Statistiques descriptives

François Mansy

Table des matières

Première partie

Statistiques descriptives

Chapitre 1

Introduction et vocabulaire

Bien que le nom de statistique soit relativement récent [1], cette activité semble exister dès la naissance des premières structures sociales. D'ailleurs, les premiers textes écrits retrouvés étaient des recensements du bétail, des informations sur son cours, et des contrats divers. On a ainsi trace de recensements en Chine au XXIII^{e} siècle av. J.-C. ou en Égypte au XVIII^{e} siècle av. J.-C.. Ce système de recueil de données se poursuit jusqu'au XVII^{e} siècle. En Europe, le rôle de collecteur est souvent tenu par des guildes marchandes puis par les intendants de l'État.

Ce n'est qu'au XVIII^{e} siècle que l'on vit apparaître le rôle prévisionnel des statistiques avec la construction des premières tables de mortalité.

La statistique mathématique s'appuya sur les premiers travaux concernant les probabilités développés par Fermat et Pascal. C'est probablement chez Thomas Bayes que l'on vit apparaître un embryon de statistique inférentielle. Condorcet et Laplace parlaient encore de probabilité là où l'on parlerait aujourd'hui de fréquence. Mais c'est à Adolphe Quételet que l'on doit l'idée que la statistique est une science s'appuyant sur les probabilités.

Le XIX^{e} siècle vit cette activité prendre son plein essor. Des règles précises sur la collecte et l'interprétation des données sont édictées. La première application industrielle des statistiques eut lieu avec le recensement américain de 1890, qui mit en oeuvre la carte perforée inventée par le statisticien Herman Hollerith. Celui-ci avait déposé un brevet au bureau américain des brevets.

Au XX^{e} siècle, ces applications industrielles se développèrent d'abord aux États-Unis, qui étaient en avance sur les sciences de gestion, puis seulement après la Première Guerre mondiale en Europe. Le régime nazi employa des méthodes statistiques à partir de 1934 pour le réarmement. En France, on était moins au fait de ces applications.

L'application industrielle des statistiques en France se développa avec la création de l'INSEE, qui remplaça le Service National des Statistiques créé par René Carmille.

1. On attribue en général l'origine du nom au XVIII^{e} siècle de l'allemand Staatskunde

L'avènement de l'informatique dans les années 1940 (aux États-Unis) puis en Europe (dans les années 1960) permit de traiter un plus grand nombre de données, mais surtout de croiser entre elles des séries de données de types différents. C'est le développement de ce qu'on appelle l'analyse multidimensionnelle.

La **statistique** est donc l'ensemble des instruments et de recherches mathématiques permettant de déterminer les caractéristiques d'un ensemble de données (généralement vaste). Les statistiques sont le produit des analyses reposant sur l'usage de la statistique. Cette activité regroupe trois principales branches :

- la collecte des données ;
- le traitement des données collectées, aussi appelé la statistique descriptive ;
- l'interprétation des données, aussi appelée l'inférence statistique, qui s'appuie sur la théorie des sondages et la statistique mathématique.

Une **enquête statistique** consiste à observer une certaine **population** (élèves d'une classe, personnes âgées de 20 à 60 ans dans une région donnée, familles dans une région donnée, exploitations agricoles, appartements, travailleurs, ...) et à déterminer la répartition d'un certain **caractère** statistique (note obtenue, taille, nombre d'enfants, superficie, nombre de pièces, secteur d'activité, ...) dans cette **population**.

Lorsque le **caractère** statistique prend un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, secteur d'activité, ...), le caractère statistique est **discret**.

Lorsque le **caractère** statistique peut prendre des valeurs multiples (taille, superficie, salaire, ...) le caractère statistique est considéré comme **continu**.

Lorsque le **caractère** statistique est un nombre (taille, note, nombre d'enfant, ...) on parle de caractère **quantitatif**, quand ce **caractère** n'est pas chiffré (langue parlée, secteur d'activité, couleur, ...) on parle de caractère **qualitatif**.

Pour résumer, la **statistique** est une méthode scientifique ayant pour objet la collecte, l'analyse et l'interprétation d'un ensemble d'observations chiffrées relatives à un même phénomène.

La **population** est l'ensemble soumis à une étude statistique.

Le **caractère** est le trait déterminé, commun à tous les éléments de la population sur laquelle porte l'étude.

Un **caractère** est dit **quantitatif** si son intensité varie et que l'on peut mesurer.

Un **caractère quantitatif** peut être
- **discret** : lorsqu'il ne prend que certaines valeurs d'un intervalle.
- **continu** : lorsqu'il peut prendre n'importe quelle valeur réelle dans un intervalle de variation.

Un **caractère** est dit **qualitatif** s'il est non mesurable et que sa nature peut varier.

Une **série** ou **distribution statistique** est l'ensemble des valeurs du caractère considéré pour une population.

Un **échantillon** est une partie de la population qu'on étudie, s'il n'est pas possible d'en étudier chaque élément.

La statistique développe un certain nombre d'outils pour traiter les résultats d'une enquête.

En mathématiques élémentaires, les statistiques sont principalement descriptives. Or la tentation est grande de partir des informations obtenues dans un échantillon pour en tirer une généralité sur la population tout entière. Passer du particulier au général est normalement une démarche interdite, si elle est faite sans précaution. En ce qui concerne les statistiques, cette démarche est un objectif.

On aborde alors la deuxième partie des statistiques : la statistique inférentielle ou théorie des sondages. Cette science permet de déterminer quelles sont les précautions à prendre pour passer du particulier au général (taille et représentativité de l'échantillon, ...), quels sont les risques d'erreur que l'on peut commettre. Elle est alors très liée et presque confondue avec la science des probabilités. Aussi, nous ne l'aborderons pas avant d'en avoir établi les bases.

Dans cette première partie du cours, nous nous limiterons à l'étude de la statistique descriptive. L'objectif de la statistique descriptive est de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses.

Chapitre 2

Séries statistiques simples (à une dimension)

2.1 Tabulation

Les résultats d'une enquête consistent en une liste désordonnée d'informations.

Ex1 - note de la classe X : 10, 9, 12, 11, 10, 8, 14 ,11 ,9 ,16 ,5 ,12 ,10 ,11 ,10 ,13

Ex2 - couleur préférée : bleu, rouge, bleu, bleu, jaune, bleu, rouge, bleu, bleu, jaune, jaune, bleu, jaune.

Il faut alors les trier, par ordre croissant, pour le caractère quantitatif, par genre, pour le caractère qualitatif.

Notes triées : 5, 8, 9, 9, 10, 10, 10, 10, 11, 11, 11, 12, 12, 13, 14, 16

Couleurs préférées triées : bleu, bleu, bleu, bleu, bleu, bleu, bleu, rouge, rouge, jaune, jaune, jaune, jaune.

Cette présentation sous forme de liste est peu exploitable, on décide alors de présenter les résultats de l'enquête sous forme d'un tableau d'effectifs. L'**effectif** ou **fréquence absolue** d'une valeur x_i est le nombre n_i de fois où cette valeur x_i apparaît.

Exemple 1 : note x_i des élèves

notes	x_i	$x_1 = 5$	$x_2 = 8$	$x_3 = 9$	$x_4 = 10$	$x_5 = 11$	$x_6 = 12$	$x_7 = 13$	$x_8 = 14$	$x_9 = 16$	Total
effectifs	n_i	$n_1 = 1$	$n_2 = 1$	$n_3 = 2$	$n_4 = 4$	$n_5 = 3$	$n_6 = 2$	$n_7 = 1$	$n_8 = 1$	$n_9 = 1$	$n_{tot} = 16$

Dans ce tableau, par exemple, on voit directement que 3 élèves sur les 16 interrogés ont obtenu un 11.

Exemple 2 : couleur x_i préférée

Couleurs x_i	Effectifs n_i
$x_1 =$ Bleu	$n_1 = 7$
$x_2 =$ Rouge	$n_2 = 2$
$x_3 =$ Jaune	$n_3 = 4$
Total	$n_{tot} = 13$

Dans ce tableau, par exemple, on voit directement que 7 personnes sur les 13 questionnées préfèrent le bleu.

Lorsque la population étudiée est trop grande, ou bien lorsque l'on cherche à faire la comparaison entre deux populations de tailles différentes, on préfère se ramener à une population de 100, donc travailler en pourcentages, appelés ici fréquences f_i.

Exemple 1 : note x_i des élèves

notes	x_i	5	8	9	10	11	12	13	14	16	Total
fréquences (en %)	f_i	6,25	6,25	12,50	25,00	18,75	12,50	6,25	6,25	6,25	100
fréq. cumulées (en %)	F_i	6,25	12,50	25,00	50,00	68,75	81,25	87,50	93,75	100	

Dans ce tableau, grâce à la fréquence cumulée, par exemple, on voit directement que 25 % des 16 étudiants interrogés ont obtenu une note strictement inférieure à 10.

Exemple 2 : couleur x_i préférée

Couleurs x_i	Fréquences (en %) f_i	Fréq. cumulées (en %) F_i
Bleu	53,85	53,85
Rouge	15,38	69,23
Jaune	30,77	100
Total	100	

Dans ce tableau, par exemple, on voit directement qu'un peu plus de 50 % de ces 13 personnes préfèrent le bleu et qu'un peu moins de 70 % préfèrent le bleu ou le rouge.

