boxed{r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}}$$

Propriété 2.4.2 *Le coefficient de régression vérifie :*

$$\boxed{-1 \leq r(X,Y) \leq 1}$$

Considérons l'exemple 2.3.1

$$r(X,Y) = \frac{2,853}{0,98 \times 8,98} \simeq 0,324$$

Commentaires :
- $r = 1$ ou $r = -1$ si et seulement si les points $M_i(x_i, y_i)$ sont alignés.
- Si r est voisin de 1 ou -1, la corrélation linéaire entre X et Y est très forte.
- Si r est proche de 0, il n'existe pas de corrélation linéaire entre X et Y. Les variables X et Y sont linéairement indépendantes ; il peut néanmoins exister une autre relation fonctionnelle entre X et Y, par exemple $Y = aX^2 + bX + c, \ldots$
- On peut présumer d'une corrélation linéaire pour $|r| \geq 0,866$ de sorte que la présomption de corrélation linéaire commence à partir de la valeur $\boxed{|r| \simeq 0,87}$

2.5 Ajustement linéaire (ou affine)

Avant de préciser ces notions par le calcul, il est bon de comprendre comment se pose le problème de la corrélation linéaire. Le problème consiste à déterminer dans quelle mesure les deux variables X et Y sont liées (c'est-à-dire dépendent l'une de l'autre). Par exemple, on peut intuitivement penser que la taille et le poids des individus d'une population sont liés, que par contre il est plus improbable que la taille et le revenu mensuel des habitants d'un pays donné soient liés. Si on arrive, à l'aide des données dont on dispose, à déterminer s'il existe une certaine fonction f telle que $\forall k \in \{1, 2, \ldots, n\}$, $y_k = f(x_k)$ on pourra répondre avec plus de précision à cette idée de lien entre X et Y.

Lorsque le nuage de points est nettement longiligne, les points étant disposés suivant une direction privilégiée, la corrélation est dite affine. Il est utile alors, dans un but d'extrapolation, de déterminer une droite rendant compte le mieux possible de la tendance observée. On dit qu'on effectue un **ajustement affine**.
On distingue deux types d'ajustement : les ajustements graphiques et les ajustements analytiques nécessitant un calcul spécifique

2.5.1 Ajustement graphique

On peut utiliser trois techniques :

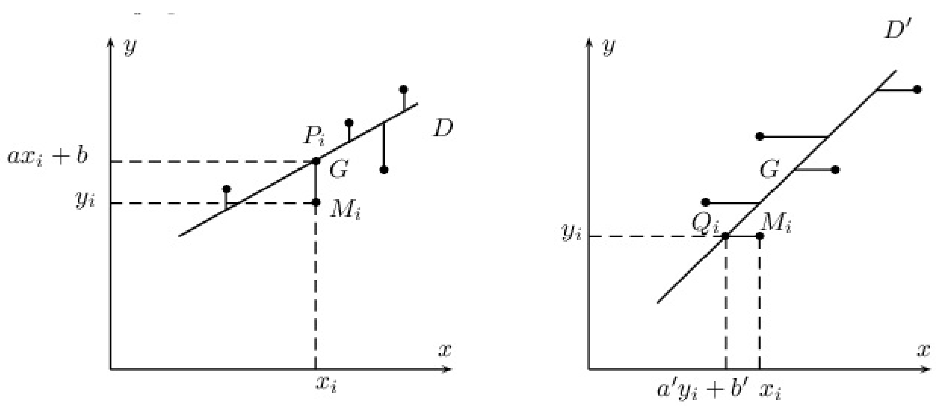
- Ajustement direct à la règle
  On utilise une règle transparente qu'on dispose de façon à l'ajuster le mieux possible suivant la direction privilégiée constatée et on s'efforce d'équilibrer le nombre de points situés de part et d'autre.

– Utilisation du point moyen
  On montre par le calcul que, pour obtenir le meilleur ajustement affine, il convient de prendre une droite passant par le point moyen G.
– Fractionnement du nuage, méthode de Mayer.

2.5.2 Ajustement analytique - Méthode des moindres carrés

On considère une série statistique à deux variables représentée par un nuage justifiant un ajustement affine. La méthode d'ajustement linéaire des moindres carrés conduit à obtenir 2 droites de régression $D_{Y/X}$ et $D_{X/Y}$ concourantes au point moyen $G(\overline{x}, \overline{y})$ mais il est nécessaire de remarquer que ces 2 droites existent quel que soit le lien existant entre X et Y (même s'il n'y a aucune dépendance entre X et Y, on peut toujours tracer la droite telle que la somme des carrés des distances des points M_i du nuage aux points de la droite de mêmes abscisses soit minimale, ce minimum pouvant d'ailleurs être grand).
Le problème peut donc être posé ainsi : « à l'aide des deux droites de régression $D_{Y/X}$ et $D_{X/Y}$, quel critère numérique permet de dire si X et Y sont plus ou moins dépendantes l'une de l'autre par l'intermédiaire d'une fonction f affine ? »

On considère les deux graphes suivants :



Définition 2.5.1 *On appelle* ***droite d'ajustement par la méthode des moindres carrés de*** Y ***par rapport à*** X *la droite*
– *passant par* G
– *qui minimise la somme des carrés des écarts* M_iP_i *entre les ordonnées des points du nuage et les ordonnées des points de la droite ayant même abscisse.*

On dit aussi ***droite de régression de*** Y ***en*** X

Propriété 2.5.1 *La droite des moindres carrés de* Y *par rapport à* X *est notée* $D_{Y/X}$*. Son équation est*

$$\boxed{y = ax + b}$$

avec

$$\boxed{\begin{cases} a = \dfrac{Cov(X,Y)}{V(X)} & \text{coefficient directeur de } D_{Y/X} \\ b = \overline{y} - a\overline{x} & \end{cases}}$$

Définition 2.5.2 *On appelle* ***droite d'ajustement par la méthode des moindres carrés de*** X ***par rapport à*** Y *la droite*
– *passant par* G
– *qui minimise la somme des carrés des écarts* Q_iM_i *entre les abscisses des points du nuage et les abscisses des points de la droite ayant même ordonnée.*

On dit aussi ***droite de régression de*** X ***en*** Y

Propriété 2.5.2 *La droite des moindres carrés de* X *par rapport à* Y *est notée* $D_{X/Y}$*. Son équation est*

$$\boxed{x = a'y + b'}$$

avec

$$\boxed{\begin{cases} a' = \dfrac{Cov(X,Y)}{V(Y)} & \textit{coefficient directeur de } D_{X/Y} \\ b' = \overline{x} - a'\overline{y} & \end{cases}}$$

Remarque 2.5.1

- Les deux droites $D_{Y/X}$ et $D_{X/Y}$ passent par le point moyen du nuage.
- Les points du nuage sont alignés si et seulement si les droites $D_{Y/X}$ et $D_{X/Y}$ sont confondues.

Propriété 2.5.3

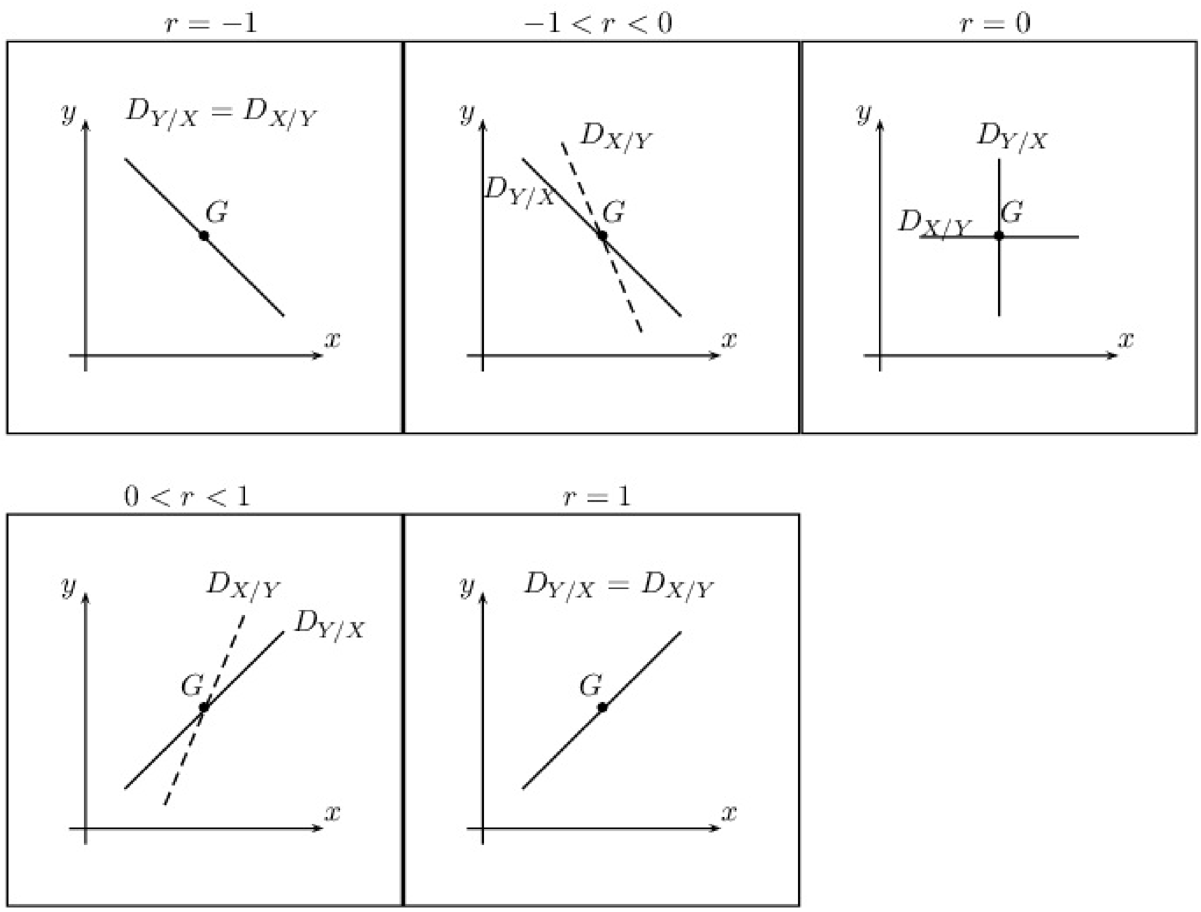
$$\boxed{D_{Y/X} = D_{X/Y} \Leftrightarrow r^2 = 1}$$

<u>*Preuve*</u> :

- L'équation de $D_{Y/X}$ peut se réécrire sous la forme $y = ax + (\overline{y} - a\overline{x}) \Leftrightarrow a(x - \overline{x}) = y - \overline{y}$
- De même, l'équation $D_{X/Y}$ peut s'écrire $x = a'y + (\overline{x} - a'\overline{y}) \Leftrightarrow x - \overline{x} = a'(y - \overline{y})$

Donc, $D_{Y/X} = D_{X/Y} \Leftrightarrow a.a' = 1 \Leftrightarrow \dfrac{Cov(X,Y)^2}{\sigma_X^2\sigma_Y^2} = 1$. Or $r = \dfrac{Cov(X,Y)}{\sigma_X\sigma_Y}$ donc on obtient le résultat voulu.

