

Guy DOTREMONT

Cours de mathématique

5^e TQ

2 périodes par semaine

2023 – 2024

Le syllabus que vous tenez entre les mains est destiné aux élèves du troisième degré de l'enseignement secondaire technique de qualification de l'Institut Émile Gryzon à Anderlecht et est basé sur le programme du CPEONS qui prévoit l'enseignement de quatre UAA (Unités d'Acquis d'Apprentissage) sur deux années, réparties comme suit :

	5 ^e	6 ^e
UAA 1 : Approche graphique d'une fonction	●	-
UAA 2 : Modèles de croissance (2 parties)	●	●
UAA 3 : Statistique	●	-
UAA 4 : Probabilité	-	●

Ce cours n'est qu'une base de travail et ne se substitue en aucun cas au cours donné en classe.

En cas de difficultés, je suis là pour vous aider. Mais ce ne sera possible que si, de votre côté, vous avez déjà fait des efforts personnels. Dans le cas contraire, toute tentative de ma part serait vouée à l'échec.

N'oubliez pas une chose importante : vous êtes arrivé au troisième degré et la matière est toujours plus conséquente. Un travail régulier est important et est gage d'une probabilité de réussite supérieure, les statistiques sont là pour le (dé)montrer...

Je vous invite à visiter régulièrement mon site, créé pour vous :

<https://ieg.gdotremont.be>



Pour me joindre : **gdotremont@spfb.edu.brussels**

J'espère que ces deux années se passeront le mieux possible. Bon travail !

Table des matières

UAA 3 : Statistique	9
1 Introduction	9
1.1 Qu'est-ce que la statistique ?	9
1.2 Bref historique	10
2 Terminologie et signe de sommation	11
2.1 Terminologie	11
2.1.1 Exemple – Auberge de jeunesse	11
2.2 Le signe de sommation	13
2.3 Exercices	14
3 Statistiques à une variable : tableaux et graphiques	17
3.1 Série statistique	17
3.2 Tableau brut	17
3.2.1 Activité – Loisirs	18
3.3 Tableau recensé	18
3.3.1 Série discrète à caractère qualitatif	18
3.3.2 Série discrète à caractère quantitatif	19
3.3.3 Série classée	21
3.4 Graphiques	24
3.4.1 Histogramme	25
3.4.2 Polygone	26
3.4.3 Attention aux échelles !	27
3.5 Exercices	30

4	Statistiques à une variable : paramètres de position	37
4.1	Mode	37
4.2	Moyenne	39
4.3	Médiane	41
4.4	Exercices	43
5	Statistiques à une variable : paramètres de dispersion	46
5.1	Activité – Résultats scolaires	46
5.2	Variance et écart-type	47
5.2.1	La variance	47
5.2.2	L'écart-type	47
5.3	Quartiles, déciles et centiles	50
5.4	Étendue	50
5.5	Boîte à moustaches	51
5.6	Exercices	52
6	Statistiques à deux variables	57
6.1	Définition	57
6.2	Nuage de points	57
6.2.1	Exemple – Dépenses publicitaires	59
6.3	Corrélation ou causalité?	60
6.3.1	Un peu d'humour	60
6.4	Ajustement affine graphique	61
6.5	Droite de Mayer	62
6.6	Exercices	63
6.7	Régression linéaire – méthode des moindres carrés	67
6.8	Coefficient de corrélation	70
UAA 1	Approche graphique d'une fonction	72
1	Notion de graphe	72
1.1	Qu'est-ce qu'un graphe?	72
1.2	Que doit contenir un graphe?	73
1.3	Prenons un exemple	74

2	Notion de fonction	75
2.1	Activité – Panneau publicitaire	75
2.2	Activité – Portail électronique	77
2.3	Activité – Aller-retour	78
2.4	Exercices	81
3	Domaine et ensemble image d'une fonction	85
3.1	Activité – Logo	85
3.1.1	Les parties de $\mathbb{R}$	89
3.1.2	Domaine d'une fonction	90
3.1.3	Ensemble image d'une fonction	90
3.2	Activité – Grue de chantier	91
3.3	Activité – Éolienne	92
4	Intersections de la fonction avec les axes	93
4.1	Activité – Fonction & intersections	93
4.2	Exercices	95
5	Image et antécédents	98
5.1	Activité – Évolution de températures (1)	98
5.2	Exercices	100
6	Signe d'une fonction	102
6.1	Activité – Évolution de températures (2)	102
6.1.1	Le signe d'une fonction, en résumé	103
6.2	Exercices	104
7	Variation d'une fonction	106
7.1	Activité – Taux d'alcoolémie	106
7.2	Exercices	108
8	Résolution graphique d'équations et d'inéquations	113
8.1	Activité – Bénéfices	113
8.2	Activité – Scooters	114
8.3	En règle générale	116
8.3.1	Droite parallèle à l'axe des abscisses	116

8.3.2	Fonction affine	118
8.3.3	Fonction quelconque	120
8.3.4	Et en cas d'inéquation ?	122
8.4	Exercices	124
9	Synthèse de l'UAA 1	128
UAA 2	: Modèles de croissance, partie 5 TQ	139
1	Rappels	139
2	Constructions...	140
2.1	Construction d'un tableau de nombres à partir d'une formule	140
2.1.1	Activité – Distance de freinage (1)	140
2.2	Construction d'un graphe à partir d'un tableau de nombres	141
2.2.1	Activité – Distance de freinage (2)	141
2.3	Exercices	142
3	Les fonctions de référence	143
3.1	Fonction constante	143
3.1.1	Définition	143
3.1.2	Représentation graphique	143
3.2	Fonction linéaire	144
3.2.1	Définition	144
3.2.2	Représentation graphique	144
3.3	Fonction affine	145
3.3.1	Définition	145
3.3.2	Représentation graphique	145
3.4	La fonction de degré 2 : la parabole	147
3.4.1	Définition	147
3.4.2	Représentation graphique	147
3.4.3	Résolution	149
3.4.4	Démonstration – pour information	150
3.4.5	Applications	151
3.5	La fonction de degré 3	152

3.5.1	Définition	152
3.5.2	Représentation graphique	152
3.6	La fonction inverse	154
3.6.1	Définition	154
3.6.2	Représentation graphique	154
3.7	La fonction racine carrée	155
3.7.1	Définition	155
3.7.2	Représentation graphique	155
3.8	La fonction puissance	156
3.8.1	Définition	156
3.8.2	Représentation graphique	156
3.9	Exercices	157

UAA 3

Statistique

1.1 Qu'est-ce que la statistique ?

La statistique, au sens général du terme, s'est invitée progressivement dans notre vie quotidienne. Elle est partout ! Et pourtant, bien souvent, on entend des commentaires du genre *La stat, c'est trop compliqué*, ou encore *La stat, ça sert à rien ; ça nous servira jamais*. Détrompez-vous, cher lecteur. Elle est vraiment partout et indispensable ; on la retrouve aussi bien dans les quotidiens que dans les revues spécialisées en économie sous la forme de tableaux, de graphes,... issus de sondages ou autres analyses de valeurs.

On utilise la statistique dans des domaines aussi divers que la démographie, la météorologie, la sociologie, la physique, la médecine, ou encore les assurances,...

Il faut savoir que la statistique est un outil mathématique. Il ne s'agit, en effet, que d'un outil, un langage qui va permettre de faire passer un message. Et ce message, c'est vous qui allez d'abord le créer et le faire passer ensuite en transformant les résultats recueillis en explications claires.

Vous allez donc recueillir des données (les statistiques), les organiser et les présenter de façon à pouvoir les analyser (avec les outils adéquats) et en tirer des informations permettant de gérer, de prévoir,...

La statistique permet donc de décrire simplement la complexité de la réalité en la synthétisant et en la résumant.

Nous pouvons donc définir la **statistique** comme étant *l'ensemble des outils et procédures qui permettent de réorganiser un ensemble de données, de les décrire et de les résumer*.

1.2 Bref historique

Je dis plus haut dans cette introduction que la statistique a envahi progressivement notre quotidien. Mais de quand date l'apparition de la statistique ? Il est difficile de le dire de manière précise mais, selon le dictionnaire Le Robert, cela viendrait de l'italien « *statista* » (homme d'État), famille du latin « *status* » qui signifie « État ».

Ce que l'on sait c'est que la statistique existe dès la naissance des premières structures sociales. Et Wikipedia d'ajouter : « [...] Les premiers textes écrits retrouvés sont des recensements de bétail, des informations sur son cours et des contrats divers. On a ainsi tracé des recensements en Chine ou en Égypte au ^{xviii}^e siècle Av. J.C. [...] Ce n'est qu'au ^{xviii}^e siècle Apr. J.C. ⁽¹⁾ que l'on voit apparaître le rôle prévisionnel des statistiques avec la construction des premières tables de mortalité [...] » ⁽²⁾.

Au début du ^{xix}^e siècle, il y a eu le développement des probabilités et des statistiques comme on les connaît encore aujourd'hui, développement favorisé par l'essor des (super) calculateurs et autres ordinateurs.

Qui sont ces gens à l'origine de la statistique ?

Dans beaucoup de disciplines, et la mathématique en fait partie, il est difficile de nommer des personnes qui en sont précisément à l'origine. Cependant, des noms reviennent souvent tels que Blaise Pascal ⁽³⁾ ou encore Adolphe Quetelet ⁽⁴⁾. D'autres sont moins connus, voire totalement inconnus du grand public tels que Ronald Aylmer Fisher ⁽⁵⁾ ou John Tukey ⁽⁶⁾ dont nous reparlerons plus loin dans le cours.

Depuis, bien des personnes ont contribué au développement de cette discipline. Une simple recherche sur Internet vous fournira des dizaines de noms.

(1). Soit environ 4000 ans plus tard...

(2). Wikipedia - page «Statistique» consultée le 15 février 2021.

(3). Blaise Pascal (1623-1662) est un mathématicien, physicien, inventeur, philosophe, moraliste et théologien français.

(4). Lambert Adolphe Jacques Quetelet (1796-1874) est un mathématicien, astronome, naturaliste et statisticien belge, précurseur de l'étude démographique et fondateur de l'observatoire royal de Belgique.

(5). Ronald Aylmer Fisher (1890-1962) est un biologiste et statisticien britannique.

(6). John Wilder Tukey (1915-2000) est l'un des plus importants statisticiens américains du siècle dernier.

Terminologie et signe de sommation

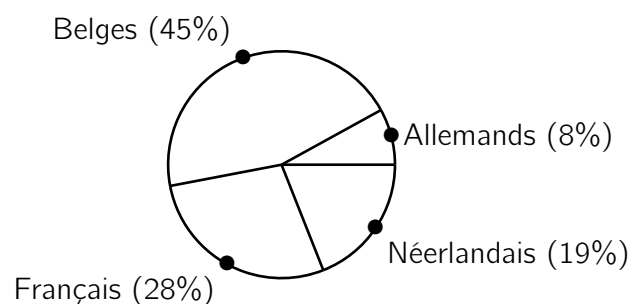
2.1 Terminologie

Pour pouvoir travailler la statistique (et les statistiques), il faut connaître un minimum de définitions, ce que j'appellerai la terminologie.

2.1.1 Exemple – Auberge de jeunesse

Une auberge de jeunesse effectue une étude sur sa fréquentation au mois d'août sur 200 personnes. Voici quelques résultats :

Nationalité des résidents



Durée du séjour

Nombre de jours	Effectif
1	56
2	83
3	43
4	18

Âge des résidents

Âge	Effectif
[18 ; 20 [	24
[20 ; 22 [	35
[22 ; 24 [	55
[24 ; 26 [	42
[26 ; 28 [	27
[28 ; 30 [	17

D'ores et déjà, nous constatons que l'on peut regrouper et classer des informations (ou des valeurs) comme cela nous intéresse et que l'on peut aussi les présenter de diverses manières (en tableau ou en graphique, par exemple).

Tout au long de ce cours, nous utiliserons divers termes qu'il est bon de connaître, il en va de la bonne compréhension de ce cours.

Voici donc ces quelques petites définitions à connaître.

Population Le terme de population (au sens mathématique et non géographique du terme !) sera utilisé pour désigner *l'ensemble des personnes, des êtres vivants ou des objets sur lequel porte l'étude statistique*.

Effectif total Le nombre total d'individus composant cette population est appelé effectif total de la population.

Caractère On appellera caractère *la propriété observée sur la population*.

Vous apprendrez également à réaliser des tableaux présentant les résultats observés, ceux-ci constituant une (ou des) série(s) statistique(s). Les séries statistiques peuvent être à caractère quantitatif ou qualitatif, discrètes ou classées et chronologiques ou non.

2.2 Le signe de sommation

Dans une série statistique, les différentes valeurs du caractères (les modalités) sont généralement notées $x_1, x_2, x_3, \dots, x_n$ (où 1, 2, 3, ..., n sont appelés « indices »).