Pour résumer, les p différentes valeurs ou natures des caractères constituent les p données x_1, x_2, ... x_p.

Le nombre de fois qu'une donnée x_i se présente dans la distribution s'appelle l'effectif (ou répétition, ou fréquence absolue) n_i.

L'effectif total n (nombre total de données) est donné par la somme des effectifs n_i. On a

$$\sum_{i=1}^{p} n_i = n_1 + n_2 + n_3 + ... + n_{p-1} + n_p = n_{tot} = n$$

L'effectif cumulé N_i de la donnée x_i est donné par

$$N_i = \sum_{j=1}^{i} n_j = n_1 + n_2 + ... + n_i$$

On a forcément que $N_p = \sum_{j=1}^{p} n_j = n$.

La fréquence relative f_i de la donnée x_i est donnée par

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

Par conséquent,

$$\begin{array}{rcc} \sum_{i=1}^{p} f_i & = & f_1 + f_2 + ... + f_p \\ & = & \frac{n_1}{n} + \frac{n_2}{n} + ... + \frac{n_p}{n} \\ & = & \frac{1}{n}(n_1 + n_2 + ... + n_p) \\ & = & \frac{1}{n}\sum_{i=1}^{p} n_i \\ & = & 1 \end{array}$$

La fréquence relative cumulée F_i de la donnée x_i est donnée par

$$F_i = \sum_{j=1}^{i} f_j = f_1 + f_2 + ... + f_i$$

On a forcément que $F_p = \sum_{j=1}^{p} f_j = 1$.

2.2 Classes

Lorsque les résultats de l'enquête statistique sont trop nombreux pour que la liste triée des valeurs soit lisible (supérieur à 20 en général), on préfère perdre de l'information et ranger les données par intervalles appelés classes. Il faut alors que, dans chaque classe, la répartition des valeurs soit régulière. Sinon, il faut affiner et prendre des classes plus petites. Il n'est pas indispensable que les classes soient de même amplitude, mais il est préférable de ne pas définir de classes de la forme « plus de ... » qui empêcherait alors tout traitement ultérieur (histogramme, moyenne, ...). On compte alors le nombre de fois où la valeur du caractère tombe dans l'intervalle $[x_i, x_{i+1}[$, ce nombre est appelé effectif de la classe $[x_i, x_{i+1}[$.

Exemple : Répartition des revenus annuels en milliers d'euros dans une population de 4370 personnes.

Salaires	$[0,8[$	$[8,12[$	$[12,16[$	$[16,20[$	$[20,30[$	$[30,40[$	$[40,60[$	Total
Effectifs	306	231	385	1180	1468	568	232	4370
Fréquences	7,0	5,3	8,8	27,0	33,6	13,0	5,3	100

Puisque l'on a estimé que la répartition dans chaque classe était régulière, on peut affirmer que le milieu de la classe est représentatif de la classe. On va donc remplacer les n_i individus de la classe $[x_i, x_{i+1}[$ par n_i individus dont le caractère statistique prendrait la valeur $m_i = \frac{x_i + x_{i+1}}{2}$.

2.3 Représentation graphique des séries quantitatives

2.3.1 Diagramme en bâtons

On porte en abscisse les bornes des classes et en ordonnées les fréquences ou les effectifs. À partir de chaque point obtenu, on abaisse un segment perpendiculairement à l'axe des abscisses. Cette représenta-

tion met par exemple en évidence des maxima ou minima éventuels.

2.3.2 Histogramme

On porte en abscisse les bornes des classes et en ordonnée les fréquences. On représente, pour chaque classe, un rectangle dont la base est proportionnelle à l'intervalle de la classe et la hauteur est proportionnelle à la fréquence ou à l'effectif correspondant et inversément proportionnelle à la largeur de l'intervalle de cette classe. L'histogramme est la ligne brisée bordant l'ensemble des rectangles. Cette représentation met en évidence le fait qu'une classe contient plusieurs éléments dont la mesure de la caractéristique a été ramenée à la valeur centrale.

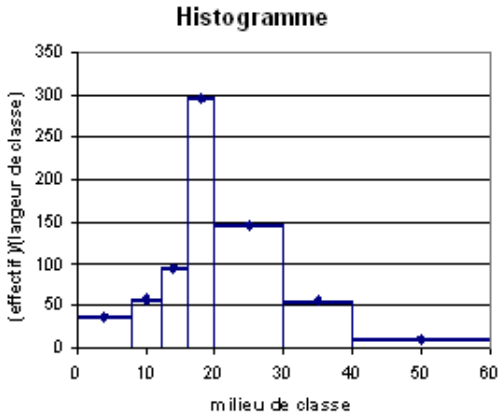

FIGURE 2.1 – Répartition des revenus annuels en milliers d'euros dans une population de 4370 personnes.

Plus généralement, pour un même effectif, si l'amplitude de la classe est deux fois plus grande, la hauteur du rectangle doit être deux fois plus petite.

2.3.3 Polygone des fréquences

On reprend les coordonnées des points du diagramme en bâtons (centre de classe, fréquence). On y ajoute les deux points correspondants à la première et dernière classe de fréquence nulle et on joint ces points par ordre d'abscisses croissants. Ce graphique met en évidence la variation de la fréquence ou de l'effectif d'une classe à la suivante.

2.3.4 Diagramme cumulatif

On porte en abscisse les données et en ordonnée les fréquences (ou effectifs) cumulées. On dessine la fonction

$$f(x) = F_i \quad \text{avec} \quad x_i \leq x < x_{i+1}$$

ou

$$g(x) = N_i \quad \text{avec} \quad n_i \leq x < n_{i+1}$$

Ce graphique permet de repérer facilement la médiane (voir plus loin).

2.3.5 Polygone cumulatif

On porte sur un graphe les points ayant pour abscisse les bornes supérieures des classes et en ordonnées les fréquences cumulées correspondantes (y compris la première classe de fréquence nulle). On joint ces points par ordre d'abscisses croissants.

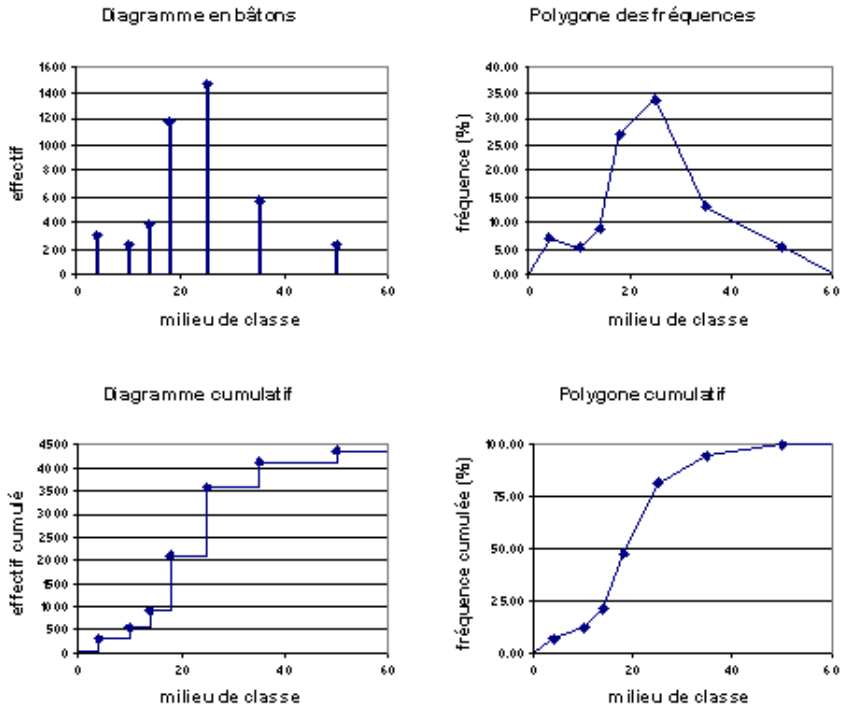

FIGURE 2.2 – Répartition des revenus annuels en milliers d'euros dans une population de 4370 personnes.

2.4 Paramètres de position

En statistiques, on est en général en présence d'un grand nombre de valeurs. Or, si l'intégralité de ces valeurs forme l'information, il n'est pas aisé de manipuler plusieurs centaines voir milliers de chiffres, ni d'en tirer des conclusions. Il faut donc calculer quelques valeurs qui vont permettre d'analyser les données.

En mesure physique (métrologie), on va en général calculer deux valeurs : la moyenne, qui représentera la « valeur » de la mesure, et l'écart type (paramètre de dispersion), qui va estimer l'erreur de mesure.

Dans d'autres domaines, on va vouloir avoir une description plus fine de la répartition des valeurs, et donc calculer d'autres paramètres de position.

2.4.1 Extrema

La valeur maximale est la plus grande valeur prise par le caractère statistique.
La valeur minimale est la plus petite valeur prise par le caractère statistique.

2.4.2 Médiane

La médiane est la valeur du caractère statistique qui coupe la population en deux populations de taille égale. Sur le polygone cumulatif, il s'agit de l'abscisse du point dont l'ordonnée, qui est la fréquence cumulée, vaut 50 %.

Cas discret

On trie les valeurs par ordre croissant.

- Si la population comporte n individus et si n est impair alors $n = 2p+1$, la médiane sera la $(p+1)^e$ valeur du caractère statistique.