On a finalement les 5 cas graphiques suivants :



L'ajustement linéaire que nous venons d'étudier avait pour objet de remplacer le nuage de points $M_i(x_i, y_i)$ par une droite D d'équation $y = ax + b$ (ou $x = a'y + b'$), résumant en partie la liaison entre X et Y. Cette droite D permet *à partir des valeurs x_i observées d'obtenir une valeur ajustée $\hat{y}_i$ pour y_i.*

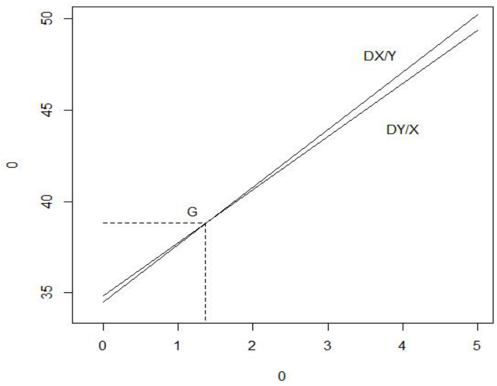
Considérons l'exemple 2.3.1 et déterminons les deux droites de régression $D_{Y/X}$ et $D_{X/Y}$:

- On a $a = \dfrac{Cov(X,Y)}{V(X)} = \dfrac{2,853}{0,98} \simeq 2,911$ et $b = \overline{y} - a\overline{x} \simeq 38,82 - 2,911 \times 1,371 \simeq 34,829$

  On déduit de ces résultats l'équation de $D_{Y/X}$: $y = 2,911x + 34,829$
- On a $a' = \dfrac{Cov(X,Y)}{V(Y)} = \dfrac{2,853}{8,98} \simeq 0,318$ et $b' = \overline{x} - a'\overline{y} \simeq 1,371 - 0,318 \times 38,82 \simeq -10,974$

  On déduit de ces résultats l'équation de $D_{Y/X}$: $x = 0,318y - 10,974$

Représentons graphiquement ces deux droites et comparons les résultats aux graphiques précédents.



On reconnaît immédiatement le cas $0 < r < 1$ (on rappelle d'ailleurs que $r \simeq 0,324$). Comme le coefficient de corrélation linéaire ne vérifie pas la condition $|r| > 0,87$, cela implique qu'il n'y ait pas de réelle corrélation linéaire entre X et Y. Il n'existe pas, par le biais des données disponibles, de relation linéaire entre l'âge d'un individu et le nombre d'arrêts de travail subi annuellement par cette personne. Il n'est pas donc possible d'extrapoler des résultats, c'est-à-dire de traiter des valeurs non contenues dans la série initiale et de répondre à des questions du type

- « Quelle estimation portant sur le nombre d'arrêts de travail peut-on donner si la personne est agée de 47 ans ? »
- « Quelle estimation de l'âge d'une personne peut on donner lorsque le nombre d'arrêts de travail qu'elle a subi est égal à 7 ? »

En supposant que le coefficient de corrélation vérifie l'inégalité $|r| > 0,87$, il est possible alors de répondre à de telles interrogations.

2.6 Exercices

Exercice 27 Après un examen de mathématiques, on a réparti 100 étudiants de 1ère année en économie selon le nombre x de minutes qu'ils ont passé lors des révisions la veille et la note y sur 20 qu'ils ont obtenue. Les résultats sont les suivants :

$x \backslash y$	$[0;4[$	$[4;8[$	$[8;12[$	$[12;16[$	$[16;20[$	$n_{i.}$
$[0;60[$	16	6	2	1	0	
$[60;120[$	5	10	5	3	0	
$[120;180[$	2	7	17	10	4	
$[180;240[$	1	1	1	5	4	
$n_{.j}$						

1. Compléter le tableau.
2. Combien d'étudiants ont entre 8 et 20 à leur examen ?
3. Quel est le pourcentage d'étudiants ayant eu une note inférieure à 8 et ayant révisé moins de 2 heures ?
4. Quel est le pourcentage d'étudiants ayant eu une note supérieure à 8 et ayant révisé plus de 2 heures ?
5. Présenter les histogrammes des séries x et y.
6. Tracer la courbe des effectifs cumulés croissants pour les deux séries statistiques.

Exercice 28 L'étude des hauteurs barométriques en fonction de l'altitude a permis d'établir le tableau suivant avec x l'altitude en km et y la hauteur barométrique en cm de mercure :

x_i	0	1	2	4	6	10
y_i	76	67	59	46	35	20

1. Représenter graphiquement la série statistique par un nuage de points.
2. Déterminer par la méthode des moindres carrés l'équation de la droite de régression de y en x. Tracer cette droite sur le graphique.
3. Sachant que la hauteur barométrique en un lieu est de 40 cm, calculer l'altitude.

Exercice 29 Le tableau ci-dessous donne l'évolution de la consommation de médicaments des ménages en France. x représente l'année, y le montant en milliards d'euros de la consommation de médicaments.

x_i	1970	1975	1980	1985	1990	1995	2002	2006
y_i	0,352	0,6660	1,073	2,026	3,369	6,420	9,592	10,89

1. Déterminer les coordonnées du point moyen $G(\overline{x}, \overline{y})$.
2. Calculer le coefficient de corrélation linéaire à $0,001$ près. Interpréter ce résultat.
3. Déterminer par la méthode des moindres carrés une équation de la droite de régression D de Y en X.
4. En supposant que l'évolution se poursuive de la même façon dans les années suivantes, donner une estimation de la consommation de médicaments des ménages en France en 2010.
   En quelle année la consommation dépassera-t-elle 12,5 milliards d'euros ?

Exercice 30 Une bibliothèque municipale établit le bilan de ses activités pour les 4 dernières années. Le tableau suivant donne en milliers pour chaque année,
- l'augmentation du nombre des prêts de livres x_i
- le nombre des nouveaux lecteurs inscrits y_i
- le nombre des nouveautés achetées z_i

	2003	2004	2005	2006
x_i	3	7	1	5
y_i	0,3	1,2	0,1	0,8
z_i	0,9	3,2	2,1	2,8

1. Calculer le coefficient de corrélation linéaire des séries (x_i) et (y_i).
2. Calculer le coefficient de corrélation linéaire des séries (x_i) et (z_i).
3. Déterminer quel élément (y_i ou z_i) est le plus susceptible d'avoir influencé l'augmentation des prêts de livres x_i.
4. Déterminer la droite d'ajustement par la méthode des moindres carrés $D_{Y/X}$.
5. En déduire une estimation du nombre de nouveaux lecteurs inscrits y_i, si l'on pense que l'augmentation du nombre des prêts sera de 9000 en 2007.

Exercice 31 50 étudiants d'une promotion ont effectué deux contrôles, l'un en méthodes quantitatives dont les notes sont x_i, l'autre en marketing dont les notes sont y_j.
On obtient la série statistique double donnée par le tableau ci-dessous :

y_j \ x_i	2	8	12	18
6	8	1	1	0
9	1	10	2	0
11	1	2	14	1
14	0	0	2	7

1. En précisant dans un tableau complet à double entrée les détails des calculs, déterminer une équation de régression de Y en X.
2. Déterminer une équation de la droite de régression de X en Y.
3. Calculer le coefficient de corrélation linéaire.
4. Si un étudiant obtient 15 au devoir de marketing, quelle note peut-on prévoir en méthodes quantitatives ?
5. Si un étudiant obtient 4 au devoir de méthodes quantitatives, quelle note peut-on prévoir en marketing ?

Exercice 32 Dans le cadre d'une enquête médicale, une étude est réalisée auprès de 60 patients concernant la taille et le poids de chacun. Les résultats sont donnés dans le tableau ci-dessous. On exprimera le poids x en kilogrammes et la taille y en centimètres. On arrondira les résultats des questions suivantes à 10^{-3} près.

y (cm) \ x (kg)	$[50;60[$	$[60;70[$	$[70;80[$	$[80;90[$	$[90;100[$
$[150;160[$	10	2	1	0	0
$[160;170[$	1	12	3	1	0
$[170;180[$	1	1	13	2	0
$[180;190[$	0	0	2	6	1
$[190;200[$	0	0	0	1	3

1. Calculer le pourcentage de personnes pesant moins de 80 kilos et mesurant plus de 180 centimètres.
2. Calculer la proportion de personnes pesant entre 60 et 90 kilos.

