Chapitre 2

Étude d'une série statistique à une variable

2.1 Concepts de base

La statistique (descriptive) est née de l'activité de recueil des données répondant aux besoins d'organisation et de gouvernement des grands empires (armée, impôts, organisation des richesses). Il s'agit de l'ensemble des méthodes permettant de décrire et d'analyser des observations (ou données). Ces observations consistent généralement en la mesure d'une ou plusieurs caractéristiques communes sur un ensemble de personnes ou d'objets équivalents.

2.1.1 Population et unité statistique - échantillon

Lors d'une étude statistique, les observations sont faites à partir d'un ensemble précis (ensemble de référence) appelé **population**. Chaque élément de cet ensemble est une **unité statistique** ou un **individu**. Les termes de population et d'individu sont employés aussi bien lorsqu'il s'agit d'êtres humains que d'êtres inanimés, d'êtres abstraits ou d'évènements.
Le **recensement** est l'étude de tous les individus d'une population. Ceci est difficile en pratique, lorsque les populations sont grandes, pour des questions de coût et de temps.
Le **sondage**, par contre, est un recueil d'une partie (ou d'un sous-ensemble) de la population. La partie des individus étudiés s'appelle un **échantillon**. Le recueil d'un échantillon à partir de la population initiale se fait par des techniques statistiques appelées méthodes d'échantillonnage.

Remarque : La population soumise à l'analyse statistique doit être définie avec précision afin que l'ensemble considéré soit déterminé sans ambiguïté, de sorte qu'un individu quelconque puisse y être affecté sans incertitude.

Exemple 2.1.1
Les étudiants de l'université de Kara, les salariés d'une entreprise, la production d'automobiles d'une année, le stock des machines à une date donnée.

2.1.2 Caractères et modalités

Pour décrire une population, on classe les individus selon certains aspects ou attributs que l'on appelle **caractères** ou **variables**.

Exemple 2.1.2
Si la population étudiée est constituée des étudiants de la faculté des sciences, les caractères étudiés peuvent être : le parcours, la spécialité, l'âge, le sexe, le nombre de crédits capitalisés, etc.

Le caractère est un critère de classement, il peut présenter plusieurs situations différentes appelées **modalités**.

Exemple 2.1.3
Les deux modalités du caractère sexe sont : masculin et féminin.

Le nombre de modalités d'un caractère dépend de l'information disponible et du but de l'étude. Il existe deux classes de caractères en Statistique : les caractères qualitatifs et les caractères quantitatifs (ou numériques).

a) Caractères qualitatifs

Les caractères qualitatifs ou variables catégorielles sont des caractères dont les différentes modalités ne sont pas mesurables. Elles sont non numériques dans le sens où les opérations de base n'ont pas de sens.
On distingue deux types de caractères qualitatifs : les caractères qualitatifs nominaux et les caractères qualitatifs ordinaux (ou ordonnés).

Caractères qualitatifs nominaux : les différentes modalités ne sont que des noms ou des catégories qui ne suivent pas un ordre naturel. C'est le cas par exemple de la race, la couleur des yeux, la marque de voiture, le sexe, etc.

Caractères qualitatifs ordonnés : les modalités suivent un ordre naturel ou peuvent être classées dans un ordre spécifique. C'est le cas par exemple du niveau d'éducation, du degré de satisfaction, etc. Ces variables sont repérables selon un type d'échelle plus ou moins légitime. Les catégories pourront alors donner lieu à un codage par les rangs qui ouvrira une autre gamme de traitements possibles proches de ceux des variables quantitatives.

b) Caractères quantitatifs ou numériques

Il s'agit des caractères (ou variables) dont les modalités sont mesurables, c'est-à-dire appartiennent à $\mathbb{R}$. On distingue deux types de variables quantitatives : les variables discrètes et les variables continues.