  Exemple : série de 13 notes : 4, 5, 7, 8, 8, 9, 10, 10, 10, 11,12, 13, 16. Médiane $= M = 10$

- Si la population comporte n individus et si n est pair alors $n = 2p$, la médiane sera la moyenne entre la p^e et $(p+1)^e$ valeur du caractère statistique.

  Exemple : série de 12 notes : 4, 5, 7, 8, 8, 9, 10, 10, 10, 11, 13, 16. Médiane $= M = 9,5$

Cas continu

On utilise le polygone des fréquences cumulées croissantes et le tableau correspondant et on détermine graphiquement ou par interpolation linéaire la valeur M pour laquelle la fréquence de l'intervalle [valeur min, M] vaut 50 %.

Dans l'exemple précédent, le tableau des fréquences cumulées croissantes est :

x_i	0	8	12	16	20	30	40	60
fréq. cumulée croissante (en %)	0	7	12,3	21,1	48,1	81,7	94,7	100

Les 50 % sont atteint entre 20 et 30 donc pour une valeur M que l'on estime à $20+10\frac{50-48,1}{81,7-48,1} = 20,56$ par interpolation linéaire.

2.4.3 Mode

Le mode est la valeur du caractère statistique qui apparaît le plus fréquemment.

Exemple 1 : note des élèves

note	x_i	5	8	9	10	11	12	13	14	16	Total
effectif	n_i	1	1	2	4	3	2	1	1	1	16

Le mode est 10. Exemple 2 : note des élèves

note	x_i	5	8	9	10	11	12	13	14	16	Total
effectif	n_i	1	1	4	2	2	4	1	1	1	16

Cette série est dite série bimodale car on voit apparaître deux modes : 9 et 12.

Dans le cas d'une variable continue, on peut entendre parler de classe modale qui serait la classe de plus grand effectif. Mais il faut se méfier de cette notion car, plus la classe est de grande amplitude, plus son effectif est important sans pour autant que cela soit significatif. Cette notion de classe modale définie par les effectifs de la classe n'a de sens que si les classes ont même amplitude. Si les amplitudes sont différentes, il faut aller chercher sur l'histogramme la classe associée au rectangle de plus grande hauteur.

Exemple : Répartition des revenus annuels en milliers d'Euros dans une population de 4370 personnes.

Salaire	$[0,8[$	$[8,12[$	$[12,16[$	$[16,20[$	$[20,30[$	$[30,40[$	$[40,60[$	Total
Effectif	306	231	385	1180	1468	568	232	4370
Fréquence (en %)	7,0	5,3	8,8	27,0	33,6	13,0	5,3	100
Fréquence/Intervalle	0,875	1,325	2,200	6,750	3,360	1,300	0,265	

L'observation de ce tableau laisse penser que la classe modale serait la classe $[20,30[$. Mais une observation de l'histogramme, et donc de la fréquence divisée par la largeur de l'intervalle, corrige cette idée fausse : La classe modale est la classe $[16,20[$.

2.4.4 Quartiles

Les quartiles sont les trois valeurs qui partagent la population en 4 sous-populations de même taille (25 %). Ces valeurs correspondent donc aux fréquences cumulées de 25 %, 50 % et 75 %.

Cas discret

On range les valeurs par ordre croissant.

On détermine le second quartile qui correspond à la médiane. Puis on cherche la médiane de la première moitié de la population qui correspond au 1er quartile. On cherche la médiane de la seconde moitié de la population qui correspond au troisième quartile.

Si la population est de taille n, on distingue 4 cas :

- Si $n = 4p$:
  - Q1 = moyenne entre la p^e et la $(p+1)^e$ valeur.
  - Q2 = moyenne entre la $(2p)^e$ valeur et la $(2p+1)^e$ valeur.
  - Q3 = moyenne entre la $(3p)^e$ valeur et la $(3p+1)^e$ valeur.

  Exemple : série de 12 notes : 4, 5, 7, 8, 8, 9, 10, 10, 10, 11, 13, 16
  Q1 = 7,5 ; Q2 = 9,5 ; Q3 = 10,5

- Si $n = 4p + 1$:
  - Q1 = moyenne entre la p^e et $(p+1)^e$ valeur.
  - Q2 = $(2p+1)^e$ valeur.
  - Q3 = moyenne entre la $(3p+1)^e$ valeur et la $(3p+2)^e$ valeur.

  Exemple : série de 13 notes 4, 5, 7, 8, 8, 9, 10, 10, 10, 11,12, 13, 16
  Q1 = 7,5 ; Q2 = 10 ; Q3 = 11,5

- Si $n = 4p + 2$:
  - Q1 = $(p+1)^e$ valeur.
  - Q2 = moyenne entre la $(2p+1)^e$ valeur et la $(2p+2)^e$ valeur.
  - Q3 = $(3p+2)^e$ valeur.

  Exemple : série de 14 notes 4, 5, 7, 8, 8, 9, 9, 10, 10, 10, 11, 12,13, 16
  Q1 = 8 ; Q2 = 9,5 ; Q3 = 11

- Si $n = 4p + 3$:
  - Q1 = $(p+1)^e$ valeur.
  - Q2 = $(2p+2)^e$ valeur.
  - Q3 = $(3p+3)^e$ valeur.

  Exemple : série de 15 notes 4, 5, 7, 8, 8, 9, 9, 10, 10, 10, 11,11, 12, 13, 16
  Q1 = 8 ; Q2 = 10 ; Q3 = 11

En pratique, on range les valeurs par ordre croissant.

$Q1$ est la première valeur pour laquelle l'intervalle $[x_{min}, Q1]$ regroupe au moins 25 % de la population.
$Q2$ est la première valeur pour laquelle l'intervalle $[x_{min}, Q2]$ regroupe au moins 50 % de la population.
$Q3$ est la première valeur pour laquelle l'intervalle $[x_{min}, Q3]$ regroupe au moins 75 % de la population.

En reprenant les exemples précédents :

Si $n = 12$: 25 % de $n = 3$, puis 50 % de $n = 6$, puis 75 % de $n = 9$.
La série de notes est 4, 5, **7**, 8, 8, **9**, 10, 10, **10**, 11, 13, 16
$Q1 = 7$, $Q2 = 9$, $Q3 = 10$

Si $n = 13$: 25 % de $n = 3,25$, puis 50 % de $n = 6,5$, puis 75 % de $n = 9,75$ que l'on arrondit à l'entier

supérieur.

La série de notes est 4, 5, **7**, 8, 8, 9, **10**, 10, 10, **11**, 12, 13, 16
$Q1 = 8$, $Q2 = 10$, $Q3 = 12$

On s'aperçoit que cette approximation rend dissymétrique la définition, que le second quartile ne correspond plus à la médiane et que les valeurs obtenues diffèrent de celles de la définition précédente. Son avantage est de rendre la recherche des quartiles (approchés) plus facile sans que l'on soit obligé de distinguer 4 cas. Les différences obtenues par l'une ou l'autre des méthodes se révèlent négligeables et justifient l'usage de cette approximation.

Cas continu

On calcule les quartiles comme la médiane, graphiquement grâce au polygone des fréquences cumulées croissantes, et par interpolation linéaire grâce au tableau correspondant.

Le tableau des fréquences cumulées croissantes est :

x_i	0	8	12	16	20	30	40	60
fréq. cumulées croissantes (en %	0	7	12,3	21,1	48,1	81,7	94,7	100

25% est atteint dans l'intervalle $[16, 20]$ soit pour une valeur de $Q1$ obtenue par interpolation linéaire $Q1 = 16 + 4\frac{25-21,1}{48,1-21,1} = 16,57$.

$Q2 = M = 20,56$.

75% est atteint dans l'intervalle $[20, 30]$ soit pour une valeur de $Q3$ obtenue par interpolation linéaire $Q3 = 20 + 10\frac{75-48,1}{81,7-48,1} = 28,00$.

2.4.5 Déciles et percentiles

Les déciles sont les 9 valeurs qui partagent la population en 10 sous-populations de même taille. Ces valeurs correspondent donc aux fréquences cumulées de 10 %, 20 %, ... 90 %.

Par conséquent, les percentiles sont les 99 valeurs qui partagent la population en 100 sous-populations de même taille. Si la fréquence cumulée est exprimée en pourcents les percentiles sont les valeurs de la variable correspondant respectivement aux fréquences cumulées.

Notons finalement que la médiane correspond au 5^e décile et au percentile 50.

2.4.6 Moyenne

Il y a plusieurs façon de calculer une moyenne d'un ensemble de nombres. Celle qu'il convient de retenir dépend de la grandeur physique que représentent ces nombres. Lorsque, dans le langage courant,

on parle de moyenne, on évoque en fait la moyenne arithmétique.

La moyenne est la valeur unique que devraient avoir tous les individus d'une population (ou d'un échantillon) pour que leur total soit inchangé. Dans la plupart des cas, le total formé par les individus d'une population est la somme de leurs valeurs. La moyenne est alors la moyenne arithmétique. Mais si le total représenté par une population ou un échantillon n'est pas la somme de leurs valeurs, la moyenne pertinente ne sera plus la moyenne arithmétique. Si, par exemple, le total d'un ensemble d'individus est calculé par l'inverse de la somme des inverses (cas des vitesses d'un ensemble de fractions d'un trajet, par exemple), on doit calculer leur moyenne harmonique. Si, par exemple, le total d'un ensemble d'individus est le produit de leurs valeurs, il convient de calculer leur moyenne géométrique. On rencontre, en physique, de multiples moyennes : La capacité moyenne d'un ensemble de condensateurs en série est la moyenne harmonique de leurs capacités.