Or, pour le calcul de plusieurs paramètres tels que la moyenne, par exemple, on sera amené à additionner ces valeurs.

Afin d'en simplifier la notation, on utilisera le symbole de sommation Σ (sigma) qui sera donc utilisé pour désigner de manière générale la somme de plusieurs termes.

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_n$$

Généralisons l'expression :

$$\sum_{i=m}^n x_i = x_m + x_{m+1} + x_{m+2} + \dots + x_n$$

avec $m \leq n$

Quelques exemples

$$\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

$$\sum_{i=3}^6 y_i = y_3 + y_4 + y_5 + y_6$$

$$\sum_{i=20}^{22} x_i^4 = x_{20}^4 + x_{21}^4 + x_{22}^4$$

$$\sum_{i=1}^4 3i = 3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3 + 3 \cdot 4 = 30$$

$$\begin{aligned} \sum_{k=3}^6 (x_k - 2) &= (x_3 - 2) + (x_4 - 2) + (x_5 - 2) + (x_6 - 2) \\ &= x_3 + x_4 + x_5 + x_6 - 8 \end{aligned}$$

2.3 Exercices

Remarque : Tous les exercices ne seront pas faits en classe. À vous de travailler également à la maison !

Exercice 1

Soit x une variable prenant les valeurs suivantes :

$$x_1 = 4 \quad x_2 = -2 \quad x_3 = 1 \quad x_4 = 0 \quad x_5 = -5$$

$$1. \quad \sum_{i=1}^5 x_i =$$

$$2. \quad \sum_{i=1}^4 x_i^2 =$$

$$3. \quad \sum_{i=1}^3 2x_i =$$

$$4. \quad \sum_{i=2}^4 (x_i - 2) =$$

$$5. \quad \sum_{i=1}^3 \frac{1}{x_i^2} =$$

$$6. \quad \sum_{i=3}^4 \frac{1}{x_i + x_{i+1}} =$$

$$7. \quad \sum_{i=1}^3 \frac{2x_i^2}{x_i} =$$

$$8. \quad \sum_{i=2}^3 \frac{-x_i^3}{2x_{i-1}} =$$

$$9. \quad \sum_i (x_i^2 - 3) =$$

$$10. \quad \sum_i (\sqrt{x_i^4} + x_i) =$$

Exercice 2

Soient x et y deux variables prenant les valeurs suivantes :

$$x_1 = 8 \quad x_2 = -2 \quad x_3 = 4 \quad x_4 = -5$$

$$y_1 = 2 \quad y_2 = -6 \quad y_3 = -1 \quad y_4 = 3$$

$$1. \quad \sum_{i=1}^4 x_i =$$

$$2. \quad \sum_{i=1}^4 y_i =$$

$$3. \quad \sum_{i=1}^4 x_i^2 =$$

$$4. \quad \sum_{i=1}^4 y_i^2 =$$

$$5. \quad \sum_{i=2}^4 (x_i + y_i) =$$

$$6. \quad \sum_i (x_i \cdot y_i) =$$

$$7. \quad \sum_i \left[(x_i + 1) \cdot (2 - y_i) \right] =$$

$$8. \quad \sum_i \frac{x_i}{y_i} =$$

$$9. \quad \left(\sum_i x_i^2 \right) \cdot \left(\sum_i y_i \right)^2 =$$

$$10. \quad \sum_i \frac{1}{x_i \cdot y_i} =$$

$$11. \quad \sum_i \frac{x_i}{y_i^2} =$$

Exercice 3

Soient x et y deux variables prenant les valeurs suivantes :

$$x_1 = -3 \quad x_2 = 9 \quad x_3 = -1 \quad x_4 = -6$$

$$n_1 = 2 \quad n_2 = 5 \quad n_3 = 0 \quad n_4 = 8$$

$$1. \quad \sum_i (x_i)^2 =$$

$$2. \quad \left(\sum_i x_i \right)^2 =$$

$$3. \quad \sum_i (x_i) \cdot \sum_i (n_i) =$$

$$4. \quad \sum_i (x_i n_i) =$$

Exercice 4

Exprime les sommes suivantes avec la notation de sommation :

$$1. \quad x_1 + x_2 + x_3 + x_4 + \cdots + x_{18} =$$

$$2. \quad \frac{x_1}{y_1} + \frac{x_2}{y_2} + \frac{x_3}{y_3} + \frac{x_4}{y_4} + \cdots + \frac{x_{100}}{y_{100}} =$$

$$3. \quad a_1^2 b_1 + a_2^2 b_2 + a_3^2 b_3 + a_4^2 b_4 + \cdots + a_9^2 b_9 =$$

$$4. \quad a_1^2 b_1 + a_2^2 b_1 + a_3^2 b_1 + a_4^2 b_1 + \cdots + a_9^2 b_1 =$$

$$5. \quad \frac{x_1}{y_1^2} + \frac{x_2}{y_2^2} + \frac{x_3}{y_3^2} + \frac{x_4}{y_4^2} + \cdots + \frac{x_{12}}{y_{12}^2} =$$

Statistiques à une variable : tableaux et graphiques

3.1 Série statistique

Je vous rappelle qu'une série statistique est l'ensemble des résultats d'une étude statistique. Par facilité, les séries statistiques sont, avant tout, représentées sous la forme d'un tableau. Il est bon de se souvenir qu'elles peuvent être à caractère qualitatif (couleur des yeux, métier pratiqué, loisirs préférés, prénom,...), quantitatif discret (taille, poids, moyenne générale au bulletin, quantité d'argent de poche, nombre d'amis sur un réseau social déterminé, température extérieure à 8h du matin,...) ou quantitatif classé.

3.2 Tableau brut

Le premier résultat de dépouillement de données est une suite de valeurs notées dans un tableau, au fur et à mesure que celles-ci se présentaient. Le tableau ainsi créé est appelé un tableau brut.

On peut considérer le tableau brut comme étant le tableau élémentaire de travail, celui qui servira de départ à toute étude statistique.

3.2.1 Activité – Loisirs

Dans une classe de 15 élèves, on a réalisé une enquête relative à leurs loisirs préférés. Les résultats ont été rapportés sur la liste alphabétique et donnent lieu au tableau brut ci-dessous :

Sport	T.V.	Lecture	Sortie	T.V.
T.V.	Sortie	Sortie	Shopping	Sortie
Lecture	Sortie	Lecture	Sortie	Lecture

Sur base du tableau brut ci-dessus, on demande :

- * Population :
- * Effectif total :
- * Caractère :
- * Modalités :

3.3 Tableau recensé

Pour pouvoir trier des informations d'un tableau brut, il est nécessaire d'établir tout d'abord un tableau résumé et clair, appelé tableau recensé.

3.3.1 Série discrète à caractère qualitatif

Reprenons l'exemple précédent.

Dans le tableau brut ci-dessus, on rencontre plusieurs fois une même **modalité**. Celle-ci sera notée x_i .

Le nombre de fois que la modalité x_i apparaît dans le tableau brut est appelé **effectif** de x_i et noté n_i .

La **fréquence** d'une modalité x_i , notée f_i , est le rapport $\frac{n_i}{n}$ (n étant l'effectif total). Dans ce cours, les fréquences seront généralement exprimées en pourcentage⁽¹⁾ et arrondies à trois décimales (arrondi « commercial »).

(1). Il est donc obligatoire de calculer la fréquence comme suit : $f_i = \frac{n_i}{n} \cdot 100$.

Complétons le tableau suivant :

Modalités (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)
Sport		
T.V.		
Lecture		
Shopping		
Sortie		
Somme :		

3.3.2 Série discrète à caractère quantitatif

Activité – L'interrogation

Lors d'une interrogation, les résultats ont été les suivants (les notes sont sur 10) :

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

Dans le cas d'une série discrète à caractère quantitatif, les **modalités** seront rangées par ordre **croissant**.

On définit ensuite les notions d'effectif x_i et de fréquence f_i d'une modalité de la même façon que précédemment. De plus, le nombre de données du tableau brut inférieures ou égales à x_i est appelé **effectif cumulé** de x_i et est noté N_i . L'effectif cumulé de x_i est donc égal à la somme des effectifs de toutes les modalités allant de x_1 à x_i .

La **fréquence cumulée** d'une modalité x_i , notée F_i , est le rapport $\frac{N_i}{n}$. Rappelons que n est l'effectif total et que la fréquence doit être multipliée par 100 pour l'avoir en pourcentage.

On obtient alors le tableau recensé suivant :

Modalités (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)	Effectifs cumulés (N_i)	Fréquences cumulées (%) (F_i)
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

3.3.3 Série classée

Activité – Spectacle

Le tableau suivant donne les âges des spectateurs relevés lors d'un spectacle dans une petite salle :

20	41	29	33	8	67	19	41	37	43
19	37	20	18	38	33	22	39	20	23
42	38	17	18	21	38	27	23	52	40
23	21	40	22	38	19	17	45	23	26
35	37	24	25	17	35	35	21	37	22
20	39	33	38	21	42	27	40	18	37
28	22	38	23	32	18	41	24	20	21
40	19	18	35	37	20	20	21	38	22
17	32	45	18	24	21	45	35	48	40
39	25	38	22	19	20	48	42	45	59
38	24	26	38	21	23	37	19	39	22
19	21	37	17	35	33	45	23	31	40
19	22	25	21	29	25	40	19	27	48

La réalisation du tableau recensé, effectué de la même manière que précédemment, donne lieu au tableau recensé de la page suivante.

Modalités (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)	Effectifs cumulés (N_i)	Fréquences cumulées (%) (F_i)
8	1	0,769	1	0,769
17	5	3,846	6	4,615
18	6	4,615	12	9,231
19	9	6,923	21	16,154
20	8	6,154	29	22,308
21	10	7,692	39	30,000
22	8	6,154	47	36,154
23	7	5,385	54	41,538
24	4	3,077	58	44,615
25	4	3,077	62	47,692
26	2	1,538	64	49,231
27	3	2,308	67	51,538
28	1	0,769	68	52,308
29	2	1,538	70	53,846
31	1	0,769	71	54,615
32	2	1,538	73	56,154
33	4	3,077	77	59,231
35	5	3,846	82	63,077
37	8	0,769	90	69,231
38	10	7,692	100	76,923
39	5	3,846	105	80,769
40	7	5,385	112	86,154
41	3	2,308	115	88,462
42	3	2,308	118	90,769
43	1	0,769	119	91,538
45	5	3,846	124	95,385
48	3	2,308	127	97,692
52	1	0,769	128	98,462
59	1	0,769	129	99,231
67	1	0,769	130	100,000

Une erreur s'est glissée dans ce tableau. La retrouverez-vous rapidement ?

Ce tableau recensé, trop détaillé, ne donne pas une vision globale du caractère relevé. Ainsi, dans ce cas, on peut regrouper les résultats dans des intervalles appelés classes ; on obtient alors la série classée suivante :

Classes	[5 ;15[	[15 ;25[	[25 ;35[	[35 ;45[	[45 ;55[	[55 ;75[
Effectifs						

Constatations :

- * Lorsque les résultats sont regroupés en classes, on ne connaît plus la valeur exacte des résultats relevés à l'intérieur de chaque classe.
- * Toutes les classes ne doivent pas nécessairement avoir la même « longueur ».

Nous avons déjà vu précédemment les notions d'effectifs, d'effectifs cumulés, de fréquences et de fréquences cumulées. Pour pouvoir réaliser un tableau recensé complet d'une série classée, il est bon d'y ajouter deux notions importantes :

- * l'**amplitude** (ou longueur) de la classe, que l'on calcule en faisant la différence entre la borne supérieure et la borne inférieure de la classe considérée ;
- * le **centre** de la classe est calculé en effectuant la moyenne arithmétique entre les bornes de la classe considérée.

On obtient alors le tableau recensé d'une série classée :

Classes						
[5 ;15[						
[15 ;25[						
[25 ;35[						
[35 ;45[						
[45 ;55[						
[55 ;75[						

Avant de passer aux exercices, nous rappellerons que le but d'une étude statistique est, non seulement, de réorganiser un ensemble de données et de les décrire mais surtout de les résumer.

Jusqu'à présent, nous nous sommes limités à (ré)organiser les données dans des tableaux. Nous apprendrons à les décrire (et les résumer) aux deux chapitres suivants grâce aux paramètres de position et de dispersion.

Nous allons maintenant voir une méthode qui permet de présenter les données sous une forme particulière : les graphiques.

3.4 Graphiques

Les résultats obtenus par une étude statistique peuvent être représentés sous forme de tableaux (recensés) mais également sous forme de graphiques. Ceux-ci sont très souvent des diagrammes à bâtonnets ou à barres (dits

aussi à rectangles) ou des diagrammes à secteurs. Ils représentent des données issues de séries discrètes. Pour décrire une série classée, les types de graphiques utilisés seront préférentiellement des histogrammes ou des polygones.