   Dans la suite du problème, on utilisera les résultats suivants :

y (cm) \ x (kg)	$[50;60[$	$[60;70[$	$[70;80[$	$[80;90[$	$[90;100[$	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
$[150;160[$	10	2	1	0	0	13	2015	312325
$[160;170[$	1	12	3	1	0	17	2805	462825
$[170;180[$	1	1	13	2	0	17	**2975**	520625
$[180;190[$	0	0	2	6	1	9	1665	308025
$[190;200[$	0	0	0	1	3	4	780	152100
$n_{i.}$	12	15	19	10	4	60	10240	1755900
$n_{i.}x_i$	660	975	1425	850	380	4290		
$n_{i.}x_i^2$	36300	63375	**106875**	72250	36100	314900		

y (cm) \ x (kg)	$[50;60[$	$[60;70[$	$[70;80[$	$[80;90[$	$[90;100[$	
$[150;160[$	85250	20150	11625	0	0	117025
$[160;170[$	9075	128700	37125	14025	0	188925
$[170;180[$	9625	11375	170625	29750	0	221375
$[180;190[$	0	0	27750	**94350**	17575	139675
$[190;200[$	0	0	0	16575	55575	72150
	103950	160225	247125	154700	73150	739150

3. Retrouver (en expliquant vos calculs) les trois valeurs encadrées dans les deux tableaux précédents.
4. Calculer les moyennes arithmétiques $\overline{x}$ et $\overline{y}$ des poids et tailles.
5. Calculer les variances $V(x)$ et $V(y)$. Déterminer alors les écart-types $\sigma(x)$ et $\sigma(y)$.
6. Calculer la covariance et en déduire le coefficient de corrélation linéaire entre le poids et la taille.
7. Un ajustement linéaire est-il justifié. Peut-on prévoir alors à l'aide de l'échantillon donné la taille d'une personne à partir de son poids ?

Chapitre 3

Le dénombrement

3.1 Notations

3.1.1 Préliminaires

- Soit E un ensemble fini, le **cardinal** de E, noté $Card(E)$ ou $|E|$, désigne le nombre de ses éléments.
- $\mathcal{P}(E)$ désigne l'ensemble des **parties** de E (y compris l'ensemble E lui-même et l'ensemble vide noté $\emptyset$).

  Exemple 3.1.1 Si on se donne $E = \{0, 1, 2\}$, l'ensemble des parties de E est donné par
  $$\mathcal{P}(E) = \{\emptyset; \{0\}; \{1\}; \{2\}; \{0, 1\}; \{0, 2\}; \{1, 2\}; E\}$$
- Soient A et B deux parties de E alors
  - $\boxed{A \cap B = \{x \in E/x \in A \text{ et } x \in B\}}$ définit l'**intersection** de A et de B,
  - $\boxed{A \cup B = \{x \in E/x \in A \text{ ou } x \in B\}}$ définit la **réunion** de A et de B,
  - $\boxed{\overline{A} = \{x \in E/x \notin A\}}$ définit le **complémentaire** de A dans E,
  - $\boxed{A - B = \{x \in E/x \in A \text{ et } x \notin B\} = A \cap \overline{B}}$ définit "A privé de B" (on écrit également A/B),
  - $A \triangle B = (A - B) \cup (B - A) = (A \cap \overline{B}) \cup (\overline{A} \cap B)$ définit la **différence symétrique** de A et de B. On a par conséquent $\boxed{A \triangle B = \{x \in E/(x \in A \text{ et } x \notin B) \text{ ou } (x \notin A \text{ et } x \in B)\}}$.

  Exemple 3.1.2 Si on se donne les ensembles $E = \{0, 1, 2, 3, 4, 5, 6\}$, $A = \{0, 1, 2\}$, $B = \{2, 3, 4\}$ alors
  - $A \cap B = \{2\}$,
  - $A \cup B = \{0, 1, 2, 3, 4\}$,
  - $\overline{A} = \{3, 4, 5, 6\}$,
  - $\overline{B} = \{0, 1, 5, 6\}$,
  - $A - B = A \cap \overline{B} = \{0, 1\}$,
  - $B - A = B \cap \overline{A} = \{3, 4\}$,
  - $A \triangle B = \{0, 1, 3, 4\}$.

3.1.2 Ensemble produit

Soient deux ensembles finis E et F.

1. On appelle **ensemble produit** ou **produit cartésien** de E par F, l'ensemble noté
$$\boxed{E \times F = \{(x, y)/x \in E,\ y \in F\}}$$

   Exemple 3.1.3 Soient les ensembles $E = \{0, 1, 2\}$ et $F = \{a, b\}$. On a alors
   $$E \times F = \{(0, a); (0, b); (1, a); (1, b); (2, a); (2, b)\}.$$

2. E^2 est le produit cartésien de E.

$$\boxed{E^2 = \{(x,y)/x \in E,\ y \in E\}}$$

Dans l'exemple précédent, on a $E^2 = \{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(2,0),(2,1),(2,2)\}$ et $F^2 = \{(a,a),(a,b),(b,a),(b,b)\}$.

3. On peut généraliser la définition du produit cartésien. Soient p ensembles finis E_1,E_2,…,E_p alors

$$\boxed{E_1 \times E_2 \times \ldots \times E_p = \{(x_1,x_2,\ldots,x_p)/x_1 \in E_1, x_2 \in E_2, \ldots, x_p \in E_p\}}$$

3.1.3 Notation factorielle

Soit n un entier naturel non nul ($n \in \mathbb{N}^*$), on définit $\boxed{n! = 1 \times 2 \times 3 \times \ldots \times n}$ qui se lit "factorielle n".

Exemple 3.1.4 $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

Par convention, on pose $\boxed{0! = 1}$.

Exercice 33 Simplifiez $a = \dfrac{n!}{(n-1)!}$ $n \in \mathbb{N}^*$, $b = \dfrac{n!}{(n-2)!}$ avec $n \in \mathbb{N} - \{0;1\}$.

Exercice 34 Écrire à l'aide de deux factorielles le produit $5 \times 6 \times 7 \times 8$.

3.2 Le dénombrement

3.2.1 Ensemble produit

1. Soient deux ensembles finis E et F de cardinaux respectifs n et p. Le cardinal du produit cartésien de E par F est donné par

$$\boxed{\mathrm{Card}(E \times F) = \mathrm{Card}(E) \times \mathrm{Card}(F)}$$

En effet, $E \times F = \{(x,y)/x \in E, y \in F\}$. Comme x et y peuvent prendre respectivement n et p valeurs, il y a $n \times p$ couples (x,y) possibles.

Exemple 3.2.1 Soient les ensembles $E = \{0,1,2\}$ et $F = \{a,b\}$. On a $\mathrm{Card}(E \times F) = 3 \times 2 = 6$.

2. Lorsque $F = E$,

$$\boxed{\mathrm{Card}(E^2) = \mathrm{Card}(E \times E) = \mathrm{Card}(E) \times \mathrm{Card}(E) = (\mathrm{Card}(E))^2}$$

Exemple 3.2.2 Dans l'exemple précédent, on a $\mathrm{Card}(E^2) = 3^2 = 9$ et $\mathrm{Card}(F^2) = 2^2 = 4$.

3. On peut généraliser la définition du cardinal. Soient p ensembles finis E_1,E_2,…,E_p alors

$$\boxed{\mathrm{Card}(E_1 \times E_2 \times \ldots \times E_p) = \mathrm{Card}(E_1) \times \mathrm{Card}(E_2) \times \ldots \times \mathrm{Card}(E_p)}$$

3.2.2 Nombre d'applications d'un ensemble E de cardinal p dans un ensemble F de cardinal n

1. Le nombre d'applications de E dans F est

$$\boxed{n^p = (\mathrm{Card}(F))^{\mathrm{Card}(E)}}$$

Exemple 3.2.3 Soient $E = \{0,1\}$ et $F = \{a,b,c\}$. Le nombre d'applications de E dans F est $3^2 = 9$.

- Considérons une application de E dans F, représentée par la Figure 3.1. Cette application est caractérisée par le couple (a,a) avec la convention "0 a pour image a" et "1 a pour image a".
- Considérons maintenant une nouvelle application, représentée par la Figure 3.2. Cette application est caractérisée par le couple (a,c) tel que $\left\{\begin{array}{ccc} 0 & \to & a \\ 1 & \to & c \end{array}\right.$.

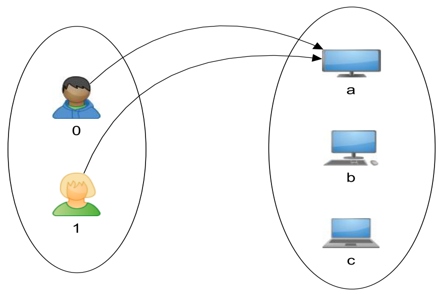


FIGURE 3.1 –

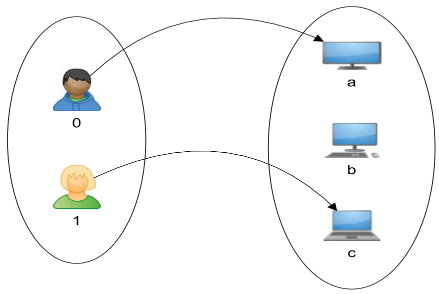


FIGURE 3.2 –

– Les neuf applications de E dans F sont donc caractérisées par les 9 couples (a, a), (b, b), (c, c), (a, b), (b, a), (a, c), (c, a), (b, c) et (c, b). Un couple (x, y) caractérise l'application
$$0 \to x$$
$$1 \to y$$
x et y pouvant prendre les valeurs a, b, c.

$\boxed{\textbf{Exercice 35}}$ Soit un ensemble de 5 balles que l'on veut répartir dans trois rangements. Chaque rangement pouvant contenir $0, 1$ ou plusieurs balles, quel est le nombre de répartitions distinctes ?

$\boxed{\textbf{Exercice 36}}$ À l'intérieur d'un bureau, un numéro de téléphone fixe est composé d'un indicatif (2 numéros) et d'une suite ordonnée de 8 numéros. Pour un indicatif donné, combien y a-t-il de numéros possibles ?