La moyenne ne peut donc se concevoir que pour une variable quantitative. On ne peut pas faire le total des valeurs d'une variable qualitative.

De manière générale, la moyenne n'est pas forcément une manière pertinente de représenter les données. On peut, par exemple, lui préférer la valeur médiane qui représente la valeur à laquelle 50 % des valeurs observées sont inférieures. La médiane n'est pas (sauf exception ou hasard) équivalente à la moyenne arithmétique de l'ensemble.

Moyenne arithmétique

La moyenne arithmétique est la moyenne ordinaire, c'est-à-dire la somme des valeurs numériques (de la liste) divisée par le nombre de ces valeurs numériques. Exemple : la hauteur moyenne des toits d'une rue.

Cas de la série statistique discrète triée mais non regroupée

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

Cas de la série statistique discrète regroupée

$$\mu = \bar{x} = \frac{\sum_{i=1}^{p} n_i x_i}{\sum_{i=1}^{p} n_i} = \sum_{i=1}^{p} f_i x_i$$

Cas de la série continue (les m_i sont les milieux de classes)

$$\mu = \bar{x} = \frac{\sum_{i=1}^{p} n_i m_i}{\sum_{i=1}^{p} n_i} = \sum_{i=1}^{p} f_i m_i$$

Moyenne harmonique

La moyenne harmonique est définie de la manière suivante :

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

Si un train fait un trajet aller-retour entre 2 villes à la vitesse constante v_1 pour l'aller et à la vitesse constante v_2 au retour, la vitesse moyenne du trajet total n'est pas la moyenne arithmétique des 2 vitesses, mais leur moyenne harmonique.

Moyenne géométrique

La moyenne géométrique est définie de la manière suivante :

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

On peut illustrer la moyenne géométrique avec les deux cas suivants :

- Si l'inflation d'un pays est de 5 % la première année et de 15 % la suivante, l'augmentation moyenne des prix se calcule grâce à la moyenne géométrique des coefficients multiplicateurs 1,05 et 1,15 soit une augmentation moyenne de 9,88 % et non grâce à la moyenne arithmétique 10 %.
- Le carré (c'est-à-dire le rectangle moyen à deux cotés égaux) qui a même surface (le total considéré ici) qu'un rectangle de cotés 3 et 7 a pour coté la moyenne géométrique des deux cotés du rectangle $\sqrt[2]{3.7} = 4,58$.

Moyenne quadratique

La moyenne quadratique est définie de la manière suivante :

$$\bar{x}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Exemple : Si un rectangle a pour côtés 3 et 7, le carré (c'est-à-dire le rectangle moyen) qui a même diagonale (le total considéré ici) que ce rectangle, a pour côté la moyenne quadratique de 3 et 7, c'est-à-dire 5,385.

Moyenne pondérée

La moyenne pondérée est utilisée, par exemple, en géométrie pour localiser le barycentre d'un polygone, en physique pour déterminer le centre de gravité ou en statistique et probabilité pour calculer une espérance. On la calcule ainsi :

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Dans le cas général le poids w_i représente l'influence de l'élément x_i par rapport aux autres.

A noter qu'il s'agit ici de la moyenne pondérée arithmétique.

Valeur moyenne d'une fonction

Pour toute fonction continue (ou même seulement continue par morceaux) sur un segment $[a, b]$ non vide et non trivial ($b > a$), la valeur moyenne de f sur $[a, b]$ est le réel $\bar{f}_{[a,b]}$ défini par :

$$\bar{f}_{[a,b]} = \frac{1}{b-a} \int_a^b f(x)\, dx$$

Cette notion généralise celle de moyenne d'un nombre fini de réels en l'appliquant à un nombre infini de valeurs prises par une fonction intégrable. Elle sert par exemple dans la décomposition en série de Fourier d'une fonction périodique : c'est la composante constante. En traitement du signal, pour les signaux périodiques, il s'agit de la composante continue.

Notons que lorsque la fonction est périodique de période T, elle a la même valeur moyenne sur toute période $[a, a + T]$. Cette valeur commune est appelée valeur moyenne de la fonction. Ainsi la fonction cosinus est de moyenne nulle, son carré de moyenne $1/2$.

2.4.7 Propriétés de la moyenne

1. Soit $x'_i = x_i - \bar{x}$, l'écart entre la donnée x_i et la moyenne $\bar{x}$. On a

$$\sum_{i=1}^{p} x'_i = 0$$

En effet,

$$\begin{array}{rcc} \sum_{i=1}^{p} x'_i & = & (x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_p - \bar{x}) \\ & = & (x_1 + x_2 + ... + x_p) - n\bar{x} \\ & = & n\bar{x} - n\bar{x} \\ & = & 0 \end{array}$$

2. La somme des carrés des écarts des données x_i par rapport à la moyenne $\bar{x}$ est minimum. On a

$$\sum_{i=1}^{p} (x_i - \bar{x})^2 \text{ est minimum}$$

En effet, avec $r \in \mathbb{R} \backslash \{\bar{x}\}$,

$$\begin{array}{rcc} \sum_{i=1}^{p} (x_i - r)^2 & = & \sum_{i=1}^{p} (x_i - \bar{x} + \bar{x} - r)^2 \\ & = & \sum_{i=1}^{p} ((x_i - \bar{x}) + (\bar{x} - r))^2 \\ & = & \sum_{i=1}^{p} ((x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - r) + (\bar{x} - r)^2) \\ & = & \sum_{i=1}^{p} (x_i - \bar{x})^2 + 2(\bar{x} - r) \sum_{i=1}^{p} (x_i - \bar{x}) + (\bar{x} - r)^2 \sum_{i=1}^{p} 1 \\ & = & \sum_{i=1}^{p} (x_i - \bar{x})^2 + 0 + (\bar{x} - r)^2 p \\ & = & \sum_{i=1}^{p} (x_i - \bar{x})^2 + p(\bar{x} - r)^2 \\ & > & \sum_{i=1}^{p} (x_i - \bar{x})^2 \end{array}$$

3. La moyenne est stable par transformation affine. C'est-à-dire si $x'_i = ax_i + b$, si $\bar{x}$ est la moyenne de la série $\{x_i\}$ alors la moyenne de la série $\{x'_i\}$ est $\bar{x}' = a\bar{x} + b$.

En effet,

$$\begin{aligned}\bar{x}' &= \sum_{i=1}^{p} \frac{x'_i}{n} \\ &= \frac{1}{n}((ax_1+b)+(ax_2+b)+...+(ax_p+b)) \\ &= \frac{1}{n}(a(x_1+x_2+...+x_p)+nb) \\ &= a\bar{x}+b\end{aligned}$$

Cette propriété est utile pour changer d'unité : si on connaît une moyenne de température en degré Fahrenheit, il est inutile de convertir toutes les valeurs en degrés Celsius pour calculer la moyenne en degrés Celsius, il suffit de ne convertir que la moyenne.

Il est aussi intéressant, pour limiter la taille des nombres, de partir d'un moyenne estimée M_{est} et de calculer la moyenne des $x'_i = x_i - M_{est}$. Dans ce cas, $\bar{x} = M_{est} + \bar{x}'$

2.5 Paramètres de dispersion

2.5.1 Intervalle de variation

L'intervalle de variation représente l'écart entre les extrema. Il s'agit donc de la différence entre la valeur maximum et minimum des valeurs.

2.5.2 Intervalle interquartile

L'intervalle interquartile contient la moitié centrale des observations. Il s'agit donc de la différence entre la valeur du 1^{er} et du 3^{eme} quartile.

2.5.3 Dérivation moyenne

La dérivation moyenne $D.M.$ est égale à la moyenne des valeurs absolues des écarts à la moyenne. On a

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

Pour une série distribuée en p fréquences, on a

$$D.M. = \frac{\sum_{i=1}^{p} f_i|x_i - \bar{x}|}{\sum_{i=1}^{p} f_i}$$

2.5.4 Variance

La variance représente la moyenne de la somme des carrés des écarts à la moyenne.

1. Pour une **population**, on utilise

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{p} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{p} f_i}$$

2. Pour un **échantillon**, on utilise

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{p} f_i(x_i - \bar{x})^2}{\sum_{i=1}^{p} f_i - 1}$$

   - Exemple : On souhaite étudier un caractère relatif à tous les belges mais on n'étudie qu'une partie de la population belge, c'est-à-dire un échantillon.
   - Exemple : L'orsqu'on effectue des mesures en chimie, les n mesures répétées ne sont qu'un échantillon de l'infinité des mesures possibles.

Remarquons que pour n suffisamment grand, on peut faire l'approximation $\sigma^2 \approx s^2$.

2.5.5 Ecart-type et déviation standart

La variance présente le désavantage d'avoir les dimensions des données élevées au carré, c'est pourquoi on utilise l'écart-type qui est la racine carrée de la variance.

1. Pour une population : $\sigma = \sqrt{\sigma^1}$
2. Pour un échantillon : $s^2 = \sqrt{s^2}$ et dans ce cas, on appelle plutôt s la **déviation standart**.