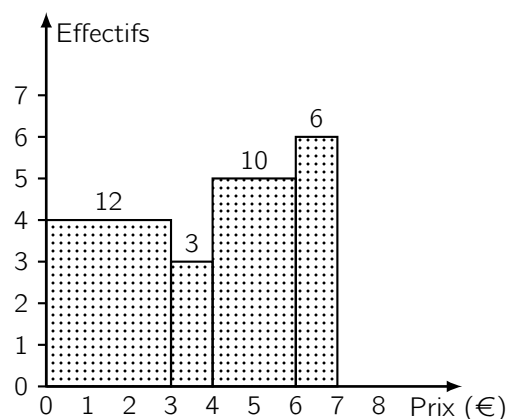
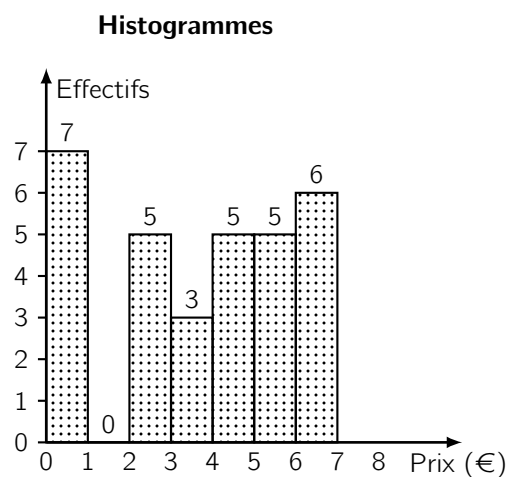
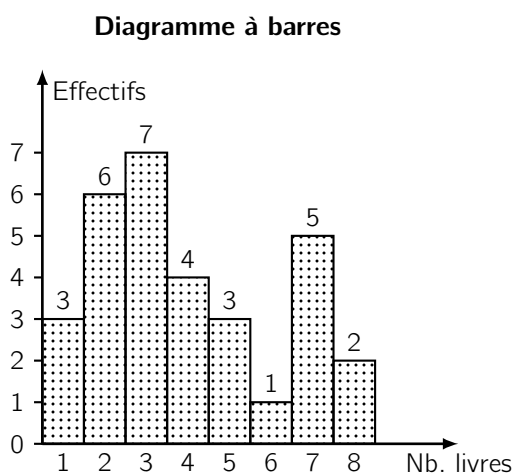
3.4.1 Histogramme

Avant d'aller plus en avant, il est bon de fixer les idées quant à la différence entre un diagramme à barres et un histogramme. Même si ceux-ci se ressemblent beaucoup, la règle utilisée pour les réaliser est différente.

Le principe utilisé dans la réalisation d'un **diagramme à barres** est la représentation des données en rectangles dont les bases sont de largeurs constantes et les hauteurs proportionnelles aux effectifs ou aux fréquences.

Quant à l'**histogramme**, le principe pour le réaliser paraît similaire mais il y a une différence fondamentale : les bases correspondent aux amplitudes des classes et les aires sont proportionnelles aux effectifs ou aux fréquences.

Voici des exemples pour les différencier :



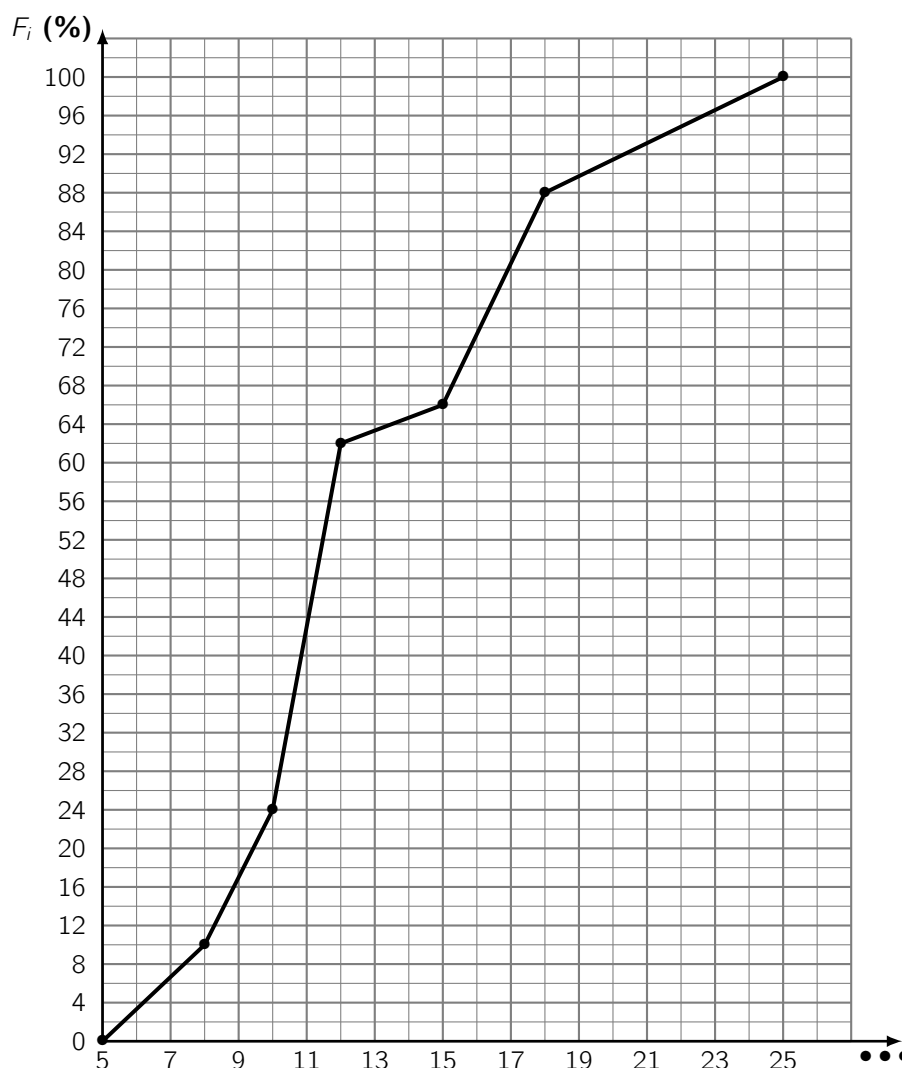
3.4.2 Polygone

Le principe pour réaliser un polygone, on supposera qu'au sein de chaque classe, tous les résultats sont répartis uniformément. On reliera donc, par des segments de droite, les points dont :

- * L'abscisse est la limite inférieure d'une classe.
- * L'ordonnée est l'effectif cumulé, ou la fréquence cumulée, de cette classe.

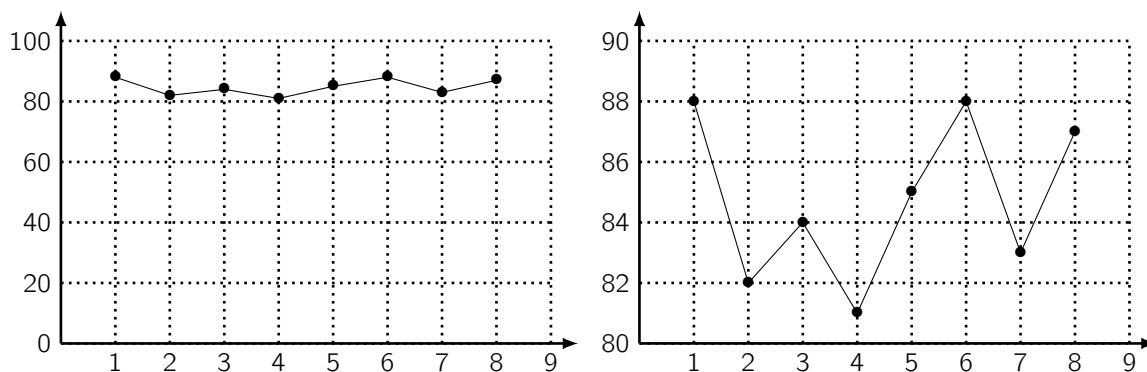
Il est à remarquer que la classe précédant la première classe étant inexistante, l'origine du segment de droite de la première classe aura donc pour origine 0 (zéro), sur l'ordonnée.

Exemple de polygone (ici, de fréquences cumulées)

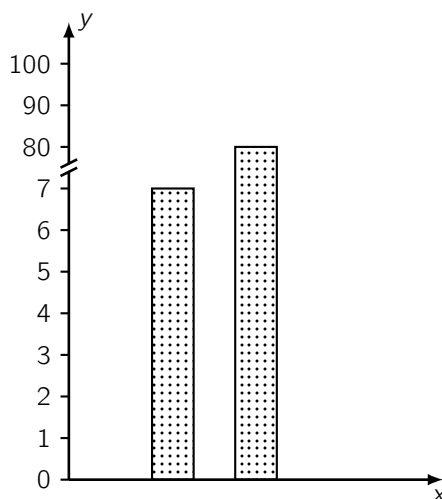


3.4.3 Attention aux échelles !

Lorsque vous êtes face à un graphique, son importance ne se limite pas à son contenu mais à l'ensemble du graphique, axes compris. En effet, dans les deux graphes ci-dessous, ce sont les mêmes données qui y sont représentées mais le graphe de droite nous donne l'impression d'une plus grande fluctuation des valeurs.

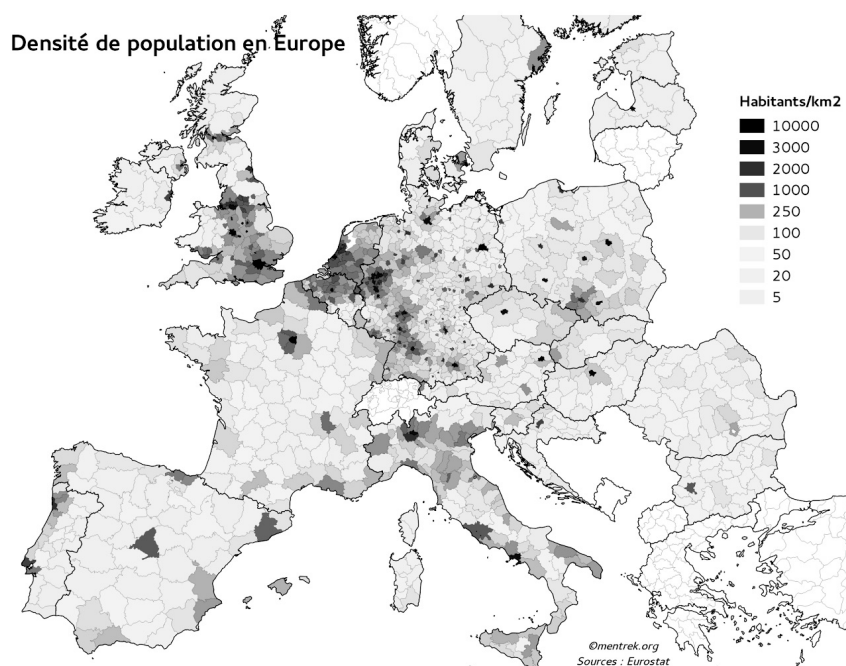


Le graphe ci-dessous nous trompe également : la différence entre les deux valeurs nous paraît insignifiante et pourtant elle est gigantesque.



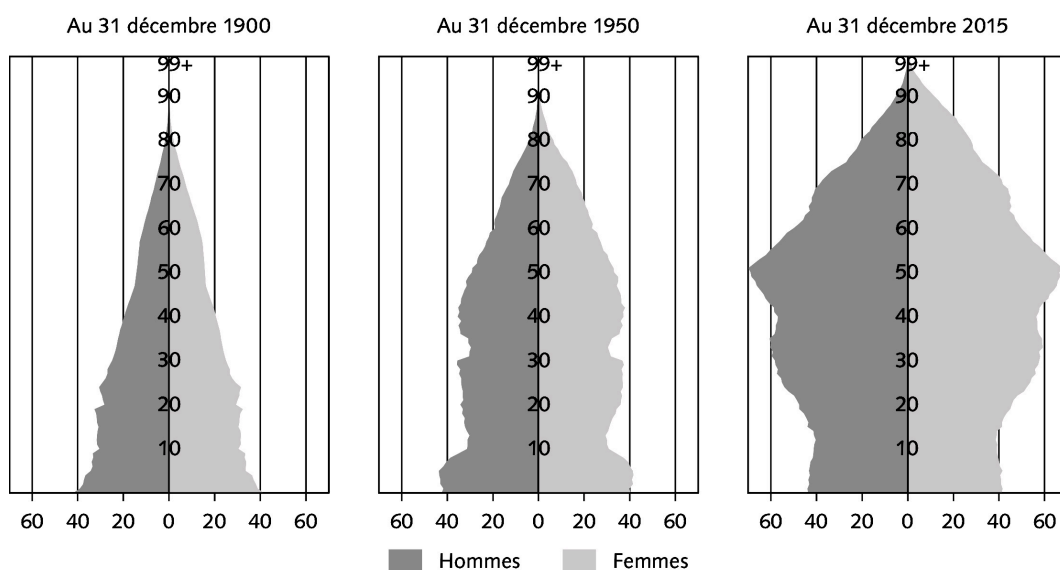
Il existe d'autres types de représentation graphique, généralement plus adaptés au sujet traité : la pyramide des âges, l'indice de masse corporelle, la densité de population, la production de lait par région,...