2. p-listes ordonnées avec remise
Soit un ensemble $F = \{e_1, e_2, \ldots, e_n\}$ de cardinal n. On appelle **p-liste** d'éléments de F, une liste ordonnée de p éléments de F (appelée encore p-uplet) de la forme $(x_1, x_2, \ldots, x_p)$, les x_i étant deux éléments *distincts ou non* de F. Ces p-uplets sont au nombre de n^p puisqu'il y a n choix possibles pour les p éléments.

Exemple 3.2.4 Soit $F = \{e_1, e_2, e_3\}$.
– Les 2-listes ordonnées avec remise (ce qui signifie que les éléments peuvent se répéter) ou 2-uplets sont au nombre de $3^2 = 9$ c'est-à-dire qu'il existe 9 couples (x, y) formés par deux éléments distincts ou non de F. Ces éléments sont (e_1, e_1), (e_1, e_2), (e_2, e_1), (e_1, e_3), (e_3, e_1), (e_2, e_2), (e_2, e_3), (e_3, e_2), (e_3, e_3).
– Les 3-listes ordonnées avec remise (ou 3-uplets, ou triplets) sont au nombre de $3^3 = 27$, on peut citer entre-autres (e_1, e_1, e_1), (e_2, e_1, e_3), (e_3, e_2, e_3).

$\boxed{\textbf{Exercice 37}}$ Combien peut-on former de sigles d'entreprises de deux lettres ? de trois lettres ? de quatre lettres ?

Exercice 38 On tire 3 boules d'un sac contenant 9 boules : 4 vertes, 3 rouges, 1 blanche et 1 noire. Combien y a-t-il de résultats successivement avec remise

(a) permettant d'obtenir 3 boules vertes ?
(b) permettant d'obtenir aucune boule rouge ?
(c) permettant d'obtenir 3 boules blanches ?
(d) permettant d'obtenir dans cet ordre : 2 boules vertes et 1 boule rouge ?
(e) permettant d'obtenir 2 vertes et 1 rouge ?
(f) permettant d'obtenir 1 verte, 1 rouge et 1 noire ?

3.2.3 Parties d'un ensemble et cardinaux

1. Le nombre de parties d'un ensemble E de cardinal n est 2^n.

$$\boxed{\text{Card}(E) = n \Rightarrow \text{Card}(\mathcal{P}(E)) = 2^n.}$$

Exemple 3.2.5 On a vu dans un exemple précédent que $\mathcal{P}(E) = \{\emptyset; \{0\}; \{1\}; \{2\}; \{0,1\}; \{0,2\}; \{1,2\}; E\}$ si $E = \{0,1,2\}$, ce qui confirme bien que $\text{Card}(\mathcal{P}(E)) = 2^3 = 8$.

2. Si A et B sont deux parties de E,

$$\boxed{\text{Card}(A \cup B) = \text{Card}(A) + \text{Card}(B) - \text{Card}(A \cap B)}$$

Exercice 39 Parmi les 20 employés d'une entreprise, 8 connaissent l'anglais, 5 l'allemand, 3 les deux langues. Dénombrez ceux qui connaissent au moins une langue.

3. Le complémentaire d'une partie A de E est défini par

$$\boxed{\overline{A} = \complement_E^A \text{ avec } \text{Card}(\overline{A}) = \text{Card}(E) - \text{Card}(A)}\ .$$

Exemple 3.2.6 Soient $E = \{0,1,2,3,4\}$ et $A = \{2,4\}$ alors $\overline{A} = \{0,1,3\}$ et $\text{Card}(\overline{A}) = 5 - 2 = 3$.

Exercice 40 Un sondage d'opinion, relatif à l'ambiance d'un cours de "Statistiques et probabilités", a été réalisé parmi 83 étudiants de l'ULCO et donne les résultats suivants :

	Très satisfaits	Assez satisfaits	Assez déçus	Très déçus
Filles de moins de 23 ans	9	6	2	2
Filles de 23 ans et plus	7	9	4	2
Garçons de moins de 23 ans	6	6	4	3
Garçons de 23 ans et plus	2	4	6	11

On pose : F les femmes, P les étudiants de 23 ans et plus, D les étudiants assez déçus ou très déçus, A les étudiants assez déçus ou assez satisfaits. On note $\overline{F}$, $\overline{A}$, $\overline{D}$ les ensembles complémentaires de F, A et D. Déterminez le nombre d'éléments des ensembles suivants en précisant ce qu'il représente :

(a) F, $\overline{P}$, $\overline{D}$ et A,
(b) $F \cap \overline{P} \cap D \cap A$,
(c) $(F \cap D) \cup (P \cap A)$,
(d) $\overline{F} \cap (P \cap \overline{D} \cap \overline{A})$.

3.2.4 Arrangements

1. Soient un ensemble F de cardinal n et p un entier naturel tel que $1 \leq p \leq n$. On appelle **arrangement, d'ordre** p**, des éléments de** F, un p-uplet $(x_1, x_2, \ldots, x_p)$ où les éléments x_i sont des éléments distincts de F.

   Déterminons le nombre d'arrangements d'ordre p, noté A_n^p :
   – si $p > n$: $A_n^p = 0$
   – si $p \leq n$: pour x_1, il y a n choix possibles. x_1 étant choisi, il y a $n-1$ choix possibles pour x_2 et ainsi de suite. Enfin, $x_1, x_2, \ldots, x_{p-1}$ étant choisis, il reste $n-(p-1) = n-p+1$ choix possibles pour x_p. Le nombre d'arrangements d'ordre p est donc

   $$\boxed{A_n^p = n(n-1)(n-2)\ldots(n-p+1)}$$

   Cette égalité peut se réécrire $A_n^p = n(n-1)(n-2)\ldots(n-p+1) \times \dfrac{(n-p)(n-p-1)\ldots 2 \times 1}{(n-p)(n-p-1)\ldots 2 \times 1}$, ce qui signifie que

   $$\boxed{A_n^p = \frac{n!}{(n-p)!}}$$

2. Cas particuliers : Avec la convention $0! = 1$ le résultat précédent implique

   $$\boxed{A_n^0 = 1} \text{ et } \boxed{A_n^n = n}$$

Exercice 41 L'équipe de direction d'une école de commerce de 20 membres est constituée d'un directeur Général, d'un directeur Adjoint et d'un directeur des Relations Internationales. Combien d'équipes de direction peut-on constituer, sachant qu'une même personne ne peut cumuler les postes ?

Exercice 42 Soit le mot $FIABILITE$. À l'aide de ce mot, combien d'anagrammes et combien de mots - au sens large - de 5 lettres distinctes peut-on former ?

3.2.5 Permutations

1. On appelle **permutation** d'un ensemble F de cardinal n, un arrangement d'ordre n de F. Le nombre de ces permutations est donné par $P_n = A_n^n = n!$.
2. Permutation avec répétition

   Exemple 3.2.7 Déterminons le nombre d'anagrammes du mot $FINI$.
   Si on considère les deux I comme différents, I_1 et I_2, les 4 lettres de FI_1NI_2 étant distinctes, il existe $4! = 24$ anagrammes. Dans ces 24 mots, les mots FI_1NI_2 et FI_2NI_1 apparaissent. Chaque mot est comptabilisé deux fois. Le nombre d'anagrammes de $FINI$ est donc $\dfrac{4!}{2!} = 12$.

   Exemple 3.2.8 Déterminons le nombre d'anagrammes du mot $ENSEMBLE$.
   Si on numérote les trois E, on obtient $E_1NSE_2MBLE_3$. Les huit lettres étant distinctes, il existe $8!$ anagrammes de $E_1NSE_2MBLE_3$. Dans ces $8!$ mots, le mot $ENSEMBLE$ apparaît $3! = 6$ fois, chaque mot est donc comptabilisé $3!$ fois. Le nombre d'anagrammes de $ENSEMBLE$ est donc $\dfrac{8!}{3!}$.

   Exemple 3.2.9 Déterminons le nombre d'anagrammes du mot $MATHEMATIQUES$.
   En procédant comme précédemment, on obtient $\dfrac{13!}{2!2!2!2!}$ anagrammes.

3. Cas général
   Soit une famille E de cardinal n, définie par $E = \{a_1, a_2, \ldots, a_k\}$, la lettre a_1 étant répétée r_1 fois, la lettre a_2 r_2 fois, la lettre a_k r_k fois avec $r_1 + r_2 + \ldots + r_k = n$. Le nombre de permutations est alors

$$\boxed{\frac{n!}{r_1!r_2!\dots r_k!}}$$

Exercice 43 Pour faire un signal maritime, on hisse 10 drapeaux le long d'un mât dans un ordre précis, au moyen d'un filin.

(a) Si on considère que tous les drapeaux sont différents, combien de signaux distincts peut-on faire ?

(b) Si 4 de ces drapeaux sont de couleur bleue et parfaitement identiques, et les 6 autres de couleur rouge, parfaitement identiques également, combien de signaux distincts peut-on faire ?

3.2.6 Combinaisons

1. Soit un ensemble F ayant n éléments distincts. On appelle **combinaison d'ordre** p **de** F toute partie à p éléments ($0 \leq p \leq n$).

   Remarque 3.2.1 Deux combinaisons distinctes d'ordre p diffèrent par la nature de leurs éléments et non pas par l'ordre.

   Exemple 3.2.10 Soit $F = \{a, b, c, d\}$. Les combinaisons d'ordre 3 de F sont $\{a, b, c\}$, $\{a, b, d\}$, $\{a, c, d\}$, $\{b, c, d\}$.

   Exemple 3.2.11 Avec les éléments a, b, c, on peut constituer
   – une combinaison de 3 éléments : $\{a, b, c\}$,
   – six arrangements de 3 éléments : (a, b, c), (a, c, b), (b, a, c), (b, c, a), (c, a, b) et (c, b, a).