Remarquons que, dans le cas d'une distribution normale (gaussienne),
- 50% des valeurs sont comprises dans l'intervalle $\bar{x} \pm \frac{2}{3}\sigma$
- 68,27% des valeurs sont comprises dans l'intervalle $\bar{x} \pm \sigma$
- 95,45% des valeurs sont comprises dans l'intervalle $\bar{x} \pm 2\sigma$
- 99,75% des valeurs sont comprises dans l'intervalle $\bar{x} \pm 3\sigma$

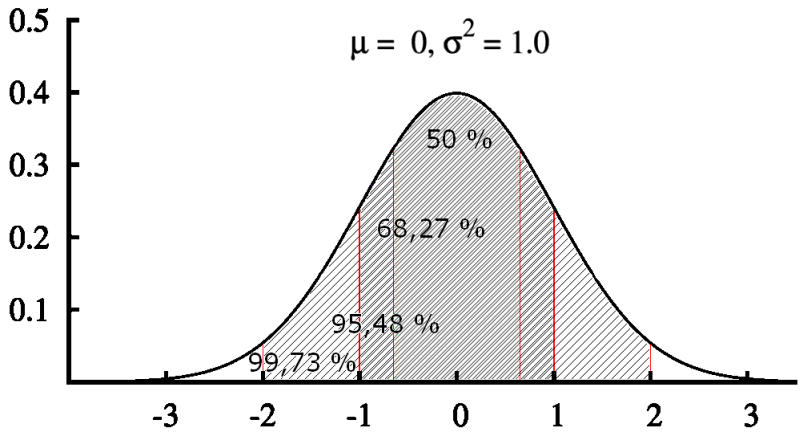

FIGURE 2.3 – Courbe normale (ou gaussienne) centrée réduite.

2.5.6 Propriétés de la variance

1. L'écart-type σ est toujours positif.

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\sum_i (x_i - \bar{x})^2 \\ &= |\frac{1}{n}| \sum_i |x_i - \bar{x}|^2 \\ &\geq 0 \end{aligned}$$

$$\boxed{\Rightarrow \sigma \geq 0}$$

2. Soit $x'_i = ax_i + b$. Donc, $\bar{x}' = a\bar{x} + b$. Dès lors, on a

$$\begin{aligned} \sigma'^2 &= \frac{1}{n}\sum_i (x'_i - \bar{x}')^2 \\ &= \frac{1}{n}\sum_i (ax_i + b - a\bar{x} - b)^2 \\ &= \frac{1}{n}\sum_i (ax_i - a\bar{x})^2 \\ &= a^2 \frac{1}{n}\sum_i (x_i - \bar{x})^2 \\ &= a^2\sigma^2 \end{aligned}$$

$$\boxed{\Rightarrow \sigma' = a\sigma}$$

3. Théorème de König-Huyghens :

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \frac{1}{n}\sum_i x_i^2 - \frac{2}{n}\bar{x}\sum_i x_i + \frac{1}{n}\sum_i \bar{x}^2 \\ &= \frac{1}{n}\sum_i x_i^2 - 2\bar{x}(\frac{1}{n}\sum_i x_i) + \frac{1}{n}n\bar{x}^2 \\ &= \frac{1}{n}\sum_i x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2 \\ &= \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 \end{aligned}$$

$$\boxed{\Rightarrow \sigma^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2}$$

Chapitre 3

Séries statistiques à deux dimensions

3.1 Introduction

Il arrive fréquemment que l'on observe conjointement deux caractères statistiques pour déterminer s'il existe une corrélation entre les deux. Par exemple,

- entre la vitesse et la température d'une réaction chimique,
- entre la longueur et la période d'un pendule,
- entre la pression, la température et le volume d'un gaz, etc.

On exprime cette relation sous forme mathématique à l'aide d'une équation reliant les variables.

Pour chaque individu, on relève la valeur de deux caractères x et y. On obtient alors une liste de couples de nombres (x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_i, y_i), ... (x_n, y_n) que l'on peut présenter sous forme d'un tableau.

Exemple : Masse appliquée (en g) et longueur du ressort (en cm).

Masse en grammes	x_i	7	10	18	20	5	24	12	3
Longueur en cm	y_i	8.5	9	10.5	11	8	11.8	9.4	7.5

Exemple : Moyenne de l'année et note à l'examen pour un échantillon de 8 personnes .

Note de l'année	x_i	8	9	7	15	12	12	10	8
Note à l'examen	y_i	7	9	4	17	13	15	9	13

On peut porter ces points (x_i, y_i) dans un plan muni d'un repère orthonormé Oxy. L'ensemble des points s'appelle le *diagramme de dispersion*.

3.2 Ajustement d'une droite à des données

Le diagramme de dispersion 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 à parier que les phénomènes ne sont pas

corrélés. S'ils semblent dessiner une courbe, on cherchera à déterminer la nature de la courbe en procédant à un ajustement.

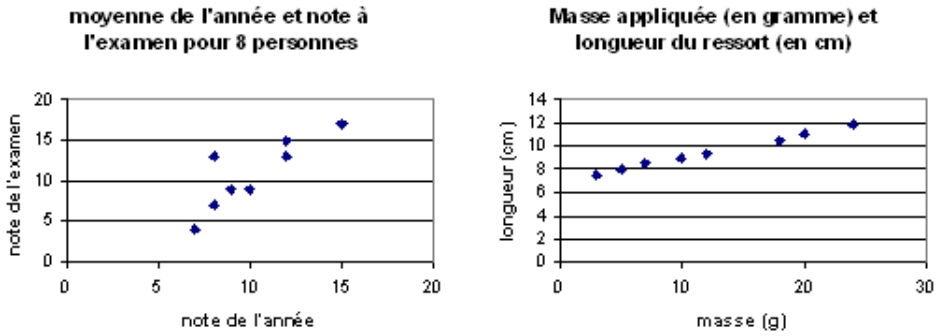

FIGURE 3.1 – Diagramme de dispersion.

Par exemple, pour une relation linéaire $y = ax + b$, deux paramètres a et b sont nécessaires. Pour une relation quadratique $y = ax^2 + bx + c$, trois paramètres a, b et c sont nécessaires, etc. On peut déterminer ces paramètres en choisissant des points sur la courbe (autant de points qu'on a de paramètres).

Le problème de cette méthode est que des observateurs différents obtiendront des courbes différentes et donc des paramètres différents. Pour mettre tout le monde d'accord, on a recherché une méthode rigoureuse permettant de déterminer la « meilleure courbe d'ajustement » aux données.

Il s'agit de la méthode des *moindres carrés*.

3.2.1 Méthode des moindres carrés

Pour obtenir une relation du type $d_{y/x} \equiv y = ax + b$, On cherche à minimiser l'écart quadratique moyen des points à la droite. On doit donc rendre minimum

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Donc, on recherche la droite eds moindres carrés $d_{y/x}$, c'est-à-dire la droite par laquelle la somme des carrés des déviations verticales d_i est minimum.

Soit $y = ax + b$ l'équation de la droite $d_{y/x}$ recherchée. Que valent les constantes a et b pour que la droite $d_{y/x} \equiv f(x) = ax + b$ réponde aux critères énoncés ? On a

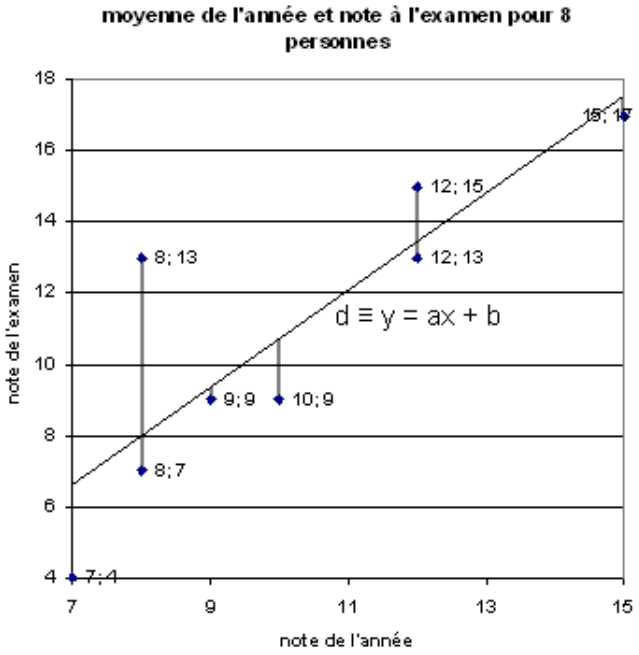

FIGURE 3.2 – On cherche à minimiser l'écart quadratique vertical entre les points et la droite.