Voici quelques représentations graphiques afin d'en montrer la diversité.



Pyramide des âges

Nombre de personnes en milliers



Sources : OFS – RFP, STATPOP

© OFS, Neuchâtel 2016

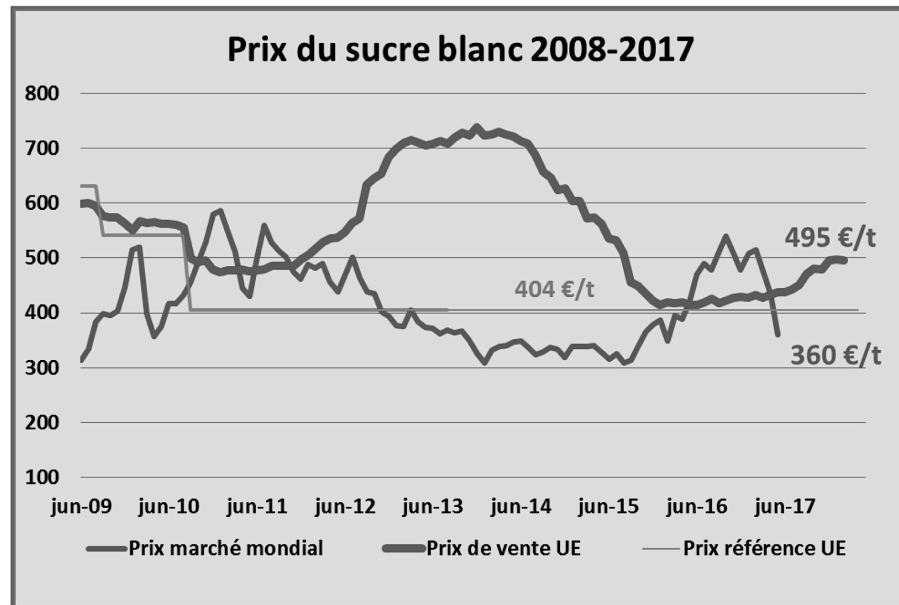
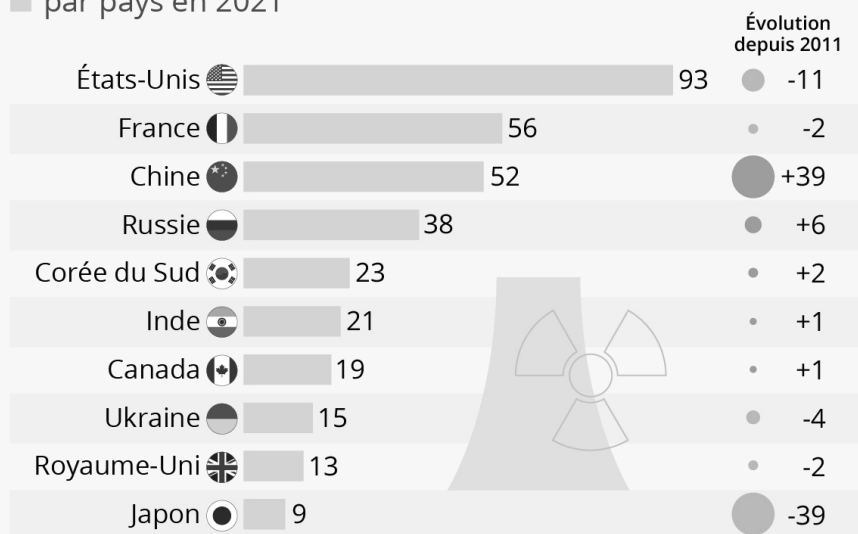


Tableau des jauges
Ventes régionales



Les pays avec le plus de réacteurs nucléaires

Nombre de réacteurs nucléaires en exploitation par pays en 2021



Source : World Nuclear Industry Status Report 2021



statista

3.5 Exercices

Remarque : Pour l'ensemble des exercices, toute valeur comportant des décimales sera arrondie au centième près (arrondi commercial à 2 décimales).

Exercice 1

Le tableau suivant donne la population belge au 1^{er} janvier 2021 ^a.

Régions	Nombre d'habitants
Bruxelles	1 214 550
Flandres	6 647 506
Wallonie	3 645 107

- Déterminer : population, caractère et modalités.
- Établir le tableau recensé (x_i , n_i , f_i %) correspondant.

a. Source : https://www.ibz.rn.fgov.be/fileadmin/user_upload/fr/pop/statistiques/population-bevolking-20210101.pdf

Exercice 2

En 2015, le nombre de travailleurs (tous secteurs confondus) en Belgique était de 4 544 milliers de personnes. Leur répartition était la suivante :

Secteurs	?
Secteur primaire (agriculture)	1,3 %
Secteur secondaire (industrie)	21,0 %
Secteur tertiaire (commerce, transport, finances, ...)	42,3 %
Secteur quaternaire (soins de santé, enseignement, ...)	35,4 %

- Déterminer : population, caractère et modalités.
- Établir le tableau recensé (x_i , n_i , f_i %) correspondant.

Exercice 3

Le tableau suivant donne la répartition de 125 agences de voyage selon leur nombre d'employés.

Nb. d'employés	Nb. d'agences	Fréquence (%)
1	22	?
2	?	28,0
3	45	?
4	?	14,4
5	5	?

- Déterminer : population, caractère et modalités.
- Établir le tableau recensé (x_i , n_i , f_i %, N_i , F_i %) correspondant sans oublier d'y indiquer le nom des colonnes.
- Quel est le pourcentage d'agences occupant :
 - * exactement 3 employés ?
 - * au plus 4 employés ?
 - * moins de 3 employés ?
 - * au moins 4 employés ?

Exercice 4

Lors d'un sondage, il a été demandé à un panel de jeunes de 18 à 20 ans la durée (en minutes) qu'ils passaient quotidiennement sur les réseaux sociaux.

Les résultats sont centralisés dans le tableau suivant.

Durée (min)	Nb. élèves
[0 ;30[	37
[30 ;60[	80
[60 ;90[	220
[90 ;120[	218
[120 ;150[	130
[150 ;180[	74
[180 ;210[	182
[210 ;240[	46
[240 ;270[	7

1. Déterminer : population, effectif total et caractère.
2. Établir le tableau recensé (Classes, x_i , n_i , f_i %, N_i , F_i %) correspondant sans oublier d'y indiquer le nom des colonnes.

Exercice 5

On a relevé le montant de 240 chèques remis à un guichet d'une agence bancaire :

Montant des chèques (€)	Effectif
$[0 ; 500[$	23
$[500 ; 1000[$	117
$[1000 ; 1500[$	59
$[1500 ; 2000[$	41

- Déterminer : population et caractère.
- Établir le tableau recensé (Classes, x_i , n_i , f_i %, N_i , F_i %) correspondant sans oublier d'y indiquer le nom des colonnes.
- Quel est le pourcentage de chèques dont le montant est :
 - * inférieur à 1.500 € ?
 - * supérieur à 1.500 € ?
 - * inférieur à 1.200 € ?
 - * inférieur à 800 € ?
 - * supérieur à 650 € ?
- Tracer le polygone des effectifs cumulés.

Exercice 6

Le tableau suivant donne le montant d'achats effectués sur un site d'e-commerce durant un week-end.

Montant (€)	Effectifs	Fréquences (%)	Effectifs cumulés
[0 ;250[	28		
[250 ;1000[		66	
[1000 ;1500[			324
[1500 ;2500[			400

- Déterminer : type de série, population et caractère.
- Établir le tableau recensé complet (Classes, amplitude, x_i , n_i , f_i %, N_i , F_i %) correspondant sans oublier d'y indiquer le nom des colonnes.
- Tracer le polygone des effectifs cumulés.
- Déterminer graphiquement le nombre d'achats dont le montant est :
 - * inférieur à 1500 € ;
 - * supérieur à 1250 € ;
 - * compris entre 250 € et 1500 € ;
- Vérifier, par calcul, les résultats ci-dessus.

Exercice 7

Le tableau suivant donne la taille d'adolescents se présentant à un casting pour une publicité.

Taille (cm)	Effectif
[160 ;163[	1
[163 ;166[	9
[166 ;169[	33
[169 ;172[	50
[172 ;175[	45
[175 ;178[	10
[178 ;190[	2

- Déterminer : population et caractère.
- Établir le tableau recensé (Classes, x_i , n_i , f_i %, N_i , F_i %) correspondant sans oublier d'y indiquer le nom des colonnes.
- Déterminer le pourcentage d'adolescents mesurant :
 - * moins de 175 cm ;
 - * plus de 169 cm ;
 - * moins de 165 cm ;
 - * plus de 173 cm.
- Tracer le polygone des fréquences cumulées.

Exercice 8

Soit le graphique suivant :

RÉSERVES PROUVÉES DE PÉTROLE BRUT DANS LE MONDE

EN TONNES, AU 1ER JANVIER 2019



Source : lafinancepourtous.com d'après INSEE

1. Déterminer : population, caractère et modalités.
2. Établir le tableau recensé $(x_i, n_i, f_i \%)$ correspondant.
3. Le tableau recensé pourrait-il contenir les colonnes des effectifs cumulés et des fréquences cumulées ? Justifier.

Exercice 9

Le tableau suivant classe la répartition des recettes (en pourcentage) de la Sécurité Sociale en France, en 1998 :

Sujet	Fréquence
Cotisations	68,2 %
Impôts et taxes	20,5 %
Autres (recettes du patrimoine immobilier, subventions État, ...)	11,3 %

Construire le diagramme à secteurs (des fréquences) correspondant.

Statistiques à une variable : paramètres de position

Les paramètres de position permettent de préciser autour de quelle valeur sont dispersées les données d'une série statistique. Ces paramètres sont au nombre de trois : le mode, la moyenne et la médiane.

4.1 Mode

* Série discrète à caractère quantitatif

- On appelle *mode*, toute modalité dont l'effectif est maximal.

* Série classée

- On appelle *classe modale*, toute classe dont l'effectif est maximal.
- On appelle *mode*, le centre d'une classe modale.

Le mode (Mo) est donc la valeur la plus fréquente, c'est-à-dire la valeur obtenue par le plus grand nombre de sujets.

Il est à remarquer qu'une série statistique peut posséder un ou plusieurs modes ; on dira alors qu'elle est unimodale, bimodale, trimodale,...

Exemples

* Série discrète à caractère quantitatif

Voici l'extrait d'un tableau (déjà vu p.20) :

Modalités (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)
1	3	2,5
2	4	3,333
3	9	7,5
4	22	18,333
5	22	18,333
6	23	19,167
7	21	17,5
8	13	10,833
9	2	1,667
10	1	0,833

→ $M_o =$

* Série classée :

Voici l'extrait d'un autre tableau, déjà vu également (p.24)

Classes	Effectifs (n_i)	Fréquences (%) (f_i)
[5 ;15[	1	0,769
[15 ;25[	57	43,846
[25 ;35[	19	14,615
[35 ;45[	42	32,308
[45 ;55[	9	6,923
[55 ;75[	2	1,538

→ Classe modale :

→ $M_o =$

4.2 Moyenne

Dans une série statistique (discrète ou classée), la moyenne peut se calculer de deux manières différentes :

* Soit en utilisant les effectifs : $\bar{x} = \frac{\sum_i x_i n_i}{n}$

* Soit en utilisant les fréquences : $\bar{x} = \frac{\sum_i x_i f_i}{100}$

Il est grandement recommandé d'utiliser la formule prenant en compte des valeurs fournies dans l'énoncé et non pas celles calculées.

Exemples

* **Série discrète à caractère quantitatif**

Voici l'extrait d'un tableau (déjà vu p.20) :

Modalités (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)
1	3	2,5
2	4	3,333
3	9	7,5
4	22	18,333
5	22	18,333
6	23	19,167
7	21	17,5
8	13	10,833
9	2	1,667
10	1	0,833

→ $\bar{x} =$

* **Série classée**

Voici l'extrait d'un autre tableau, déjà vu également (p.24)

Classes	Centres (x_i)	Effectifs (n_i)	Fréquences (%) (f_i)
[5 ;15[	10	1	0,769
[15 ;25[	20	57	43,846
[25 ;35[	30	19	14,615
[35 ;45[	40	42	32,308
[45 ;55[	50	9	6,923
[55 ;75[	65	2	1,538

→ $\bar{x} =$

Remarque

La moyenne est un paramètre très souvent utilisé car elle est unique (contrairement au mode) et elle tient compte de tous les résultats (contrairement au mode, également), mais cela peut parfois être un inconvénient.