   Le nombre de combinaisons d'ordre p est noté C_n^p ou $\begin{pmatrix} n \\ p \end{pmatrix}$ et est défini par

   $$\boxed{C_n^p = \frac{A_n^p}{p!} = \frac{n!}{p!(n-p)!} = \frac{n(n-1)(n-2)\dots(n-p+1)}{p!}}$$

   Les C_n^p sont encore appelés **coefficients binomiaux**.

Exercice 44 Un porte-monnaie contient 2 pièces de 2 euros, 4 pièces d'1 euro, 4 pièces de 50 centimes et 6 pièces de 20 centimes. On prend au hasard 5 pièces dans le porte-monnaie.
De combien de façons différentes peut-on obtenir un total de 4 euros ?

2. Propriétés des nombres C_n^p
   – On a le résultat
   $$\boxed{C_n^p = C_n^{n-p}} \text{ pour } n \geq 0 \text{ et } 0 \leq p \leq n.$$
   En particulier, $C_n^0 = C_n^n = 1$ et $C_n^1 = C_n^{n-1} = n$.

   – On a également
   $$\boxed{C_n^p = C_{n-1}^p + C_{n-1}^{p-1}} \text{ pour } n \geq 1 \text{ et } 1 \leq p \leq n.$$

3. Binôme de Newton
   On a la formule
   $$\boxed{(x+y)^n = \sum_{p=0}^{n} C_n^p x^p y^{n-p}} \quad \text{pour } n \geq 0.$$

   Exemple 3.2.12
   $$\begin{aligned}(x+y)^5 &= C_5^0 x^0 y^5 + C_5^1 x^1 y^4 + C_5^2 x^2 y^3 + C_5^3 x^3 y^2 + C_5^4 x^4 y^1 + C_5^5 x^5 y^0 \\ &= y^5 + 5y^4 x + 10y^3 x^2 + 10y^2 x^3 + 5yx^4 + x^5\end{aligned}$$

Exercice 45 Développez à l'aide de la formule du binôme les expressions suivantes où a est un nombre réel quelconque

(a) $E = (a+1)^4$

(b) $F = (a-3)^5$

4. **Le triangle de Pascal**
Ce triangle permet par simple addition de récupérer les coefficients C_n^p à partir des C_{n-1}^p.

$n \backslash p$	0	1	2	3	4	5	$\cdots$	$p-1$	p
0	1								
1	1	1							
2	1	2	1						
3	1	3	3	1					
4	1	4	6	4	1				
5	1	5	10	10	5	1			
$\vdots$	$\vdots$						$\ddots$		
$n-1$								C_{n-1}^{p-1}	C_{n-1}^p
n									C_n^p

3.2.7 Combinaisons avec répétition

Soit un ensemble F de cardinal n. On nomme **combinaison d'ordre p, avec répétition des éléments de E**, une liste de p éléments tous extraits de E, les répétitions étant autorisées mais l'ordre dans la liste n'intervient pas.

Le nombre de combinaisons avec répétition, d'ordre p, est noté

$$\boxed{\Gamma_n^p = C_{n+p-1}^p = C_{n+p-1}^{n-1} = \frac{(n+p-1)!}{p!(n-1)!}}$$

Exemple 3.2.13 5 clients vous commandent 5 composants électroniques parmi les 8 que vous fabriquez. Quel est le nombre de commandes globales distinctes dans les cas suivants ?

– Les composants sont tous différents : $C_8^5 = \dfrac{8!}{5!3!} = 56$.
– Les composants sont tous quelconques : $C_{8+5-1}^5 = C_{12}^5 = 792$.

3.2.8 Modèle fondamental : schéma d'urne

Il est très pratique dans les exercices de considérer l'ensemble E impliqué comme une **urne** contenant n boules numérotées de 1 à n, chacune des boules s'interprétant comme un élément de E, et de laquelle on tire p boules. On aura très souvent les cas suivants :

– *Tirages successifs avec remise* : On tire au hasard une boule dans l'urne puis on la remet dans l'urne avant d'effectuer le tirage suivant. Si on effectue ainsi p tirages avec remise, le résultat global s'interprète comme une p-liste. Il y a donc n^p tirages avec remise (de p éléments) possibles.
– *Tirages successifs et sans remise* : On tire au hasard une boule dans l'urne que l'on conserve, la boule n'est donc pas remise dans l'urne qui contient ainsi après chaque tirage une boule de moins. Si on effectue ainsi p tirages sans remise ($p \leq n$), le résultat global s'interprète comme une p-liste d'éléments 2 à 2 distincts ou encore comme un arrangement de p éléments de E. Il y a donc A_n^p tirages sans remise (de p éléments) possibles.

- *Tirages simultanés* : On tire simultanément p boules de l'urne (et non plus successivement, cela revient à dire que l'ordre du tirage des boules est sans importance). Un tel tirage s'interprète comme un sous-ensemble de E et donc comme une combinaison de p éléments de E. Il y a donc C_n^p tirages simultanés (de p éléments) possibles.

3.3 Exercices

Exercice 46 Soit Ω un ensemble. On note $\mathcal{P}$ l'ensemble de ses parties.

1. Considérons l'ensemble $\Omega_1 = \{1\}$. Déterminez l'ensemble $\mathcal{P}(\Omega_1)$.
2. Considérons l'ensemble $\Omega_2 = \{1, 2\}$. Déterminez l'ensemble $\mathcal{P}(\Omega_2)$.
3. Considérons l'ensemble $\Omega_3 = \{1, 2, 3\}$. Déterminez l'ensemble $\mathcal{P}(\Omega_3)$.

Exercice 47 Une classe de Licence 1 se compose de 20 garçons et 8 filles.

1. De combien de façons peut-on désigner trois garçons pour tenir les rôles des 3 premiers de la promotion ?
2. De combien de façons peut-on désigner trois garçons et deux filles pour tenir les 5 premiers rôles de la promotion ?

Exercice 48 Un syndicat réunit quatre délégués qui doivent décider de l'éventualité d'une grève. Chaque délégué peut voter "oui", "non " ou s'abstenir.

1. De combien de manières dictinctes les délégués peuvent-ils répondre ?
2. Parmi toutes les réponses possibles, quelles sont celles favorables à la grève, c'est-à-dire 3 "oui" ou plus ?

Exercice 49 Une urne contient des jetons numérotés de 1 à 20. On tire un jeton.

1. Calculer les cardinaux des événements suivants :
   A : "obtenir un jeton pair",
   B : "obtenir un jeton impair",
   C : "obtenir un jeton divisible par 3",
   D : "obtenir un jeton au moins égal à 2".
2. Déterminer $(B \cap C)$ et en déduire $Card(B \cup C)$.

Exercice 50 On tire 3 cartes simultanément d'un jeu de 32 cartes.

1. Combien y a-t-il de tirages possibles ?
2. Calculer les cardinaux des événements suivants :
   A : "un roi et un seul",
   B : "un roi rouge et un seul",
   C : "trois dames exactement",
   D : "au moins un as",
   E : "un valet et une dame noire",
   F : "un seul roi et deux piques seulement".

Exercice 51 On dispose de 3 dés cubiques (un rouge, un vert, un blanc) dont les faces sont numérotées de 1 à 6. On jette les 3 dés sur une table et on forme un nombre de 3 chiffres :
- le chiffre des centaines est donné par le dé rouge,
- le chiffre des dizaines est donné par le dé vert,
- le chiffre des unités est donné par le dé blanc.

1. Combien y a-t-il de nombres possibles ?
2. Calculer les cardinaux des événements suivants :
   A : “les trois chiffres du nombre sont égaux”,
   B : “les trois chiffres du nombre sont différents”,
   C : “deux des trois chiffres au moins sont égaux”,
   D : “le nombre formé est pair”,
   E : “le nombre formé commence par 3”,
   F : “le nombre formé est divisible par 9”.

Exercice 52 Un voyageur de commerce veut visiter les cinq villes suivantes : Arras, Boulogne, Cherbourg, Digne et Epernay.

1. Combien de circuits différents peut-il réaliser ?
2. Il commence son voyage par Arras. Combien de circuits différents peut-il réaliser ?
3. Il commence son voyage par Arras et doit terminer par Epernay. Combien de circuits différents peut-il réaliser ?

Exercice 53 Résoudre dans l'ensemble des entiers naturels les équations suivantes :

1. $A_n^3 = 210n$
2. $C_n^2 = 6$
3. $C_n^1 + C_n^2 + C_n^3 = 5n$

Exercice 54 On effectue 20 fois une même expérience qui n'a que deux issues possibles : soit elle réussit, soit elle échoue.

1. Combien y a-t-il de suites d'observations des réussites et échecs ?
2. Combien y a -t-il de suites d'observations comptant 5 réussites ?

Chapitre 4

La probabilité

4.1 Le vocabulaire

4.1.1 Expérience aléatoire et univers

La théorie des probabilités fournit des modèles mathématiques permettant l'étude d'expériences dont on ne peut prévoir le résultat avec certitude. Une telle expérience est appelée **expérience aléatoire**.
On appelle **univers**, noté Ω, tout ensemble dont les éléments représentent tous les résultats possibles d'une expérience aléatoire. L'ensemble des résultats possibles est connu.

4.1.2 Evénements

Soit Ω un univers correspondant à une certaine expérience aléatoire. Il sera noté généralement

$$\boxed{\Omega = \{w_1, w_2, \dots, w_p\}}$$

Les w_i, éléments de Ω, sont des **éventualités**. Le singleton $\{w_i\}$ est appelé **événement élémentaire**. Une partie de Ω est appelée **événement**, c'est un élément de $\mathcal{P}(\Omega)$ (qui on le rappelle, définit l'ensemble des parties de Ω).

Exemple 4.1.1 Dans le cas du "lancer d'un dé à 6 faces ", les éventualités sont $1, 2, 3, 4, 5, 6$, les événements élémentaires $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$ et $\{6\}$. Les événements sont les parties de Ω, par exemple, l'événement "obtenir un nombre pair " peut s'écrire $A = \{2, 4, 6\}$.