$$\begin{array}{rrcc} & d_i & = & y_i - f(x_i) \\ \Leftrightarrow & d_i & = & y_i - ax_i - b \\ \Leftrightarrow & d_i^2 & = & (y_i - ax_i - b)^2 \\ \Leftrightarrow & d_i^2 & = & y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i \\ \Leftrightarrow & \sum_i d_i^2 & = & \sum_i y_i^2 + a^2\sum_i x_i^2 + nb^2 - 2a\sum_i x_iy_i - 2b\sum_i y_i + 2ab\sum_i x_i \\ \Leftrightarrow & \sum_i d_i^2 & = & nb^2 + 2(a\sum_i x_i - \sum_i y_i)b + (\sum_i y_i^2 + a^2\sum_i x_i^2 - 2a\sum_i x_iy_i) \\ \Leftrightarrow & \sum_i d_i^2 & = & \alpha b^2 + \beta b + \gamma \end{array}$$

qui doit être minimum. Comme la dérivée en un extréma est nulle, on va rechercher la valeur de b pour laquelle

$$\frac{d(\sum_i d_i^2)}{db} = 2\alpha b + \beta = 0$$

On a donc que

$$\begin{array}{rcc} 2nb + 2(a\sum_i x_i - \sum_i y_i) & = & 0 \\ \Leftrightarrow \quad b & = & -\frac{2(a\sum_i x_i - \sum_i y_i)}{2n} \\ \Leftrightarrow \quad b & = & \frac{1}{n}\sum_i y_i - a\frac{1}{n}\sum_i x_i \\ \Leftrightarrow \quad b & = & \bar{y} - a\bar{x} \end{array}$$

De même, on a que

$$\begin{array}{rrcc} & d_i & = & y_i - f(x_i) \\ \Leftrightarrow & d_i & = & y_i - ax_i - b \\ \Leftrightarrow & d_i^2 & = & (y_i - ax_i - b)^2 \\ \Leftrightarrow & d_i^2 & = & y_i^2 + a^2x_i^2 + b^2 - 2ax_iy_i - 2by_i + 2abx_i \\ \Leftrightarrow & \sum_i d_i^2 & = & \sum_i y_i^2 + a^2\sum_i x_i^2 + nb^2 - 2a\sum_i x_iy_i - 2b\sum_i y_i + 2ab\sum_i x_i \\ \Leftrightarrow & \sum_i d_i^2 & = & a^2\sum_i x_i^2 + 2(b\sum_i x_i - \sum_i x_iy_i)a + (\sum_i y_i^2 + nb^2 - 2b\sum_i y_i \\ \Leftrightarrow & \sum_i d_i^2 & = & \alpha' a^2 + \beta' a + \gamma' \end{array}$$

qui doit être minimum. Comme la dérivée en un extrémum est nulle, on va rechercher la valeur de a pour laquelle

$$\frac{d(\sum_i d_i^2)}{da} = 2\alpha' a + \beta' = 0$$

On a donc que

$$\begin{array}{rcc} 2\sum_i x_i^2 a + 2(b\sum_i x_i - \sum_i x_iy_i) & = & 0 \\ \Leftrightarrow \quad a & = & -\frac{2(b\sum_i x_i - \sum_i x_iy_i)}{2\sum_i x_i^2} \\ \Leftrightarrow \quad a & = & \frac{\sum_i x_iy_i - b\sum_i x_i}{\sum_i x_i^2} \\ \Leftrightarrow \quad a\sum_i x_i^2 & = & \sum_i x_iy_i - (\bar{y} - a\bar{x})\sum_i x_i \\ \Leftrightarrow \quad a\sum_i x_i^2 - a\bar{x}\sum_i x_i & = & \sum_i x_iy_i - \bar{y}\sum_i x_i \\ \Leftrightarrow \quad a\frac{1}{n}(\sum_i x_i^2 - \bar{x}\sum_i x_i) & = & \frac{1}{n}(\sum_i x_iy_i - \bar{y}\sum_i x_i) \\ \Leftrightarrow \quad a(\frac{1}{n}\sum_i x_i^2 - \bar{x}^2) & = & \frac{1}{n}\sum_i x_iy_i - \bar{x}\bar{y} \\ \Leftrightarrow \quad a\sigma_x^2 & = & \frac{1}{n}\sum_i x_iy_i - \bar{x}\bar{y} \end{array}$$

Or, la quantité $\frac{1}{n}(\sum_i x_iy_i) - \bar{x}\bar{y} = Cov(x,y) = \sigma_{xy}$ est appellée *covariance* de x et de y. On a en réalité

$$\begin{array}{rcc} \sigma_{xy} & = & \frac{1}{n}\sum_i(x_i - \bar{x})(y_i - \bar{y}) \\ & = & \frac{1}{n}\sum_i(x_iy_i - \bar{x}\sum_i y_i - \bar{y}\sum_i x_i + \bar{x}\bar{y}) \\ & = & \frac{1}{n}\sum_i x_iy_i - \bar{x}\frac{1}{n}\sum_i y_i - \bar{y}\frac{1}{n}\sum_i x_i + n\bar{x}\bar{y} \\ & = & \frac{1}{n}\sum_i x_iy_i - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y} \\ & = & \frac{1}{n}\sum_i x_iy_i - \bar{x}\bar{y} \end{array}$$

Dès lors, on peut dire que

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{et} \quad b = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}$$

La droite de régression $d_{y/x}$ de y par rapport à x aura donc comme équation

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x - \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}$$

On peut, par un procédé analogue, construire la droite de régression $d_{x/y} \equiv x = a'y + b'$ de x par rapport à y. Dans ce cas, on cherchera à minimiser les distances horizontales $d_i' = x_i - (a'y_i + b')$.

Selon un raisonnement similaire, on arrive à

$$a' = \frac{\sigma_{xy}}{\sigma_y^2} \quad \text{et} \quad b' = \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2}\bar{y}$$

La droite de régression $d_{x/y}$ de x par rapport à y aura donc comme équation

$$x = \frac{\sigma_{xy}}{\sigma_y^2} y - \bar{x} - \frac{\sigma_{xy}}{\sigma_y^2} \bar{y}$$

Pour résumer :

$$\boxed{\begin{array}{l} d_{y/x} \equiv y = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}) + \bar{y} \\ d_{x/y} \equiv x = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}) + \bar{x} \end{array}}$$

$$\boxed{\sigma_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}$$

3.2.2 Propriété des droites de régression

Soient les droites de régression $d_{y/x}$ et $d_{x/y}$. Quel est leur point d'intersection ?

Pour répondre à cette question, il convient de résoudre le système

$$\left\{ \begin{array}{rcl} y & = & ax + b \\ x & = & a'y + b' \end{array} \right.$$

On a $y = a(a'y + b') + b$

$$\begin{array}{rcrcl} \Leftrightarrow & y - aa'y & = & ab' + b \\ \Leftrightarrow & y(1 - aa') & = & a(\bar{x} - a'\bar{y}) + \bar{y} - a\bar{x} \\ \Leftrightarrow & y(1 - aa') & = & \bar{y} - aa'\bar{y} + a\bar{x} - a\bar{x} \\ \Leftrightarrow & y(1 - aa') & = & \bar{y}(1 - aa') \\ \Leftrightarrow & y & = & \bar{y} \end{array}$$

De même, $x = a'(ax + b) + b'$

$$\begin{array}{rcrcl} \Leftrightarrow & x - aa'x & = & a'b + b' \\ \Leftrightarrow & x(1 - aa') & = & a'(\bar{y} - a\bar{x}) + \bar{x} - a'\bar{y} \\ \Leftrightarrow & x(1 - aa') & = & \bar{x} - aa'\bar{x} + a'\bar{y} - a'\bar{y} \\ \Leftrightarrow & x(1 - aa') & = & \bar{x}(1 - aa') \\ \Leftrightarrow & x & = & \bar{x} \end{array}$$

Dès lors, on peut conclure que le point d'intersection des deux droites de régression est le point de coordonnées $(\bar{x}, \bar{y})$.

$$d_{y/x} \cap d_{x/y} = \{(\bar{x}, \bar{y})\}$$

3.3 Corrélation linéaire

En statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. Donc, il vient naturellement la question : sous quelle conditions est-il justifié d'ajuster une droite à des données ?

3.3.1 Parfaite corrélation

Les droites $d_{y/x}$ et $d_{x/y}$ sont confondues.

$d_{y/x} \equiv y = ax + b$ est confondue avec $d_{x/y} \equiv y = \frac{1}{a'}x - \frac{b'}{a'}$. Elles ont donc la même pente

$$a = \frac{1}{a'} \Leftrightarrow aa' = 1$$

et la même ordonnée à l'origine $b = -\frac{b'}{a'}$.

Par exemple, soit x le rayon d'un cercle et y sa circonférence. On a $d_{y/x} \equiv y = 2\pi x$ et $d_{x/y} \equiv x = \frac{1}{2\pi}y$ qui sont confondues. On vérifie bien que $aa' = 2\pi\frac{1}{2\pi} = 1$.

3.3.2 Indépendance

Les droites $d_{y/x}$ et $d_{x/y}$ sont perpendiculaires.

$d_{y/x} \equiv y = b$ est perpendiculaire à $d_{x/y} \equiv x = b'$. Ce qui signifie que

$$\sigma_{xy} = 0$$

Par exemple, soit x les points obtenus en jettant un dé et y les points obtenus en jettant un autre dé. Ces deux variables sont indépendantes l'une de l'autre et il n'existe pas de relation entre elles.

3.3.3 Corrélation partielle

Une mesure de la corrélation est obtenue par le calcul du coefficient de corrélation linéaire r. Ce coefficient est égal au rapport de leur covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation r est compris entre -1 et 1.