4.3 Médiane

La médiane (Me) est le résultat qui correspond à un effectif cumulé de $n/2$ ou à une fréquence cumulée de 50%. On appelle donc médiane la valeur centrale de la série statistique. Il s'agit de la valeur qui sépare l'ensemble des données ordonnées en deux groupes de même effectif.

Exemples

* **Série discrète à caractère quantitatif :**

Voici l'extrait d'un tableau (déjà vu p.20) :

x_i	n_i	N_i
1	3	3
2	4	7
3	9	16
4	22	38
5	22	60
6	23	83
7	21	104
8	13	117
9	2	119
10	1	120

Le principe est le suivant : on recherche $n/2$ dans la colonne des effectifs cumulés, et la médiane est la modalité correspondante.

→ $Me =$

Remarque : dans le cas où aucun effectif cumulé ne vaut $n/2$, on choisira l'effectif cumulé qui lui est directement supérieur (en valeur), et la médiane sera la modalité correspondante.

* **Série classée :**

Voici l'extrait d'un autre tableau, déjà vu également (p.24) :

Classes	Amplitudes	n_i	N_i
[5 ;15[	10	1	1
[15 ;25[	10	57	58
[25 ;35[	10	19	77
[35 ;45[	10	42	119
[45 ;55[	10	9	128
[55 ;75[	20	2	130

Le principe est le suivant : on va déterminer la classe médiane dans laquelle les N_i atteignent la valeur de $n/2$.

→ classe médiane :

Puis, dans la classe médiane, on calcule la valeur de la médiane par interpolation linéaire avec la formule suivante :

$$Me = \text{borne inf cl méd} + \frac{\text{ampl cl méd}}{\text{eff cl méd}} \cdot \left(\frac{n}{2} - \text{eff cumul cl précéd} \right)$$

Avec :

- * « borne inf cl méd » : Borne inférieure de la classe médiane ;
- * « ampl cl méd » : Amplitude de la classe médiane ;
- * « eff cl méd » : Effectif de la classe médiane ;
- * « eff cumul cl précéd » : Effectif cumulé de la classe précédente.

→ Me =

4.4 Exercices

Exercice 1

Une entreprise, spécialisée dans la rénovation intérieure d'appartements, désire faire quelques statistiques sur la durée de ses interventions. Les résultats de son enquête, réalisée sur deux mois, sont consignés dans le tableau suivant.

Durées (jours)	1	2	3	4	5
Effectifs	250	220	154	78	25

- Déterminer le mode puis calculer la moyenne et la médiane.

Exercice 2

Le tableau ci-dessous donne le nombre de buts marqués par match au cours du premier week-end du mois d'avril en division 1 nationale, provinciale et corporative.

Nb. buts	0	1	2	3	4	5	6	7	8	9	10	13
Nb. matches	4	20	35	23	15	20	6	6	2	1	1	1

Avant de se lancer dans les calculs demandés, je conseille de réaliser le tableau recensé (x_i , n_i , f_i %, N_i , F_i %).

1. Calculer $\bar{x}$, Mo et Me à partir des effectifs.
2. Calculer $\bar{x}$, Mo et Me à partir des fréquences.
3. Quelle conclusion peut-on en tirer ?

Exercice 3

Le tableau suivant donne la taille (en cm) de 150 adolescents se présentant à un casting pour une publicité.

Taille (cm)	$[160;163[$	$[163;166[$	$[166;169[$	$[169;172[$	$[172;175[$	$[175;178[$	$[178;190[$
Effectifs	1	9	33	50	45	10	2

1. Déterminer la classe modale et le mode.
2. Calculer $\bar{x}$.
3. Tracer le polygone des effectifs cumulés et y lire une valeur approximative de Me.
4. Calculer Me par interpolation linéaire.

Exercice 4

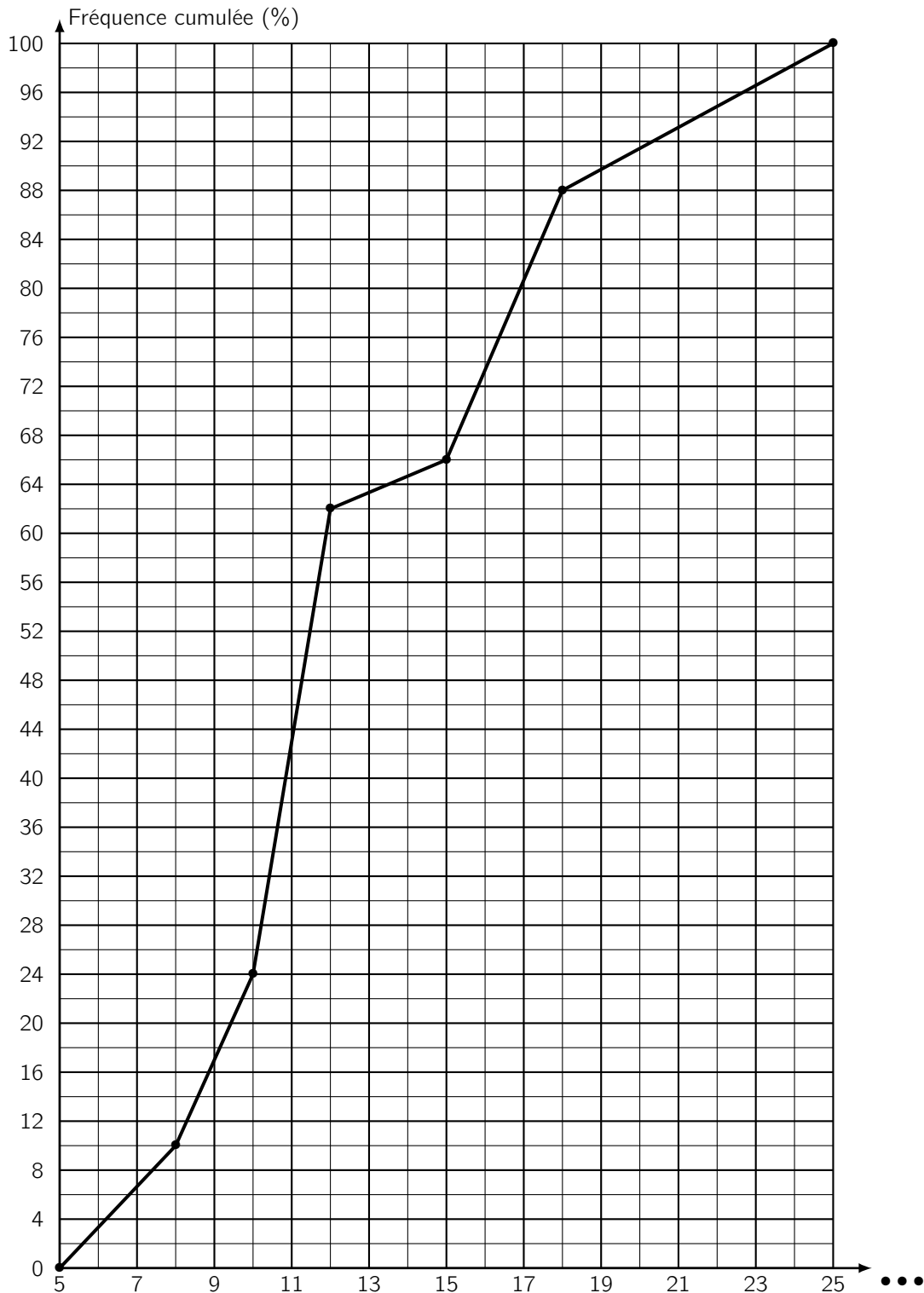
Le tableau suivant donne le salaire horaire des employés d'une entreprise.

Salaire (€)	$[10,00;10,25[$	$[10,25;10,50[$	$[10,50;10,75[$	$[10,75;11,00[$	$[11,00;11,25[$	$[11,25;11,50[$	$[11,50;11,75[$	$[11,75;12,00[$
Effectifs	6	8	16	76	36	16	7	2

1. Déterminer le mode et la moyenne.
2. Calculer Me par interpolation linéaire, à partir des effectifs et à partir des fréquences.
3. Vérifier la valeur obtenue ci-dessus sur un graphique.

Exercice 5

Le graphe suivant montre l'évolution d'une fréquence cumulée.



1. De quel type de graphe s'agit-il ?
2. Établir le tableau recensé (Classes, f_i %, F_i %) correspondant.
3. Calculer $\bar{x}$, Mo et Me.

Statistiques à une variable : paramètres de dispersion

5.1 Activité – Résultats scolaires

À la fin d'une année scolaire, un enseignant désire se faire une idée de l'ensemble des résultats obtenus par ses élèves. Dans ce but, il calcule la moyenne arithmétique des pourcentages. Mais cette moyenne suffit-elle, à elle seule, pour décrire de manière globale les résultats des élèves... ?

Dans la classe, il y a 6 élèves et les pourcentages acquis aux 1^{er} et 2^d semestres sont :

Semestre	Élève 1	Élève 2	Élève 3	Élève 4	Élève 5	Élève 6	Moyenne
N° 1	35 %	55 %	65 %	70 %	85 %	95 %	67,5 %
N° 2	59 %	61 %	64 %	66 %	75 %	80 %	67,5 %

Les moyennes arithmétiques sont donc égales, alors qu'il y a eu une évolution différente dans les résultats ! Ceux du 1^{er} semestre sont plus dispersés que ceux du 2^d semestre.

Cet exemple prouve la nécessité de caractériser une série statistique, non seulement par un paramètre de position, mais également par un paramètre de dispersion qui mesure le degré de confiance que l'on peut accorder au paramètre de position calculé.

5.2 Variance et écart-type

Si l'on désire étudier la dispersion des résultats autour de la moyenne, l'idée qui vient est de calculer l'écart moyen entre ces résultats et la moyenne.

On définit ainsi :

5.2.1 La variance

La variance correspond à la moyenne des carrés des différences entre les observations et leur moyenne.

Elle peut se calculer de deux manières différentes⁽¹⁾

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

ou

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

5.2.2 L'écart-type

Il s'interprète mieux que la variance, car il donne les unités utilisées dans l'échelle originale et non dans leur forme au carré.

$$\sigma = \sqrt{\sigma^2}$$

Reprenons le tableau du début de chapitre et calculons l'écart-type de chacun des semestres.

(1). Il ne sera pas fait ici la démonstration de l'équivalence des deux formules, cela sort du cadre de ce cours. Par contre, on apprendra à les utiliser.

Calcul de l'écart-type pour le premier semestre

Formule utilisée :

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

* Rappel, la moyenne est de 67,5 %

* On complète le tableau suivant :

x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
35	1			
55	1			
65	1			
70	1			
85	1			
95	1			

* Somme des $n_i (x_i - \bar{x})^2$:

* Calcul de σ^2 :

* Calcul de σ :

Calcul de l'écart-type pour le second semestre

Formule utilisée :

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

* Rappel, la moyenne est de 67,5 %

* On complète le tableau suivant :

x_i	n_i	x_i^2	$n_i x_i^2$
59	1		
61	1		
64	1		
66	1		
75	1		
80	1		

* Somme des $n_i x_i^2$:

* Calcul de σ^2 :

* Calcul de σ :

5.3 Quartiles, déciles et centiles

Nous avons vu que la médiane divise une série statistique ordonnée en deux parties de même effectif. De manière analogue, les quartiles divisent la série statistique en quatre parties de même effectif, les déciles la divisent en dix parties de même effectif et les centiles en cent parties de même effectif.

Rappelons-nous que la médiane (Me) est le résultat qui correspond à un effectif cumulé de $n/2$ ou à une fréquence cumulée de 50 %.

Sur le même principe :

- * Le 1^{er} quartile, noté Q_1 , est le résultat qui correspond à un effectif cumulé de $n/4$ ou à une fréquence cumulée de 25 %.
- * Le 3^e quartile, noté Q_3 , est le résultat qui correspond à un effectif cumulé de $3n/4$ ou à une fréquence cumulée de 75 %.

L'intervalle interquartile est l'intervalle $[Q_1; Q_3]$.

L'écart interquartile est l'amplitude de cet intervalle.