On a également les définitions suivantes : soit A un événement de Ω,
- on dit que A **se réalise** si le résultat obtenu à l'issue de l'expérience aléatoire est un élément de A.
- Soit A un événement de Ω. On appelle **événement contraire de** A (ou **complémentaire**), noté $\bar{A}$, l'ensemble des éléments de Ω qui n'appartiennent pas à A. Autrement dit, l'événement $\bar{A}$ se réalise lorsque l'événement A ne se réalise pas.
- Soient A et B deux événements de Ω. On appelle **intersection de** A **et** B le sous-ensemble des éléments de Ω qui appartiennent à la fois à A et à B et on le note $A \cap B$. Autrement dit, l'événement $A \cap B$ se réalise lorsque les événements A et B se réalisent.
- Soient A et B deux événements de Ω. On appelle **réunion de** A **et** B le sous-ensemble des éléments de Ω qui appartiennent à A ou à B et on le note $A \cup B$. Autrement dit, l'événement $A \cup B$ se réalise lorsque au moins l'un des événements A et B se réalise.

4.1.3 Propriétés de $\mathcal{P}(\Omega)$

1. $\Omega \in \mathcal{P}(\Omega)$, Ω est appelé **événement certain**.
2. $\forall A \in \mathcal{P}(\Omega) \Rightarrow \overline{A} \in \mathcal{P}(\Omega)$
   En particulier $\Omega \in \mathcal{P}(\Omega) \Rightarrow \overline{\Omega} = \emptyset \in \mathcal{P}(\Omega)$. L'ensemble vide est appelé **événement impossible**.

3. $\left\{\begin{array}{l} A \in \mathcal{P}(\Omega) \\ B \in \mathcal{P}(\Omega) \end{array}\right. \Rightarrow \left\{\begin{array}{l} A \cup B \in \mathcal{P}(\Omega) \\ A \cap B \in \mathcal{P}(\Omega) \end{array}\right.$
4. $\overline{A \cap B} = \overline{A} \cup \overline{B}$
5. $\overline{A \cup B} = \overline{A} \cap \overline{B}$
6. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \;\; A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

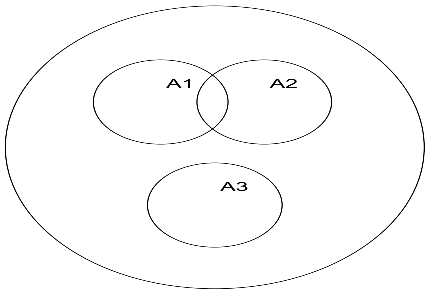
4.1.4 Opérations sur les événements

Les opérations sur les ensembles ont une interprétation en matière d'événements.

1. $A \subset B$: la réalisation de A implique celle de B.
2. $A \cup B$ désigne l'événement "A ou B", il se produit si au moins un des événements est réalisé.
3. $A \cap B$ désigne l'événement "A et B", il se produit si A et B sont tous les deux réalisés.
4. Soient deux parties disjointes A et B, c'est-à-dire telles que $A \cap B = \emptyset$. Les événements A et B sont dits **incompatibles**.
5. Soit un système d'événements $A_i, \; i \in I$. Les événements A_i sont dits **globalement incompatibles** si $\bigcap_{i \in I} A_i = \emptyset$.

   Remarque 4.1.1 Si les A_i sont deux à deux incompatibles alors ils sont globalement incompatibles, par contre la *réciproque* est fausse.

   Exemple 4.1.2 Avec 3 événements A_1, A_2 et A_3 :

   on a donc bien $A_1 \cap A_2 \cap A_3 = \emptyset$ mais $A_1 \cap A_2 \neq \emptyset$, ce qui illustre la remarque précédente.

4.2 Probabilité

4.2.1 Axiome de probabilité

Soient une expérience aléatoire et Ω l'univers associé, $\mathcal{P}(\Omega)$ l'ensemble des événements. On définit une probabilité comme une application qui à un événement associe un nombre qui mesure les chances de réalisation de cet événement. Mathématiquement parlant, une probabilité est une application

$$\begin{array}{rrcl} p: & \mathcal{P}(\Omega) & \rightarrow & \mathbb{R}^+ \\ & A & \mapsto & p(A) \in [0;1] \end{array}$$

On dit que $p(A)$ est la probabilité de l'événement A. L'application p vérifie les axiomes suivants

1. $p(\Omega) = 1$
2. si $A \cap B = \emptyset$ (les événements sont incompatibles) alors $p(A \cup B) = p(A) + p(B)$.

Remarque 4.2.1 Le couple $(\Omega, \mathcal{P}(\Omega))$ est **probabilisé** par la définition de la probabilité p. Pour probabiliser $(\Omega, \mathcal{P}(\Omega))$, il existe une infinité de probabilités possibles. Le choix de la probabilité résulte d'hypothèses faites sur l'épreuve aléatoire.

4.2.2 Conséquences

1. $\boxed{p(\emptyset) = 0}$

   Preuve : On a $\emptyset \cup \emptyset = \emptyset$ et $\emptyset \cap \emptyset = \emptyset$ donc $p(\emptyset) = p(\emptyset) + p(\emptyset) \Rightarrow p(\emptyset) = 0$. ∎

2. $\boxed{p(A) + p(\overline{A}) = 1}$

   Preuve :
   $$\left.\begin{array}{l} A \cap \overline{A} = \emptyset \\ A \cup \overline{A} = \Omega \end{array}\right\} \Rightarrow p(\Omega) = 1 = p(A) + p(\overline{A}).$$ ∎

3. $\boxed{A \subset B \Rightarrow p(A) \leq p(B)}$

   Preuve :
   $$\left.\begin{array}{l} A \cup (B - A) = b \\ A \cap (B - A) = \emptyset \end{array}\right\} \Rightarrow p(B) = p(A) + p(B - A).$$ Comme $p(B - A) \in \mathbb{R}^+$ on a $p(A) \leq p(B)$. ∎

4. $\boxed{A \subset \Omega \Rightarrow 0 \leq p(A) \leq 1}$

   Preuve : $\emptyset \subset A \subset \Omega \Rightarrow p(\emptyset) \leq p(A) \leq p(\Omega)$ donc $0 \leq p(A) \leq 1$. ∎

5. $\boxed{p(A \cup B) = p(A) + p(B) - p(A \cap B)}$

   Preuve :
   $$\left.\begin{array}{l} A \cup B = B \cup (A/B) \\ B \cap (A/B) = \emptyset \end{array}\right\} \Rightarrow p(A \cup B) = p(B) + p(A/B)$$ d'après le deuxième axiome de la probabilité,
   $$\left.\begin{array}{l} (A/B) \cup (A \cap B) = A \\ (A/B) \cap (A \cap B) = \emptyset \end{array}\right\} \Rightarrow p(A) = p(A \cap B) + p(A/B),$$
   $p(A \cup B) - p(B) = p(A/B) = p(A) - p(A \cap B)$ donc $p(A \cup B) = p(A) + p(B) - p(A \cap B)$. ∎

6. $\boxed{p(A \cup B \cup C) = p(A) + p(B) + p(C) - p(A \cap B) - p(A \cap C) - p(B \cap C) + p(A \cap B \cap C)}$

   Preuve : En posant $B \cup C = X$ on obtient $p(A \cup B \cup C) = p(A \cup X) = p(A) + p(X) - p(A \cap X)$ d'après le résultat précédent. Or $p(X) = p(B \cup C) = p(B) + p(C) - p(B \cap C)$ et $A \cap X = A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. Par conséquent, $p(A \cap X) = p(A \cap B) + p(A \cap C) - p(A \cap B \cap A \cap C) = p(A \cap B) + p(A \cap C) - p(A \cap B \cap C)$. En remplaçant cette expression dans $p(A \cup B \cup C)$ on obtient l'égalité. ∎

   Remarque 4.2.2
   – Dans le cas où les événements A, B et C sont incompatibles deux à deux, on obtient
   $$\boxed{p(A \cup B \cup C) = p(A) + p(B) + p(C)}$$
   – On peut généraliser cette propriété à n événements incompatibles deux à deux
   $$\boxed{p\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} p(A_i) \text{ avec } A_i \cap A_j = \emptyset \text{ pour } i \neq j}$$

4.3 Ensembles probabilisés

4.3.1 Ensembles finis probabilisés

1. Soit Ω un ensemble fini de cardinal $n \in \mathbb{N}^\star$ donné par :

$$\Omega = \{w_1, w_2, \ldots, w_n\}$$

On définit une probabilité de la façon suivante

- $p_i = p(\{w_i\}) \geq 0$
- $\displaystyle\sum_{i=1}^{n} p_i = 1$

Remarque 4.3.1 Si A est une partie de Ω, $p(A) = \displaystyle\sum_{w_i \in A} p(\{w_i\})$.

2. On dit qu'on a **équiprobabilité sur** Ω si tous les événements élémentaires ont la même probabilité, autrement dit si

$$p(\{\omega_1\}) = p(\{\omega_2\}) = \ldots = p(\{\omega_n\}).$$

Remarque 4.3.2 Si on a équiprobabilité sur Ω alors, d'après la définition d'une probabilité, on a

$$p(\{\omega_1\}) = p(\{\omega_2\}) = \ldots = p(\{\omega_n\}) = \frac{1}{n}.$$

et, pour tout événement A,

$$\boxed{p(A) = \frac{k}{n} = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega)}}$$

où, on le rappelle, $\mathrm{Card}(A)$ désigne le nombre d'éléments contenus dans l'ensemble A.

Remarque 4.3.3 Considérons l'expérience aléatoire qui consiste à lancer un dé cubique. Supposons que le dé soit équilibré (ou non truqué ou non pipé), dans ce cas, si p_i est la probabilité d'obtenir la face i avec $i \in \{1, 2, \ldots, 6\}$, $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \dfrac{1}{6}$.