Il s'agit, plus simplement, de la moyenne géométrique des coefficients a et a' des deux droites de régression $d_{y/x}$ et $d_{x/y}$

On a

$$\boxed{r = \pm\sqrt{aa'} = \pm\frac{\sigma_{xy}}{\sigma_x\sigma_y}}$$

- Le signe de r est positif si a et a' sont tous les deux positifs.
- Le signe de r est négatif si a et a' sont tous les deux positifs.
- La covariance seule détermine le signe de a, a' et r : σ_x et σ_y étant toujours positifs.

On admet généralement qu'un ajustement entre des données x et y est valable lorsque

$$0,7 \leq |r| \leq 1$$

Si $|r| < 0,7$, on déduit qu'il n'existe pas de rapport de dépendance entre les deux séries de mesures x et y.

3.3.4 Récapitulatif

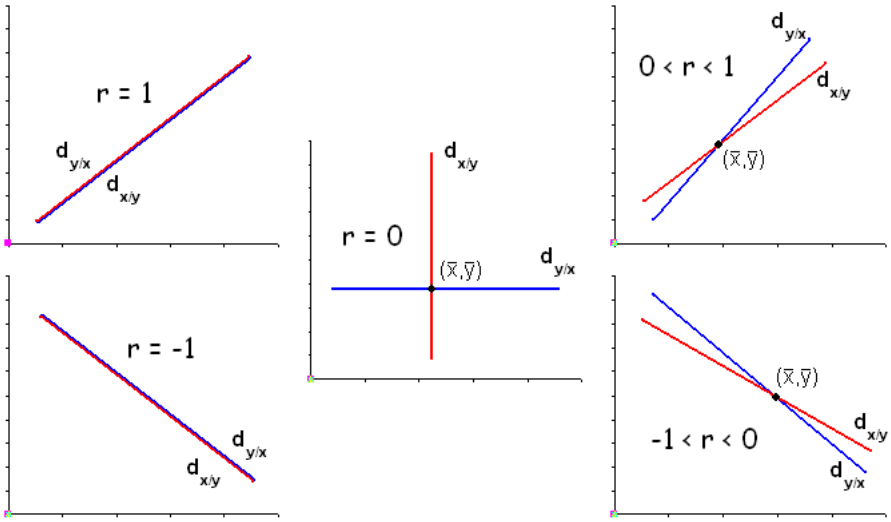

FIGURE 3.3 – Ajustement d'une courbe à des données : récapitulatif.

3.4 Ajustement d'une courbe à des données

Lorsque les points d'un diagramme de dispersion ne semble pas se positionner sur une droite mais plutôt sur une courbe, on recherche l'équation de la courbe qui décrit le mieux le phénomène observé.

3.4.1 Exponentielle, logarithme et puissance

Soient

1. $y = b.a^x$ (Courbe exponentielle)
2. $y = b + a\ln x$ (Courbe logarithmique)
3. $y = b.x^a$ (Courbe puissance)

qu'on peut toujours ramener à une régression linéaire par un changement de variables.

1. $y = b.a^x$: On pose $Y = \ln y$, $A = \ln a$ et $B = \ln b$. Dès lors, $y = e^Y$, $a = e^A$ et $b = e^b$. Il vient alors, en accord avec les propriétés des logarithmes
   - $e^{ln\alpha} = \ln e^\alpha = \alpha$,
   - $\ln(\alpha\beta) = \ln\alpha + \ln\beta$ et
   - $\ln\alpha^\beta = \beta\ln\alpha$ que

   $$\ln y = \ln(b.a^x) = \ln b + x\ln a \Rightarrow Y = Ax + B$$

   Remarquons que si l'équation est de la forme $y = b.e^{ax}$, on aura $\ln y = ax\ln e + \ln b \Rightarrow Y = ax + B$.

2. $y = b + a\ln x$: On pose $X = \ln x$. Dès lors, $x = e^X$ et il vient simplement

   $$y = aX + b$$

3. $y = b.x^a$: On pose $Y = \ln y$, $X = \ln x$ et $B = \ln b$. Dès lors, $y = e^Y$, $x = e^X$ et $b = e^B$. Il vient

   $$\ln y = \ln b + a\ln x \Rightarrow Y = aX + B$$

Remarquons qu'à la place du ln, nous aurions pu utiliser le log. Cependant, il est très simple de passer de l'un à l'autre en considérant la propriété

$$\log_a x = \frac{\ln x}{\ln a}$$

où $\log_a x$ est le logarithme en base 10 de x. (notation : $\log_{10} = \log$)

Remarquons aussi qu'à la place de l'exponentielle e, nous aurions pu utiliser les puissances de 10. Ici aussi, on peu considérer la propriété $a^x = e^{x\ln a}$. (Exemple : $10^x = e^{x\ln 10}$)

3.4.2 Courbe parabolique

Si la courbe est parabolique, son équation nous sera donnée par $y = ax^2 + bx + c$. On cherchera alors à minimiser la somme des carrés des écarts d_i des points à cette courbe, c'est-à-dire

$$\sum_i d_i^2 = \sum_i (y_i - ax_i^2 - bx_i - c)^2$$

En bref, on devra déterminer les constantes a, b et c de la parabole de régression en résolvant le système de 3 équations à 3 inconnues suivant

$$\left\{ \begin{array}{rcl} \sum_i y_i & = & a\sum_i x_i^2 + b\sum_i x_i + nc \\ \sum_i x_i y_i & = & a\sum_i x_i^3 + b\sum_i x_i^2 + c\sum_i x_i \\ \sum_i x_i^2 y_i & = & a\sum_i x_i^4 + b\sum_i x_i^3 + c\sum_i x_i^2 \end{array} \right.$$

3.4.3 Courbe polynomiale

Si la courbe à ajuster est du type $y = a_p x^p + a_{p-1}x^{p-1} + ... + a_2 x^2 + a_1 x + a_0$, un polynôme de degré p, on peut écrire $y = \sum_{j=0}^{p} a_j x^j$ et il faudra résoudre le système de $p+1$ équations à $p+1$ inconnues a_j suivant

$$\left\{ \begin{array}{rcl} \sum_i y_i & = & \sum_{j=0}^{p} a_j \sum_i x_i^j \\ \sum_i x_i y_i & = & \sum_{j=0}^{p} a_j \sum_i x_i^{j+1} \\ \sum_i x_i^2 y_i & = & \sum_{j=0}^{p} a_j \sum_i x_i^{j+2} \\ & ... & \\ \sum_i x_i^p y_i & = & \sum_{j=0}^{p} a_j \sum_i x_i^{j+p} \end{array} \right.$$

Deuxième partie

Exercices

Chapitre 4

Statistiques descriptives

4.1 Rappels fondamentaux

4.1.1 Signe sommatoire

Exercice 1

Calculer $\sum_{i=1}^{6}(2i+1)$

Exercice 2

Ecrire en utilisant la notation du signe sommatoire Σ : 3+5+7+9+11+13+15+17

Exercice 3

Soient

i	1	2	3	4
x_i	5	7	-2	-10

Calculer

$\sum_{i=1}^{4} x_i$, $\sum_{i=2}^{3} x_i$, $\sum_{i=1}^{4} |x_i|$,
$\sum_{i=1}^{4} x_i^2$, $(\sum_{i=1}^{4} x_i)^2$, $\sum_{i=1}^{4}(x_i+2)(x_i-2)$

Exercice 4

Soient

i	1	2	3	4	5	6	7
x_i	32	-16	8	-4	2	-1	$0,5$
y_i	0	2	4	6	8	10	12

1. Calculer $\sum_{i=1}^{7} x_i$ et $\sum_{i=1}^{7} y_i$
2. En déduire $\sum_{i=1}^{7}(x_i+y_i)$, $\sum_{i=1}^{7}(x_i-2)$, $\sum_{i=1}^{7} 2x_i$ et $\sum_{i=1}^{7} \frac{y_i}{2}$
3. Calculer $\sum_{i=1}^{7} x_i^2$ et $\sum_{i=1}^{7} y_i^2$
4. En déduire $\sum_{i=1}^{7}(x_i+y_i)^2$ et $\sum_{i=1}^{7}(x_i-2)(y_i-4)$

4.1.2 Moyennes

Exercice 1 : Moyenne arithmétique

Durant la projection d'un film à la télévision sont apparus quatre spots de publicité dont on a noté la durée : 3 minutes, 5 minutes, 6 minutes, 2 minutes.

Quelle est la durée moyenne d'un spot publicitaire ?

Exercice 2 : Moyenne harmonique

Un véhicule automobile accomplit un parcours de 500 km à la vitesse horaire de 125 km/h. Le trajet retour est accompli à la vitesse de 100 km/h. Calculez la vitesse moyenne de l'automobile.

Exercice 3 : Moyenne quadratique

Un courant électrique alternatif a une intensité efficace de 6 A pendant la moitié de sa période, et une intensité efficace de 3 A pendant l'autre moitié : quelle est son intensité efficace [1] moyenne sur toute la période ?

Exercice 4 : Moyenne géométrique

Le chef du bureau d'achat de poudre d'or de la compagnie Goldfout possède une balance Roberval dont les bras n'ont pas exactement la même longueur (on notera a la longueur d'un bras et b la longueur de l'autre bras). Il s'en suit que les masses marquées placées dans l'un des plateaux équilibrent une masse différente placée dans l'autre plateau. Pour effectuer une pesée, le chef du bureau décide d'opérer deux mesures successives. La première, qui est réalisée en plaçant l'or à gauche donne : $M_1 = 1040$ g. La seconde pesée opérée en plaçant l'or à droite donne : $M_2 = 1160$ g. Le chef de bureau annonce au mineur une masse de 1100 g. Quel est la masse réele de l'or ?