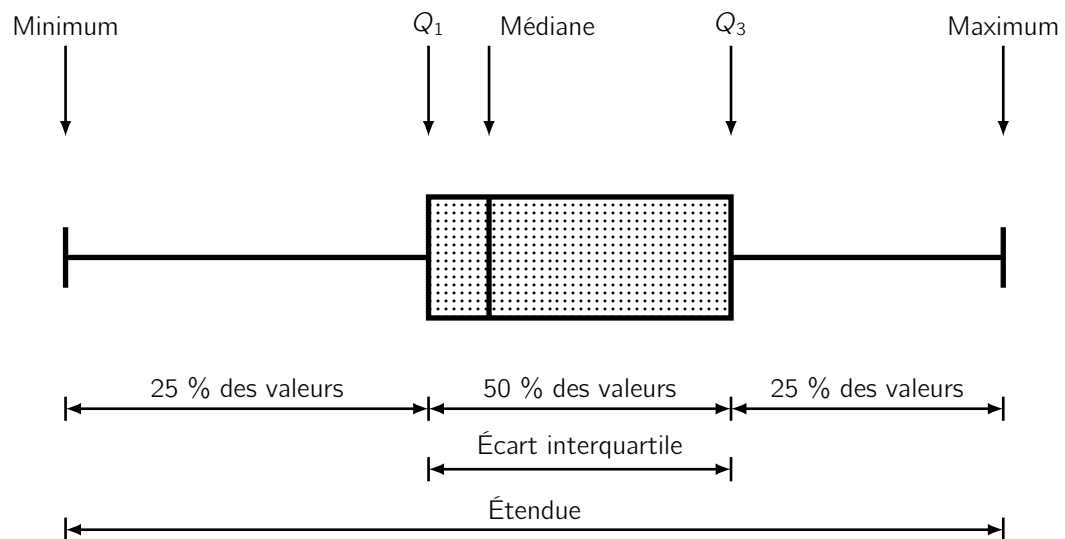
L'intervalle interquartile contient donc 50 % de l'effectif dont les valeurs sont les plus proches de la médiane. L'écart interquartile est donc une mesure de la dispersion des données autour de la médiane : plus il est grand, plus les données sont dispersées autour de la médiane ; plus il est petit, plus les données en sont proches.

5.4 Étendue

L'étendue d'une série statistique est la différence entre ses valeurs extrêmes.

5.5 Boite à moustaches

Les différentes caractéristiques vues jusqu'à présent peuvent être « résumées » par une boite à moustaches, aussi appelée *diagramme en boite*, ou encore *diagramme de Tukey*⁽²⁾, du nom de son concepteur.



(2). John Wilder Tukey (1915-2000) est un statisticien américain.

5.6 Exercices

Exercice 1

Lors d'une étude de la qualité des eaux destinées à la consommation, on analyse le contenu de soixante flacons dont on détermine la concentration en nitrates en mg/l.

Le tableau suivant résume les résultats de l'analyse :

Concentration (mg/l)	45,8	46,6	47,0	47,5	48,2	49,2	49,5	50,0	50,1	50,2
Nb. flacons	1	9	2	10	9	10	7	7	3	2

1. Déterminer le mode, la moyenne et l'écart-type.
2. Calculer la médiane.
3. Pour que l'eau soit propre à la consommation, sa concentration en nitrates doit être inférieure ou égale à 50 mg/l. Quel est le pourcentage de flacons contenant de l'eau impropre à la consommation ?

Exercice 2

Une auberge de jeunesse a effectué une étude sur sa fréquentation au mois d'août sur deux-cents personnes.

Le relevé des durées de séjour a donné les résultats suivants :

Nb jours	1	2	3	4	5
Fréquence (%)	28,0	31,5	29,0	7,5	4,0

Déterminer le mode, la moyenne, la médiane et l'écart-type.

Exercice 3

Le tableau suivant donne la répartition des voyagistes français (en 1990) selon leur taille en flux de vacanciers dans le marché métropolitain français. Cette étude ne concerne que les voyagistes ayant transporté plus de 15 000 personnes dans l'année.

Nb vacanciers (milliers)	$[15;25[$	$[25;50[$	$[50;75[$	$[75;100[$	$[100;150[$	$[150;200[$	$[200;500[$
Nb voyagistes	26	44	7	9	5	0	9

- Déterminer : population et caractère.
- Déterminer le pourcentage de voyagistes dont le flux de vacanciers fut :
 - * inférieur à 50 000
 - * compris entre 20 000 et 50 000
- Calculer $\bar{x}$ et σ
- Déterminer le mode.
- Calculer la médiane.
- Calculer les premier et troisième quartiles ; ensuite vérifier les valeurs obtenues pour Me , Q_1 et Q_3 sur un graphique.
- Calculer l'intervalle interquartile.
- Déterminer l'étendue de la série.
- Déterminer le pourcentage de voyagistes dont le flux de vacanciers fut supérieur à la moyenne.

Exercice 4

L'âge au 1^{er} janvier 1991 du personnel d'une entreprise liégeoise est repris dans le tableau suivant :

Âge des salariés	$[20;25[$	$[25;30[$	$[30;35[$	$[35;40[$	$[40;45[$	$[45;50[$	$[50;55[$	$[55;60[$	$[60;65[$
Hommes	10	26	44	52	38	20	17	7	0
Femmes	0	0	3	14	14	6	11	2	0

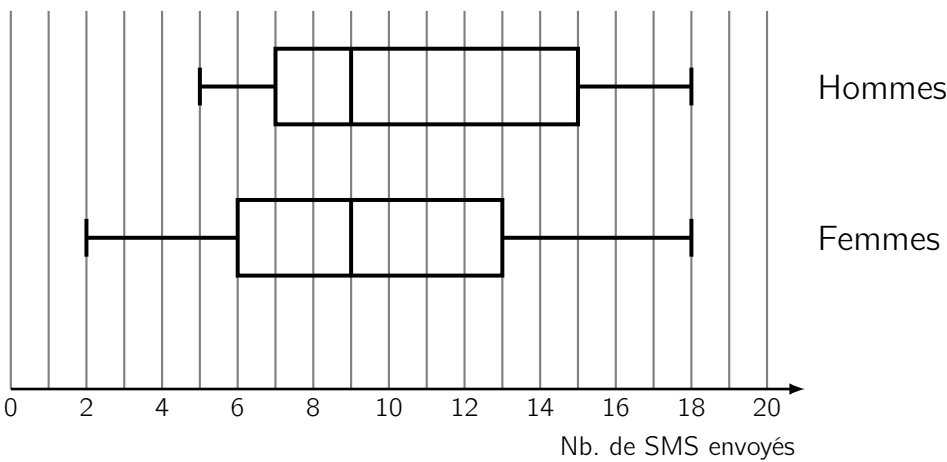
1. Déterminer Mo_H , Mo_F et Mo_{H+F} .
2. Calculer $\bar{x}_H$, $\bar{x}_F$ et $\bar{x}_{H+F}$.
3. Déterminer Me_H , Me_F et Me_{H+F} .
4. Calculer σ_H , σ_F et σ_{H+F} .
5. Calculer les intervalles interquartiles pour les H, les F et les H+F.

Exercice 5

Exercice sur deux pages !

On demande à un groupe de personnes le nombre de SMS envoyés en une journée.

Les résultats sont synthétisés par le diagramme ci-dessous :



1. Complète le tableau suivant :

	Hommes	Femmes
Valeur minimale		
Valeur maximale		
Valeur de la médiane		
Valeur de Q_1		
Valeur de Q_3		
Valeur de l'écart interquartile		
Valeur de l'étendue		

La suite de l'exercice se trouve page suivante...

2. Vrai ou faux ? Place une croix dans la case correspondante.

	VRAI	FAUX
La moitié des femmes a envoyé plus de 9 SMS en une journée.		
10 % des hommes ont envoyé entre 5 et 7 SMS.		
En moyenne, les hommes et les femmes ont envoyé le même nombre de SMS.		
La moitié des femmes a envoyé entre 6 et 13 SMS en une journée.		
50 % des hommes ont envoyé entre 9 et 18 SMS sur la journée.		
Un quart des hommes a envoyé entre 15 et 18 SMS.		
Un quart des femmes a envoyé entre 6 et 9 SMS.		

Statistiques à deux variables

6.1 Définition

On appelle série statistique à deux variables l'étude simultanée de deux caractères, comme par exemple la taille et le poids, les températures minimale et maximale,...

Il ne sera étudié dans ce cours que des séries statistiques dont les caractères seront quantitatifs.

6.2 Nuage de points

Les données obtenues lors d'expériences comportent souvent des erreurs dues à l'imprécision des mesures ou au processus expérimental. Lorsque les résultats sont portés en graphique, ils forment généralement ce que l'on appelle un nuage de points.

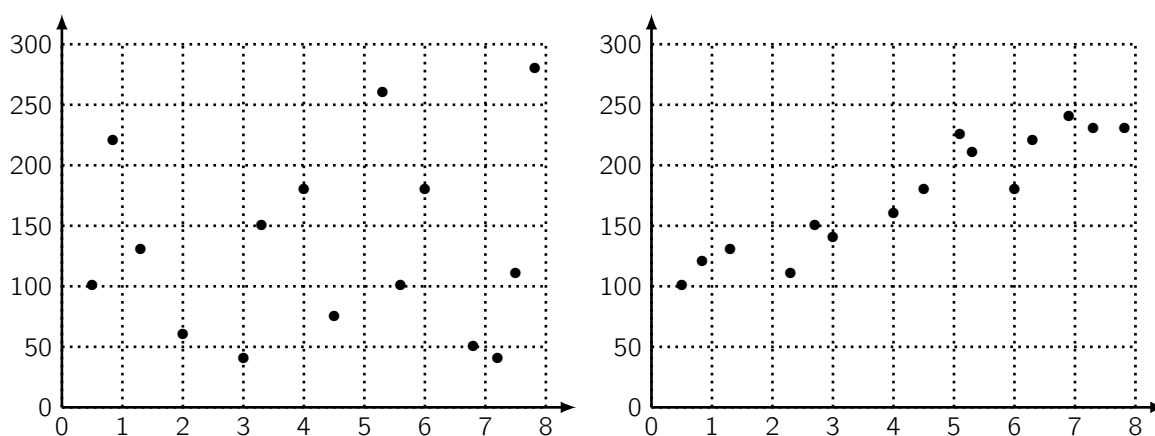
Le nuage de points est un graphique qui met en relation les valeurs de deux variables sur un repère de coordonnées cartésiennes.

Si, pour chacun des n individus d'une population, on note x_i et y_i les valeurs prises par les deux caractères, on peut alors les représenter sous la forme d'un tableau :

	M_1	M_2	M_3	...	M_n
Caractère x	x_1	x_2	x_3	...	x_n
Caractère y	y_1	y_2	y_3	...	y_n

En plaçant l'ensemble des points M_i de coordonnées $(x_i; y_i)$ dans un repère orthogonal, ceux-ci constituent un nuage de points, associé à la série statistique à deux variables, comme dans les exemples sur cette page.

Le nuage de points est un bon indicateur pour vérifier une corrélation entre les caractères x et y . Si les points sont sous la forme d'un nuage, il est fort probable que les phénomènes ne soient pas corrélés. S'ils semblent, au contraire, dessiner une courbe, on cherchera à déterminer la nature de la courbe en procédant à un ajustement.



Dans le cas où la courbe est une droite, ce qui nous intéresse dans la présente partie du cours, on dira que l'on réalise une *régression linéaire*, également appelée *ajustement des moindres carrés*.

Quel que soit le domaine dans lequel on va réaliser une expérience, l'exploitation de résultats expérimentaux nécessite très souvent l'emploi de la régression linéaire. Cet outil permet notamment de valider un modèle théorique grâce aux données expérimentales.

Puisque la droite est la fonction mathématique la plus simple à manipuler, on cherchera toujours à tracer graphiquement une grandeur Y en fonction d'une grandeur X de telle sorte que le graphe $Y = f(X)$ soit une droite, donc d'équation $Y = aX + b$.

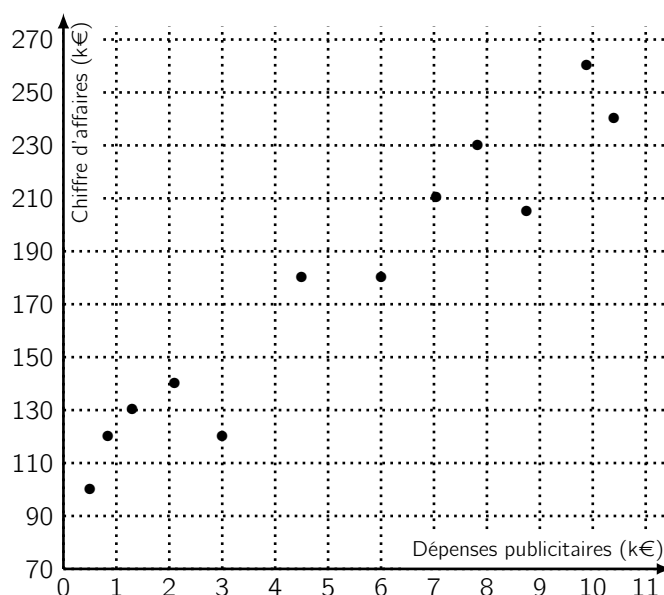
6.2.1 Exemple – Dépenses publicitaires

En tant qu'entrepreneur, je voudrais savoir s'il existe un lien entre mes dépenses mensuelles de publicité et mon chiffre d'affaires mensuel.