Soit A l'événement "le numéro de la face supérieure du dé est pair". Alors $p(A) = \dfrac{3}{6}$ de par l'équiprobabilité des événements élémentaires. En effet, $p(A) = p(\{2\} \cup \{4\} \cup \{6\}) = p(\{2\}) + p(\{4\}) + p(\{6\})$ car les événements (élémentaires) $\{2\}$, $\{4\}$ et $\{6\}$ sont incompatibles deux à deux.

4.3.2 Ensembles infinis dénombrables probabilisés

Un ensemble Ω est un ensemble **dénombrable infini** s'il existe une bijection de $\mathbb{N}$ ou de $\mathbb{N}^*$ dans Ω. Il peut s'écrire sous la forme :

$$\Omega = \{w_1, w_2, w_3, \ldots, w_n, \ldots\ldots\}$$

On définit une probabilité p sur $(\Omega, \mathcal{P}(\Omega))$ en attribuant à chaque w_i une probabilité $p_i = p(\{w_i\}) \geq 0$ telle que $\displaystyle\sum_{i \in \mathbb{N}^*} p_i = 1$.

Exemple 4.3.1 On réalise une expérience qui n'a que deux issues possibles (l'échec ou la réussite) jusqu'à ce qu'elle réussisse. On s'intéresse au nombre de réalisations nécessaires à la réussite. Ce nombre est variable, $\Omega = \mathbb{N}^*$ donc Ω est un ensemble infini dénombrable.

$$\begin{array}{llll}
\bullet\, p_1 & = & p(\{1\}) & = \dfrac{1}{2}, \text{ la "réussite" apparaît lors de la première réalisation,} \\
\bullet\, p_2 & = & p(\{2\}) & = \dfrac{1}{2} \times \dfrac{1}{2} = \dfrac{1}{4}, \text{ la "réussite" apparaît lors de la deuxième réalisation,} \\
 & \vdots & & \\
\bullet\, p_n & = & p(\{n\}) & = \left(\dfrac{1}{2}\right)^{n-1}\left(\dfrac{1}{2}\right) = \left(\dfrac{1}{2}\right)^{n}, \text{ la "réussite" apparaît lors de la } n\text{-ième réalisation} \\
 & & & \text{donc les}(n-1) \text{ premières réalisations ont donné un "échec"} \\
 & \vdots & &
\end{array}$$

Ainsi, $\sum_{i\in\mathbb{N}^*} p_i = \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \ldots + \left(\frac{1}{2}\right)^n + \ldots$.

Pour déterminer cette somme, on rappelle la formule

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

En effet, $\sum_{i=1}^{n} p_i$ est la somme des n premiers termes d'une suite géométrique de raison $\frac{1}{2}$.

De plus, $\lim_{n\to\infty} \left(\frac{1}{2}\right)^{n+1} = 0$ donc $\sum_{i\in\mathbb{N}^*} p_i = \lim_{n\to+\infty} \sum_{i=1}^{n} \left(\frac{1}{2}\right)^i = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1$.

4.4 Probabilité conditionnelle

4.4.1 Définition

1. Étant donné un espace probabilisé $(\Omega, \mathcal{P}(\Omega), p)$, A étant un événement de probabilité non nulle, considérons l'application notée p_A telle que

$$\boxed{\begin{array}{rccl} p_A: & \mathcal{P}(\Omega) & \to & \mathbb{R}^+ \\ & X & \mapsto & p_A(X) = \dfrac{p(A\cap X)}{p(A)} \end{array}}$$

2. L'application p_A est une probabilité. En effet,
   - $p_A(\Omega) = \dfrac{p(A\cap\Omega)}{p(A)} = \dfrac{p(A)}{p(A)} = 1$
   - Si $X\cap Y = \emptyset$, $p_A(X\cup Y) = \dfrac{p(A\cap(X\cup Y))}{p(A)}$ mais on peut remarquer que $A\cap(X\cup Y) = (A\cap X)\cup(A\cap Y)$ et $(A\cap X)\cap(A\cap Y) = A\cap X\cap Y = X\cap Y = \emptyset$. Par conséquent,
   $$p_A(X\cup Y) = \frac{p(A\cap X) + p(A\cap Y)}{p(A)} = \frac{p(A\cap X)}{p(A)} + \frac{p(A\cap Y)}{p(A)}.$$
   Finalement, si $X\cap Y = \emptyset$ alors $p_A(X\cup Y) = p_A(X) + p_A(Y)$.

3. $p_A(X) = \dfrac{p(A\cap X)}{p(A)}$ est encore notée $p(X/A)$ et est appelée **probabilité conditionnelle de X sachant A** ou **probabilité de X sachant A** ou **probabilité de X une fois A réalisé**.

Remarque 4.4.1

- L'application p_A est une probabilité, elle vérifie donc les propriétés de la probabilité, en particulier
  - $p_A(\emptyset) = 0$
  - $p_A(X) + p_A(\overline{X}) = 1$
- Si $p(A) \neq 0$ et $p(B) \neq 0$ alors
$$\boxed{p_A(B) = \frac{p(A\cap B)}{p(A)} = p(B/A) \text{ et } p_B(A) = \frac{p(A\cap B)}{p(B)} = p(A/B)}$$
  Donc
$$\boxed{p(A\cap B) = p(A)\times p(B/A) = p(B)\times p(A/B)}$$
- Pour trois événements A, B et C, on a :
$$p(A)\times p(B/A)\times p(C/A\cap B) = p(A)\times\frac{p(B\cap A)}{p(A)}\times\frac{p(A\cap B\cap C)}{p(A\cap B)}$$

donc

$$\boxed{p(A \cap B \cap C) = p(A) \times p(B/A) \times p(C/A \cap B)}$$

4.4.2 Exemple

Un technicien doit régler de toute urgence un problème électrique sur une machine. La partie défaillante étant hors de vue mais pas hors d'atteinte, notre réparateur a la possibilité de déconnecter 3 fusibles noirs (numérotés N_1, N_2, N_3) et 2 fusibles rouges (numérotés R_1, R_2), indiscernables au toucher malheureusement pour lui. Il doit déconnecter deux fusibles rouges successivement pour régler le problème.
On a les éventualités suivantes :

(N_1, R_1) ; (N_1, R_2) ; (N_2, R_1) ; (N_2, R_2) ; (N_3, R_1) ; (N_3, R_2) ; (R_1, N_1) ; (R_2, N_1) ;
(R_1, N_2) ; (R_2, N_2) ; (R_1, N_3) ; (R_2, N_3) ; (R_1, R_2) ; (N_1, N_2) ; (N_1, N_3) ; (N_2, N_3) ;
(R_2, R_1) ; (N_2, N_1) ; (N_3, N_1) ; (N_3, N_2)

On dénombre donc 20 résultats possibles. On peut retrouver ce nombre en utilisant des techniques de dénombrement : notre problème peut être asssimilé à un arrangement d'ordre 2 des 5 fusibles. On en dénombre $A_5^2 = \frac{5!}{3!} = 5 \times 4 = 20$.

- On considère les événements :
  - A : “déconnecter un fusible rouge au premier essai”.
  - B : “déconnecter un fusible rouge au second essai”.

  On a $p(A) = \frac{2}{5}$, $p(A \cap B) = \frac{2}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$, $p(B) = \frac{8}{20} = \frac{2}{5}$. Ainsi, en utilisant la définition de la probabilité conditionnelle, $p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{\frac{2}{20}}{\frac{8}{20}} = \frac{1}{4}$.
- Sans utiliser la définition de $p(B/A)$, l'événement A étant réalisé (on a déconnecté un fusible rouge au premier essai) dans les 8 cas suivants

  (R_1, N_1) ; (R_1, N_2) ; (R_1, N_3) ; (R_1, R_2) ; (R_2, N_1) ; (R_2, N_2) ; (R_2, N_3) ; (R_2, R_1),

  quelle est la probabilité de B ? $p(B/A) = \frac{2}{8} = \frac{1}{4}$ c'est-à-dire que sur les 8 cas de réalisation de A, 2 cas donnent la réalisation de B, les cas (R_1, R_2) et (R_2, R_1).

4.4.3 Indépendance en probabilité

1. Étant donné un espace probabilisé $(\Omega, \mathcal{P}(\Omega), p)$, on dit que deux événements A et B sont **indépendants** si et seulement si

$$\boxed{p(A \cap B) = p(A) \times p(B)}$$

2. Propriétés équivalentes : Si $p(A) \neq 0$ et $p(B) \neq 0$,
   - les événements A et B sont indépendants si et seulement si $p(A) = p(A/B) \Leftrightarrow p(B) = p(B/A)$,
   - la réalisation de A n'influence pas B ; de même la réalisation de B n'influence pas A.

   Remarque 4.4.2 On ne confondra pas “indépendance en probabilité” et“événements incompatibles” qui se traduisent respectivement par $p(A \cap B) = p(A)p(B)$ et $A \cap B = \emptyset$.

3. Généralisation : Trois événements sont dits **globalement indépendants** en probabilité s'ils sont deux à deux indépendants en probabilité c'est-à-dire si
   - $p(A \cap B) = p(A)p(B)$
   - $p(B \cap C) = p(B)p(C)$
   - $p(A \cap C) = p(A)p(C)$

4. Soient A et B deux événements indépendants alors $\overline{A}$ et B sont indépendants.
En effet $p(\overline{A} \cap B) = p(B) \times p(\overline{A}/B)$ or $p(\overline{A}/B) = 1 - p(A/B)$ donc $p(\overline{A} \cap B) = p(B)[1 - p(A/B)] = p(B)[1 - p(A)]$ car A et B sont indépendants. Enfin, $p(\overline{A} \cap B) = p(B)p(\overline{A})$ ce qui signifie que $\overline{A}$ et B sont indépendants.

Remarque 4.4.3 Il en est de même pour A et $\overline{B}$, $\overline{A}$ et $\overline{B}$.

Exercice 55 Une classe de Prépa 1 - ISCID compte 4 garçons et 6 filles en première année, 6 garçons en seconde année. Combien doit-il y avoir de filles de seconde année si l'on veut que "sexe" et "année" soient des facteurs indépendants lors du choix au hasard d'un étudiant ?