Exercice 5 : Résumé

Soit la distribution de 40 entreprises selon le nombre de micro-ordinateurs utilisés.

Nombre d'ordinateurs	1	2	3	4
Nombre d'entreprises	5	15	10	10

Calculez les valeurs des moyennes arithmétique, harmonique, quadratique et géométrique et classez-les par ordre croissant.

1. L'intensité efficace est celle qui donne le même Effet Joule $W = R\,I^2 t$, avec W en Joules, R = résistance en Ohm, I = intensité en Ampères, t = durée en Secondes.

4.2 Statistiques à 1 dimension

Exercice 1

Un contrôle de connaissances effectué sur une population de 120 étudiants a donné les résultats suivants (les travaux ont été notés sur 10).

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

1. Calculez le mode, la médiane, la moyenne et les quartilles de cette distribution.
2. Calculez l'intervalle de variation, la variance et l'écart-type de cette population.
3. Dessinez l'histogramme, le diagramme en bâtons et le polygone des fréqueces.
4. Dessinez le diagramme et le polygone cumulatif.
5. Estimez le mode, la médiane, la moyenne et les quartilles sur base des graphiques et comparez les avec les résultats calculés précédemment.

Exercice 2

Le kilométrage (exprimé en milliers de km) parcouru par 5000 voitures lors de leur mise hors circulation a donné les résultats suivants :

Classes	$[0,20[$	$[20,40[$	$[40,60[$	$[60,80[$	$[80,100[$	$[100,120[$	$[120,140[$	$[140,160[$	total
n_i	400	650	850	800	1600	400	200	100	5000

1. Calculez le mode, la médiane, la moyenne et les quartilles de cette distribution.
2. Calculez l'intervalle de variation, la variance et l'écart-type de cette population.
3. Dessinez l'histogramme, le diagramme en bâtons et le polygone des fréqueces.
4. Dessinez le diagramme et le polygone cumulatif.
5. Estimez le mode, la médiane, la moyenne et les quartilles sur base des graphiques et comparez les avec les résultats calculés précédemment.

Exercice 3

Le pourcentage de 30 élèves en fin d'année est donné par le tableau suivant :

$84,6$	$79,0$	$76,3$	$60,5$	$61,2$	$74,3$	$85,9$	$83,8$	$83,4$	$66,3$
$77,8$	$76,2$	$81,2$	$68,5$	$80,9$	$59,1$	$53,6$	$71,6$	$67,9$	$71,7$
$63,3$	$64,3$	$53,2$	$62,6$	$67,0$	$73,4$	$69,7$	$57,9$	$65,9$	$65,2$

1. Regroupez les données sous forme de classes. (conseil : utilisez 7 classes $[52,5;57,5[$ ... $[82,5;87,5[$)
2. Calculez le mode, la médiane, la moyenne et les quartilles de cette distribution.
3. Calculez l'intervalle de variation, la variance et l'écart-type de cette population.
4. Dessinez l'histogramme, le diagramme en bâtons et le polygone des fréqueces.
5. Dessinez le diagramme et le polygone cumulatif.
6. Estimez le mode, la médiane, la moyenne et les quartilles sur base des graphiques et comparez les avec les résultats calculés précédemment.

Exercice 4

Le taux de glycémie exprimé en cg/l mesuré sur un échantillon de 20 malades a donné les résultats suivants :

95	93	220	245	100	115	130	112	180	150
215	125	80	95	105	90	120	95	90	85

1. Calculez la moyenne et estimez grossièrement la médiane de cet échantillon.
2. Calculez l'intervalle de variation, la variance et la déviation standart de cet échantillon.

Exercice 5

Le quotient intellectiel des 480 enfants d'une école maternelle est donné dans le tableau suivant :

x_i	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126
n_i	4	9	16	28	45	66	85	72	54	38	27	18	11	5	2

1. Calculez la moyenne et estimez grossièrement la médiane de cet échantillon.
2. Calculez l'intervalle de variation, la variance et la déviation standart de cet échantillon.

Exercice 6

Le tableau ci-dessous représente la distribution exprimée en tonnes des charges maximales supportées par un échantillon des câbles fabriqués par une usine spécialisée.

Classes	n_i
$[9,25;9,75[$	2
$[9,75;10,25[$	5
$[10,25;10,75[$	12
$[10,75;11,25[$	17
$[11,25;11,75[$	14
$[11,75;12,25[$	6
$[12,25;12,75[$	3
$[12,75;13,25[$	1

1. Calculez la charge moyenne supportée par les câbles et estimez grossièrement la médiane de cet échantillon.
2. Calculez l'intervalle de variation et l'intervalle inter-quartille.
3. Calculez la variance et la déviation standart de cet échantillon.

Exercice 7

Lors d'une enquête sur la composition de l'émail dentaire, on a mesuré la concentration en fluor d'un échantillon de 170 dents provenant de différentes régions. On a obtenu les résultats suivants :

Centre de classe	50	150	250	350	450	550	650	750	850
n_i	1	9	20	36	33	16	18	15	6
Centre de classe	950	1050	1150	1250	1350	1450	1550	1650	1750
n_i	4	3	0	1	0	1	3	2	2

1. Calculez la charge moyenne supportée par les câbles et estimez grossièrement la médiane de cet échantillon.
2. Calculez l'intervalle de variation et l'intervalle inter-quartille.
3. Calculez la variance et la déviation standart de cet échantillon.

4.3 Statistiques à 2 dimensions

Exercice 1

On a mesuré la longueur totale (antenne non comprise) x et l'envergure (aile déployée) y de 12 papillons de même espèce. On a obtenu les résultats suivants :

$x\ (mm)$	62	63	64	65	66	67	67	68	68	69	70	71
$y\ (mm)$	66	66	65	68	65	67	68	71	69	68	68	70

1. Calculez $\bar{x}$, $\bar{y}$, σ_x^2, σ_y^2 et σ_{xy}^2 et recherchez l'équation des droites de régression $d_{y/x}$ et $d_{x/y}$.
2. Tracez les deux droites sur le diagramme de dispersion.
3. Calculez le point d'intersection $(\bar{x}, \bar{y})$ et le coefficient de corrélation r.

Exercice 2

On étudie l'influence de la température sur la durée d'incubation des oeufs de grenouille. On a constaté que sur huit échantillons de 200 oeufs chacun, le nombre d'éclosions obtenues au 22^{eme} jour était de :

x	$T^o\ (^oC)$	6,0	6,2	6,4	6,6	6,8	7,0	7,2	7,4
y	Nombre d'éclosions	135	132	150	156	152	155	180	178

1. Recherchez l'équation des droites $d_{y/x}$ et $d_{x/y}$. Tracez ces droites sur le diagramme de dispersion.

2. Calculez le point d'intersection $(\bar{x}, \bar{y})$ et le coefficient de corrélation r.
3. Déterminez le nombre d'éclosions relatif à une température de $7,1\ ^oC$.

Exercice 3

Soit les données (x_i, y_i) suivantes :

x_i	1	3	4	6	8	9	11	14
y_i	1	2	4	4	5	7	8	9

1. Recherchez l'équation des droites $d_{y/x}$ et $d_{x/y}$. Tracez ces droites sur le diagramme de dispersion.
2. Calculez le point d'intersection $(\bar{x}, \bar{y})$ et le coefficient de corrélation r.
3. Déterminez y quand $x = 12$ à partir de $d_{y/x}$.
4. Déterminez x quand $y = 3$ à partir de $d_{x/y}$.

Exercice 4

1. Soit la distribution suivante :

x_i	8	14	27	29	34	43	61
y_i	23	61	160	189	244	330	612

   Ajuster une courbe $y = b\, x^a$ à ces données et calculez le coefficient de corrélation r.

2. Soit la distribution suivante :

x_i	6,9	12,9	19,8	26,7	35,1
y_i	21,4	15,7	12,1	8,5	5,7

   Ajuster une courbe $y = b\, a^x$ à ces données et calculez le coefficient de corrélation r.

3. Soit la distribution suivante :

x_i	29	50	74	103	118
y_i	1,6	23,5	38,0	46,4	48,9

   Ajuster une courbe $y = b + a \ln x$ à ces données et calculez le coefficient de corrélation r.

Exercice 5

Le tableau suivant donne les valeurs expérimentales de la pression p d'une masse d'un gaz pour différentes valeurs du volume V. D'après les principes de la thermodynamique, on a la relation $p\, V^\gamma = C^{te}$ où γ et C^{te} sont des constantes dépendantes des conditions de l'expérience.

Volume	cm^3	54,3	61,8	72,4	88,7	118,6	194,0
pression	$bar(10^5 Pa)$	61,2	49,2	37,6	28,4	19,2	10,1

1. Déterminez l'équation reliant p et V.
2. Estimez la pression p correspondant à un volume V de $100,0\ cm^3$.

Exercice 6

Ajuster les données du tableau suivant par une parabole des moindres carrés :

x_i	0	1	2	3	4	5	6
y_i	$2,4$	$2,1$	$3,2$	$5,6$	$9,3$	$14,6$	$21,9$