Je récolte donc les données mensuelles sur un an (en milliers d'euros) et les étudie :

Mois	Dépenses (k€)	Chiffre d'affaires (k€)
Janvier	7,04	210
Février	10,4	240
Mars	8,75	205
Avril	2,1	140
Mai	4,5	180
Juin	6	180
Juillet	3	120
Août	0,5	100
Septembre	0,84	120
Octobre	1,3	130
Novembre	7,82	230
Décembre	9,88	260

Vous conviendrez qu'un tel tableau ne nous apporte pas directement une réponse claire. Nous allons donc réaliser un graphique en forme de nuage de points du chiffre d'affaires en fonction des dépenses publicitaires, chacun exprimé en milliers d'euros.



6.3 Corrélation ou causalité ?

Il est important de ne pas confondre *corrélation* et *causalité*.

- Une corrélation est un lien, un rapport réciproque, entre deux variables statistiques.
- Une causalité est un lien qui s'établit entre une cause et son effet, c'est-à-dire une variable qui agit sur une autre.

Notons que la causalité implique toujours la corrélation alors que la corrélation n'implique pas la causalité. La nuance peut sembler ténue ; et pourtant elle est de taille :

- Le médecin suisse Franz Messerli montrait dans son article⁽¹⁾ cette corrélation entre la consommation de chocolat par habitant et par an et le nombre de prix Nobel pour 10 millions d'habitants. De là à dire que manger du chocolat rend les gens intelligents est un pas qu'il ne faut pas franchir. En effet, il y a une variable manquante : le niveau de vie. Celui-ci explique les dépenses en éducation et en recherche, augmentant ainsi les chances d'avoir des prix Nobel dans la population mais aussi le pouvoir d'achat qui permet de consommer du chocolat.
- Le lien entre une vitre brisée et un caillou qui la heurte est un lien de causalité : parce que le caillou a heurté la vitre, celle-ci s'est brisée.

6.3.1 Un peu d'humour



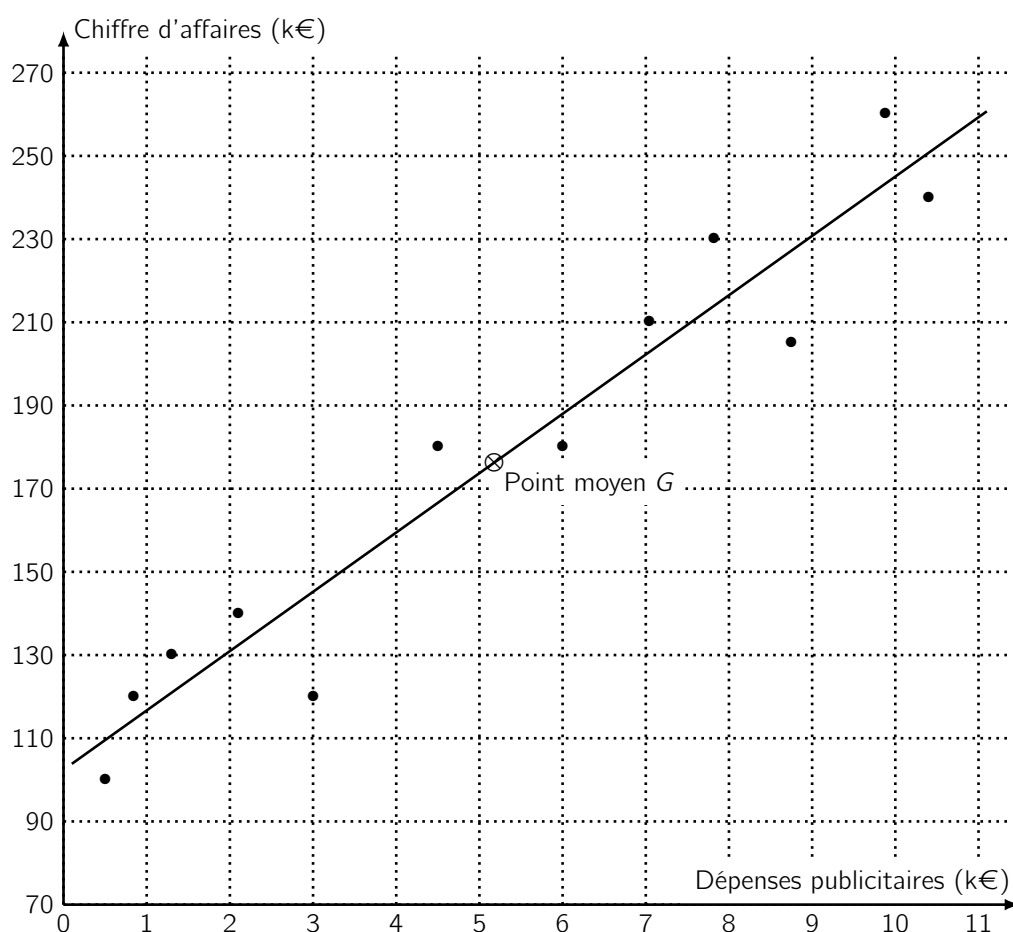
(1). Franz Messerli, « Chocolate Consumption, Cognitive Function, and Nobel Laureates », The New England Journal of Medicine, 2012.

6.4 Ajustement affine graphique

Nous remarquons qu'il semble exister un lien entre les dépenses publicitaires et le chiffre d'affaires. Ce lien pourrait (et c'est ce qui semble le plus logique) être une droite

Nous allons essayer de placer manuellement sur le graphique une droite qui passerait entre les points et qui serait la plus « correcte » possible. Comme le but est de pouvoir, par la suite, interpoler ou extrapoler les données, il faut connaître l'équation de la droite.

Cet ajustement affine graphique (c'est-à-dire effectué manuellement) sera réalisé en plaçant sur le graphe le point moyen $G^{(2)}$. La droite doit passer par ce point⁽³⁾ et on essaie d'être le plus près possible de chacun des autres points, ce qui donne :



(2). Pour déterminer le point moyen G , on calcule la moyenne des coordonnées x et la moyenne des coordonnées y . Autrement dit : $G(\bar{x}; \bar{y})$. Il a été attribué la lettre G à ce point car il s'agit du centre de gravité du nuage de points.

(3). Le point moyen G a pour coordonnées $(5,1775; 176,25)$ et est représenté par une croix encadrée.

6.5 Droite de Mayer

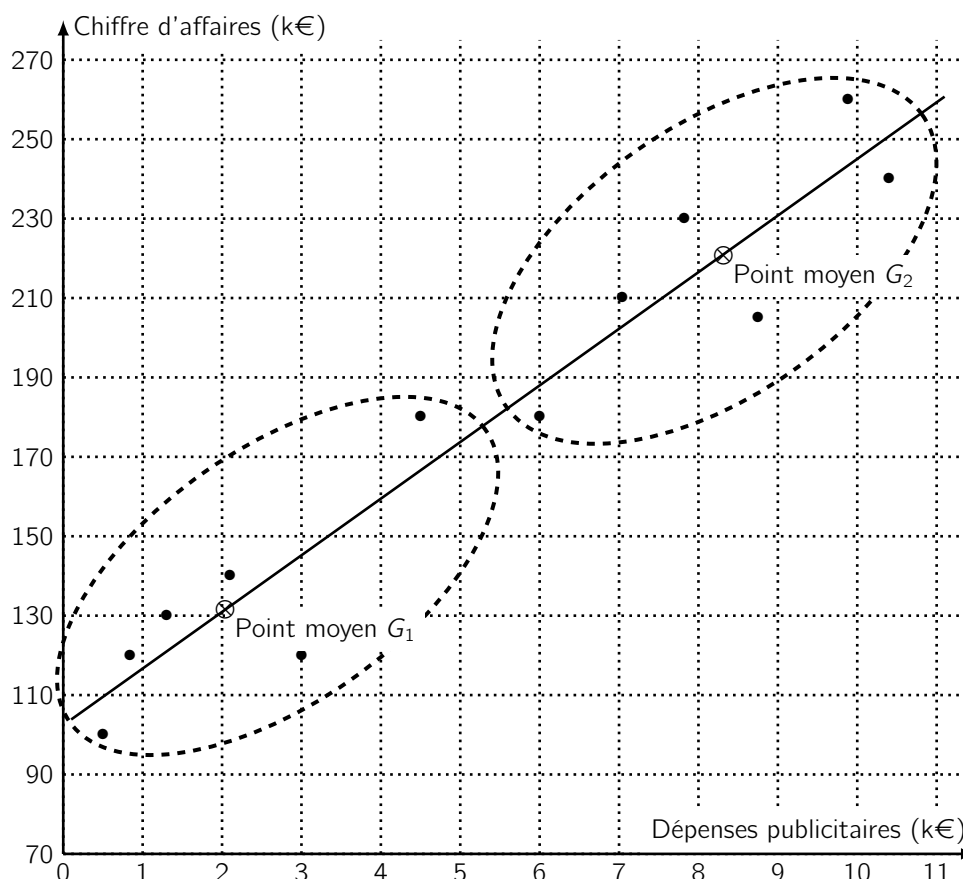
La méthode de Mayer⁽⁴⁾ consiste à découper le nuage de points en deux sous-ensembles de même effectif. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et celle des y_i . On obtient ainsi deux points moyens $G_1(\bar{x}_1; \bar{y}_1)$ et $G_2(\bar{x}_2; \bar{y}_2)$. Il reste à déterminer l'équation de la droite passant par ces deux points et à la représenter.

Je rappelle la formule à utiliser pour déterminer aisément l'équation de la droite passant par les points $A(x_A; y_A)$ et $B(x_B; y_B)$:

$$d \equiv y - y_B = \frac{y_B - y_A}{x_B - x_A} \cdot (x - x_B)$$

Ce qui revient, dans le cas qui nous intéresse, à déterminer l'équation de la droite passant par les points $G_1(\bar{x}_1; \bar{y}_1)$ et $G_2(\bar{x}_2; \bar{y}_2)$:

$$d \equiv y - \bar{y}_2 = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} \cdot (x - \bar{x}_2)$$



(4). Tobias Mayer (1723-1762) est un mathématicien, cartographe et astronome allemand.

6.6 Exercices

Exercice 1

Pour chaque couple de points, déterminer l'équation de la droite passant par ces points :

- | | |
|-------------------------|--------------------------|
| a) $(0; 2) \& (-4; 9)$ | g) $(-1; 3) \& (2; 4)$ |
| b) $(3; -2) \& (-6; 4)$ | h) $(2; -1) \& (4; -3)$ |
| c) $(1; 7) \& (0; 4)$ | i) $(-4; -1) \& (-2; 5)$ |
| d) $(1; 3) \& (-2; 4)$ | j) $(3; 1) \& (1; 3)$ |
| e) $(2; 5) \& (-3; -7)$ | k) $(-2; 4) \& (1; 3)$ |
| f) $(-3; 1) \& (4; -2)$ | l) $(7; 7) \& (0; 0)$ |

Exercice 2

Pour chacune des séries statistiques suivantes, déterminer la droite de Mayer.