4.4.4 La formule de Bayes

1. Soient $(\Omega, \mathcal{P}(\Omega), p)$ un espace probabilisé et un système complet d'événements $B_1, B_2, \ldots, B_n$ deux à deux incompatibles ($B_i \cap B_j = \emptyset$ pour $i \neq j$ et $\bigcup_{i \in I} B_i = \Omega$), on dit que les $(B_i)_{i \in I}$ forment une **partition** de Ω.

   Soit A un événement de probabilité non nulle, on peut alors écrire la formule de Bayes donnant la probabilité pour que B_i se réalise sachant A :

$$\boxed{p(B_i/A) = \frac{p(B_i) \times p(A/B_i)}{\sum_{j \in I} p(B_j) \times p(A/B_j)}}$$

   Preuve : $A = A \cap \Omega = A \cap \left(\bigcup_{i \in I} B_i \right) = \bigcup_{i \in I} (A \cap B_i)$. De plus, $(A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset$ pour $i \neq j$. Par conséquent, $p(A) = \sum_{i \in I} p(A \cap B_i) = \sum_{i \in I} p(B_i) \times p(A/B_i)$. Or on sait que $p(B_i/A)$ peut s'écrire sous la forme $p(B_i/A) = \dfrac{p(A \cap B_i)}{p(A)} = \dfrac{p(B_i) \times p(A/B_i)}{p(A)}$ d'où la formule de Bayes. ■

2. Dans le cas d'un système complet de deux événements B_1 et B_2 ($B_1 \cap B_2 = \emptyset,\ B_1 \cup B_2 = \Omega$),
   - $p(B_1/A) = \dfrac{p(B_1) \times p(A/B_1)}{p(B_1) \times p(A/B_1) + p(B_2) \times p(A/B_2)}$
   - $p(B_2/A) = \dfrac{p(B_2) \times p(A/B_2)}{p(B_2) \times p(A/B_2) + p(B_1) \times p(A/B_1)}$

3. Dans le cas d'un système complet de trois événements B_1, B_2 et B_3 ($B_1 \cap B_2 = B_2 \cap B_3 = B_1 \cap B_3 = \emptyset,\ B_1 \cup B_2 \cup B_3 = \Omega$), on a par exemple

$$p(B_1/A) = \frac{p(B_1) \times p(A/B_1)}{p(B_1) \times p(A/B_1) + p(B_2) \times p(A/B_2) + p(B_3) \times p(A/B_3)}.$$

Exercice 56 On suppose que les essais d'un test médical sur une population ont conduit à admettre pour un individu les probabilités suivantes, le test servant à dépister une certaine maladie.

- Probabilité pour qu'un malade ait un test positif (donc probabilité pour que le test soit positif sachant que la personne est malade) : $p(T/M) = 0,95$.
- Probabilité pour qu'un non-malade ait un test négatif (donc probabilité pour que le test soit négatif sachant que la personne est saine) : $p(\overline{T}/\overline{M}) = 0,95$.
- Probabilité pour qu'un individu soit atteint de la maladie $p(M) = 0,01$.

Quelle est la probabilité pour qu'un individu qui a donné lieu à un test positif soit atteint de la maladie ?

4.5 Exercices

Exercice 57 On mesure les longueurs des boulons d'une certaine boîte de 100.
On obtient les résultats suivants :

Longueur en cm	$[4; 4,2[$	$[4,2; 4,4[$	$[4,4; 4,6[$	$[4,6; 5[$
Effectifs	17	24	51	8

On tire au hasard un boulon. Calculez les probabilités des événements suivants :

1. Le boulon mesure moins de $4,2$ cm.
2. Le boulon mesure plus de $4,4$ cm.

Un boulon est utilisable si sa longueur est comprise entre $4,2$ et $4,6$ cm.

3. Quelle est la probabilité qu'un boulon soit utilisable ?
4. On achète 50 boîtes de 100 boulons. Combien peut-on espérer de boulons utilisables ?

Exercice 58 Deux ateliers, notés A et B, d'une même entreprise produisent chaque jour respectivement 1000 et 800 puces électroniques d'un même modèle. 2% des pièces produites par l'atelier A et 3% des pièces produites par l'atelier B sont défectueuses.

1. Complétez le tableau suivant qui décrit la production journalière .

	Nombre de puces défectueuses	Nombre de puces non défectueuses	Total
Nombre de puces produites par l'atelier A			
Nombre de puces produites par l'atelier B			
Total			1800

2. Un jour donné, on choisit au hasard une puce parmi les 1800 puces produites par les deux ateliers. On est dans une situation d'équiprobabilité. On considère les événements suivants :
   A : "la puce choisie provient de l'atelier A",
   B : "la puce choisie provient de l'atelier B",
   D : "la puce choisie est défectueuse",
   $\overline{\mathrm{D}}$: "la puce choisie n'est pas défectueuse".
   Déterminez exclusivement à l'aide du tableau précédent les probabilités suivantes :
   (a) $p(D)$, $p(A \cap D)$, $p(A/D)$.
   (b) $p(\overline{D})$, $p(B \cap \overline{D})$, $p(B/\overline{D})$.
3. Vérifiez que $p(A \cap D) = p(A/D) \times p(D)$ et que $p(B \cap \overline{D}) = p(B/\overline{D}) \times p(\overline{D})$.

Exercice 59 Dans un lot de pièces fabriquées, il y a 3% de pièces défectueuses. Le mécanisme de contrôle des pièces est aléatoire. Si la pièce est bonne, elle est acceptée avec une probabilité de $0,96$ et si elle est défectueuse elle est refusée avec une probabilité de $0,98$. Calculez les probabilités suivantes :
p_0 : pour qu'une pièce soit mauvaise et acceptée,

p_1 : pour qu'il y ait une erreur dans le contrôle,
p_2 : pour qu'une pièce soit acceptée,
p_3 : pour qu'une pièce soit mauvaise, sachant qu'elle est acceptée.

Exercice 60 Trois machines A, B et C produisent respectivement 60%, 30% et 10% de la production des pièces d'une entreprise. La machine A (respectivement B et C) produit 2% (respectivement 3% et 4%) d'objets défectueux.

1. On choisit une pièce au hasard à la sortie de l'usine. Calculez la probabilité de l'événement : "La pièce est défectueuse".
2. On choisit une pièce au hasard à la sortie de l'usine et on voit qu'elle est défectueuse. Calculez la probabilité de l'événement : "Cette pièce a été fabriquée par la machine B".

Exercice 61 Monsieur et Madame A ont quatre enfants. On suppose que la probabilité de naissance d'un garçon est la même que celle d'une fille. Calculez la probabilité des événements suivants :
A : "Monsieur et Madame A ont quatre filles",
B : "Monsieur et Madame A ont trois filles et un garçon",
C : "Monsieur et Madame A ont deux filles et deux garçons",
D : "Monsieur et Madame A n'ont pas de fille",
E : "Monsieur et Madame A ont au moins une fille",
F : "Monsieur et Madame A ont au moins une fille et un garçon".

Exercice 62 Soit $(\Omega, \mathcal{A}, p)$ un espace de probabilité et soient $A \in \mathcal{A}$ et $B \in \mathcal{A}$ tels que

$$p(A) = \frac{1}{3}, \quad p(B) = \frac{1}{2}, \quad p(A \cap B) = \frac{1}{5}.$$

Calculez $p(A \cup B)$, $p(\overline{A})$, $p(\overline{B})$, $p(\overline{A} \cap B)$, $p(\overline{A} \cup B)$, $p(\overline{A} \cap \overline{B})$.

Exercice 63 Une urne contient cinq boules, trois rouges, numérotées 1, 2, 3 et deux noires, numérotées 1 et 2. On tire au hasard et simultanément deux boules de cette urne. Les tirages sont équiprobables.

1. Quelle est la probabilité de l'événement A : "les deux boules tirées sont de la même couleur " ?
2. Quelle est la probabilité de l'événement B : "la somme des numéros portés sur chacune des deux boules tirées est égale à 3 " ?
3. Quelle est la probabilité de B sachant que A est réalisé ?

Exercice 64 Une urne A contient 4 boules rouges et 2 boules bleues. Une urne B contient 5 boules rouges et 6 boules bleues et une urne C contient 1 boule rouge et 9 boules bleues. On jette un dé parfait numéroté de 1 à 6.

- Si le résultat est impair, on tire au hasard une boule de A.
- Si le résultat est '2' ou '4', on tire au hasard une boule de B.
- Si le résultat est '6', on tire au hasard une boule de C.

Sachant que la boule tirée est bleue, quelle est la probabilité qu'elle ait été tirée de l'urne C ?

Exercice 65 Dans un jeu de 52 cartes, on choisit simultanément 3 cartes.

1. Déterminez le nombre de tirages possibles.
2. Calculez la probabilité des événements suivants :
   A : "le tirage contient le roi de coeur",
   B : "le tirage contient un roi exactement",
   C : "le tirage contient un coeur exactement",
   D : "le tirage contient deux coeurs dont le roi",
   E : "le tirage contient au moins un coeur" (pensez au complémentaire),

F : “le tirage contient exactement deux coeurs et exactement un roi”.

Exercice 66 On tire 3 boules d'un sac contenant 9 boules : 4 vertes, 3 rouges, 1 blanche et 1 noire

1. successivement avec remise. Calculez la probabilité des événements suivants :
   A : “le tirage contient 3 boules vertes”,
   B : “le tirage ne contient aucune boule rouge”,
   C : “le tirage contient 3 boules blanches”,
   D : “le tirage contient dans cet ordre : 2 boules vertes et 1 boule rouge”,
   E : “le tirage contient 2 vertes et 1 rouge”,
   F : “le tirage contient 1 verte, 1 rouge et 1 noire”.
2. Mêmes questions si le tirage se fait successivement sans remise.
3. Mêmes questions si le tirage se fait simultanément.