- Série 1 :

x_i	1	2	3	4	5	6	7	8
y_i	24,4	26,3	27,8	29,5	30,7	32,8	34,4	35,7

- Série 2 :

x_i	3	7	1	20	12	11	14	6	10	9
y_i	105	104	118	79	90	92	91	109	103	100

- Série 3 :

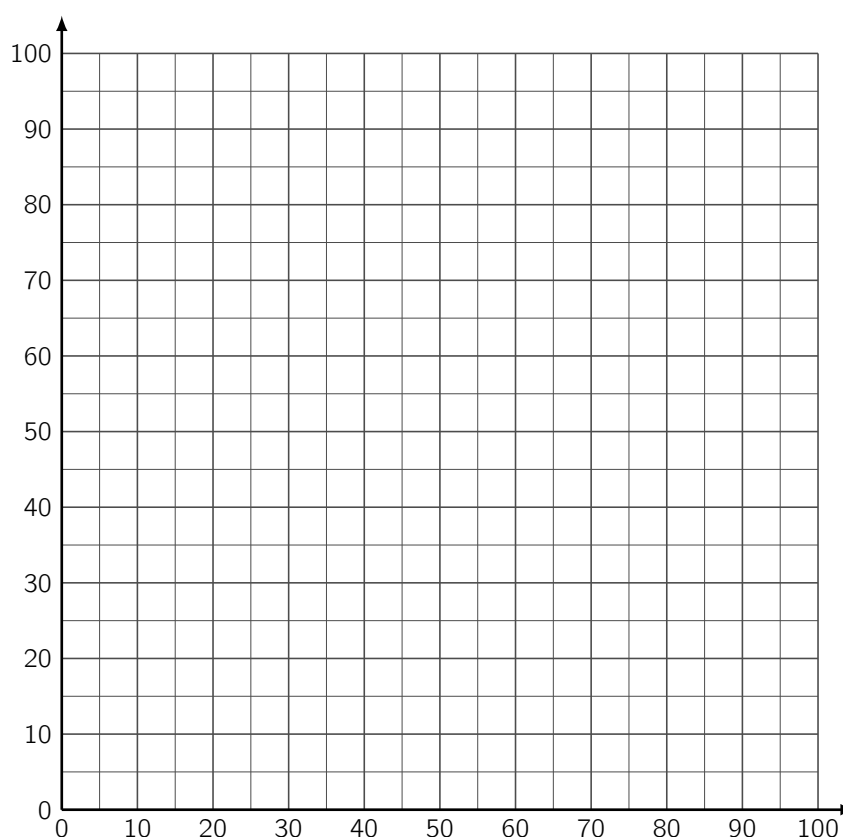
x_i	122	177	77	135	109	88	185	128	120	146
y_i	185	221	114	164	125	118	193	160	151	172

Exercice 3

La résistance aux antibiotiques a un impact mondial de plus en plus important sur notre santé. Des chercheurs testent en laboratoire la résistance d'une bactérie à un nouvel antibiotique. Les résultats sont consignés dans le tableau suivant :

Quantité de médicament administrée (en cg)	12	24	32	34	47	52	56	59	72	90
Pourcentage de bactéries survivantes en fin de test	89	84	85	87	66	63	43	44	38	26

- Représenter le nuage de points :



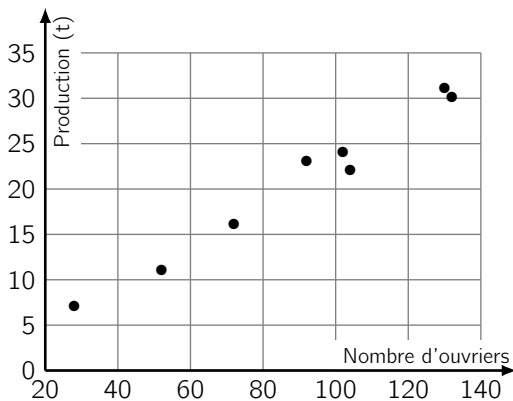
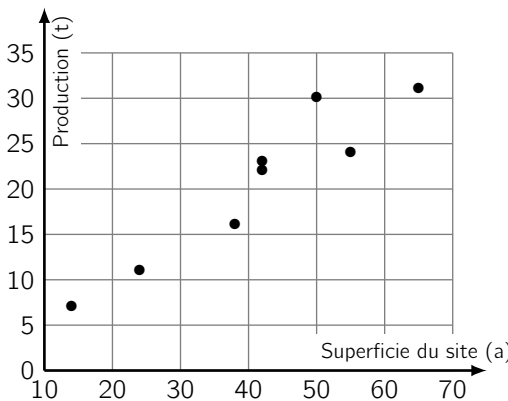
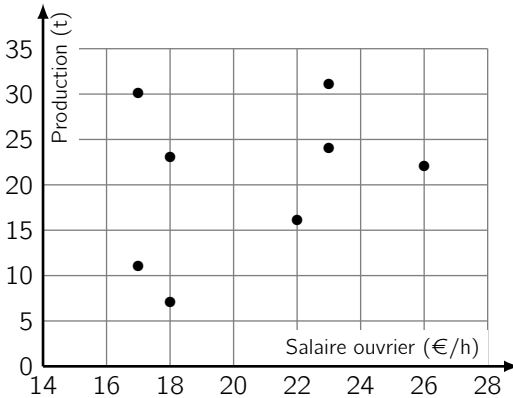
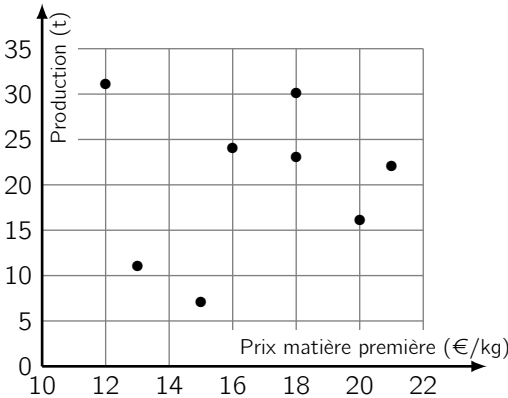
- Déterminer l'équation de la droite de Mayer puis la représenter sur le graphe ci-dessus.
- À partir de l'équation de la droite de Mayer, estimer :
 - le pourcentage de bactéries restantes avec 40 cg d'antibiotique.
 - la quantité d'antibiotique nécessaire pour tuer 75 % des bactéries.
- Selon l'expérience, quelle serait la quantité d'antibiotique à administrer pour qu'il ne reste plus aucune bactérie ?

Exercice 4

Un chef d'entreprise a réalisé une étude concernant la productivité de ses sites européens. Le tableau suivant donne les informations concernant les divers sites ainsi que quelques variables susceptibles d'influencer la production.

Sites	Liège	Paris	Bilbao	Rome	Venise	Amsterdam	Cologne	Londres
Production hebdomadaire (t)	11	31	30	22	7	23	16	24
Prix matière première (€/kg)	13	12	18	21	15	18	20	16
Salaire moyen des ouvriers (€/h)	17	23	17	26	16	18	22	23
Superficie du site (a)	24	65	50	42	14	42	38	55
Nombre d'ouvriers	52	130	32	104	28	92	72	102

- Sur base des graphiques suivants, déterminer les variables qui influencent la production des sites.



Exercice 4 (suite)

- Le chef d'entreprise a acquis un nouveau site de production de 70 ares et désire y avoir une production de 35 tonnes. Est-ce faisable selon les données statistiques ? Justifier.
- À l'aide de l'équation de la droite de Mayer, estimer le nombre d'ouvriers que le chef d'entreprise doit employer sur le nouveau site.
- Serait-il judicieux d'utiliser les données statistiques pour estimer la production d'un site de plus de 200 ares ? Justifier.

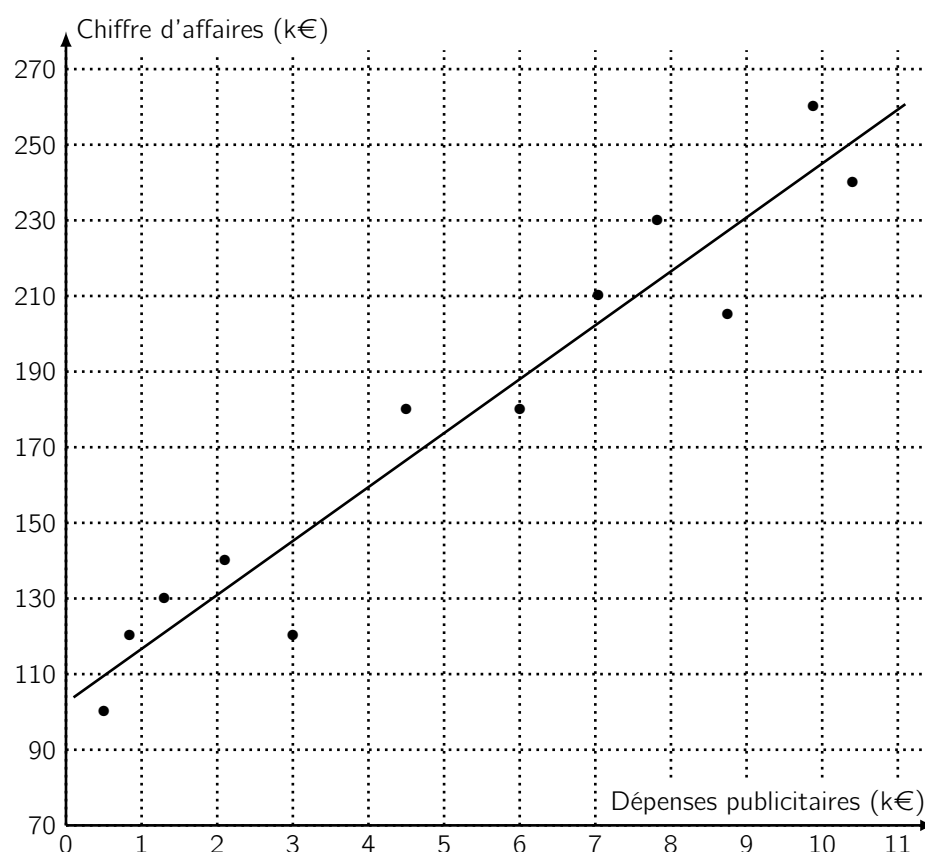
6.7 Régression linéaire – méthode des moindres carrés

Lorsque l'on a un grand nombre de données, il devient très vite fastidieux de réaliser ce travail sans l'aide d'une application informatique. Nous demanderons donc à un logiciel (généralement un tableur) de réaliser ce travail à notre place.

Dans le cas de l'activité ci-dessus, le logiciel nous fournit comme équation de la droite de régression :

$$y = 14,261x + 102,42$$

Ce qui donne sur le graphique :



On note que certaines des observations (les points) sont au-dessus de la droite de régression et d'autres sont en dessous. Sauf cas particulier, il y aura toujours un écart positif ou négatif entre la valeur réelle de Y qui correspond à X et la valeur de Y donnée par la droite de régression.

Il peut parfois arriver que certains points tombent exactement sur la droite de régression, mais ce n'est pas le cas dans notre exemple.

Dans le cas tout-à-fait exceptionnel où tous les points sont alignés, tous les écarts sont nuls ; dans ce cas particulier, la droite de régression épouserait parfaitement l'ensemble des points. Le coefficient de corrélation vaudrait alors 1 (un) mais on en reparlera plus loin au point 6.8 de ce chapitre (p.70).

L'objectif de la méthode des moindres carrés, aussi appelée *régression linéaire* est de minimiser la somme des écarts (ou « erreurs »). Le problème est que certains écarts sont positifs tandis que d'autres écarts sont négatifs. Il faut donc élever tous les écarts au carré. Ensuite on fait la somme de ces écarts au carré. Puis on minimise cette somme.

Le résultat de cette procédure mathématique complexe est le duo de formules ci-dessous qui permet de calculer les coefficients a et b de la droite de régression d'équation $y = ax + b$:

$$a = \frac{\sum_{i=1}^n (x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i^2) - n\bar{x}^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

Par souci de facilité et d'efficacité, nous disposerons les calculs dans un tableau (voir page suivante).

x_i	y_i	$x_i y_i$	x_i^2
7,04	210	1478,4	49,5616
10,4	240	2496	108,16
8,75	205	1793,75	76,5625
2,1	140	294	4,41
4,5	180	810	20,25
6	180	1080	36
3	120	360	9
0,5	100	50	0,25
0,84	120	100,8	0,7056
1,3	130	169	1,69
7,82	230	1798,6	61,1524
9,88	260	2568,8	97,6144
Somme :	Somme :	Somme :	Somme :
62,13	2115	12999,35	465,3565
Moyenne ($\bar{x}$) :	Moyenne ($\bar{y}$) :	$n\bar{x}\bar{y}$:	$n\bar{x}^2$:
5,1775	176,25	10950,4125	321,678075

Maintenant que notre tableau est complété, il est bien plus facile de calculer les coefficients a et b . On obtient alors l'équation suivante :

$$y = 14,261x + 102,42$$

ou encore, ce qui nous intéresse plus,

$$\text{chiffre d'affaires} = 14,26 \times \text{dépenses publicitaires} + 102,42$$

6.8 Coefficient de corrélation

La qualité de la corrélation ci-dessus, c'est-à-dire la fiabilité de la droite déterminée par rapport aux valeurs dont on dispose, peut être mesurée par un coefficient que l'on appelle *coefficient de corrélation*, généralement noté R^2 . Celui-ci a toute son importance. En effet, il ne s'agit pas simplement de déterminer l'équation d'une droite, il faut vérifier que cette droite se rapproche bien des points expérimentaux, c'est-à-dire qu'il y a bien un lien de corrélation entre mes dépenses mensuelles de publicité et mon chiffre d'affaires mensuel.

Lorsque l'on a estimé la droite de régression, on doit se demander, avant de pouvoir accepter cette droite comme modèle, si cette estimation est de bonne qualité. Pour y répondre, on détermine le coefficient de corrélation avec la formule suivante :

$$R^2 = \frac{\left[n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) \right]^2}{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}$$

Ce coefficient est compris entre 0 (zéro) et 1 (un). Plus il est proche de 1, plus la qualité globale de la régression est bonne.

Après avoir calculé le coefficient de corrélation de l'équation de régression déterminée plus haut (et ce, afin de confirmer ou d'infirmer le lien de corrélation entre mes dépenses mensuelles de publicité et mon chiffre d'affaires mensuel), on obtient la valeur de :

$$R^2 = 0,9259$$

Le coefficient de corrélation étant proche de 1, nous pouvons affirmer que la droite de régression déterminée plus haut est de bonne qualité et qu'il y a donc bien un lien de corrélation entre mes dépenses mensuelles de publicité et mon chiffre d'affaires mensuel.