Variable discrète : les valeurs sont obtenues par dénombrement. C'est le cas par exemple du nombre d'élèves. Une variable discrète peut ne prendre que certaines valeurs isolées (dans $\mathbb{N}$). C'est le cas du nombre de personnes qui composent un ménage. Elle peut prendre une infinité de valeurs dénombrables, mais elle peut aussi n'en prendre que quelques unes.

Variable continue : peut prendre toutes les valeurs à l'intérieur d'un intervalle. Le nombre de modalités possibles d'une telle variable est alors infini. C'est le cas par exemple de la taille, la température, le salaire, le PIB par habitant, etc.

2.2 Étude d'une variable qualitative

2.2.1 Tableau statistique

Modalité	Effectif n_i	Fréquence f_i	Pourcentage p_i	FCC F_i
M_1	n_1	f_1	p_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_i	n_i	$f_i = \frac{n_i}{n}$	$p_i = \frac{n_i}{n} \times 100$	$F_i = \sum_{k=1}^{i} f_k$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
M_m	n_m	f_m	p_m	$F_m = 1$
Total	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

2.2.2 Représentations graphiques

On peut représenter une variable qualitative soit par un diagramme en tuyaux d'orgue ou en barres, soit par un diagramme à secteurs ou camembert.

Exercice 2.1

On considère le brin d'ADN suivant : GGGAGTGTBTATTAABTBBGAA- BTBBBAGBGB-TAGBTBGBGBGGAGTGABBGAGBBTABATGAGGGTA BTGTBAATAABGBATGT-TABBAGAAGGA. En considérant comme modalités les lettres A, B, G et T, faire le dépouillement et donner le tableau statistique. Faire les représentations graphiques précédemment mentionnées.

2.3 Étude d'une variable quantitative discrète

2.3.1 Tableau statistique

Modalité	Effectif n_i	Fréquence f_i	Pourcentage p_i	FCC F_i
x_1	n_1	f_1	p_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	$f_i = \frac{n_i}{n}$	$p_i = \frac{n_i}{n} \times 100$	$F_i = \sum_{k=1}^{i} f_k$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_m	f_m	p_m	$F_m = 1$
Total	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

Exercice 2.2

Le tableau suivant contient la distribution du nombre d'employés par masse salariale.

Masse salariale	*1000000*	*1500000*	*1200000*	*1507000*	*500000*
Nombre d'employés	*20*	*25*	*30*	*35*	*40*

1 *Faire le tableau statistique correspondant.*

2) Représentation en bâtonnets
3) Fonction de répartition

2.3.2 Caractéristiques de tendance centrale

a) Mode

Le mode ($\boldsymbol{Mo}$) d'une distribution est la valeur la plus fréquente dans la série. Il correspond à la valeur de la variable pour laquelle la fréquence est la plus élevée.

b) Percentile

Le $\boldsymbol{p}$-ième percentile est la valeur telle qu'au moins $\boldsymbol{p}$ pour cent des observations ont une valeur inférieure ou égale à cette valeur, et $(\mathbf{100} - \boldsymbol{p})$ pour cent des observations ont une valeur supérieure ou égale à cette valeur.

Calcul du p-ième percentile

Étape 1 : classer les données dans l'ordre croissant.
Étape 2 : calculer l'index $\boldsymbol{i} = \dfrac{\boldsymbol{p}}{\mathbf{100}} \times \boldsymbol{n}$ où $\boldsymbol{n}$ le nombre d'observations.
Étape 3 (décision) : si $\boldsymbol{i}$ n'est pas un nombre entier naturel, la position du $\boldsymbol{p}$-ième percentile correspond à l'entier $\boldsymbol{E(i)} + \mathbf{1}$, où $\boldsymbol{E(i)}$ désigne la partie entière de $\boldsymbol{i}$; si $\boldsymbol{i}$ est un nombre entier, le $\boldsymbol{p}$-ième percentile correspond à la moyenne des valeurs des observations $\boldsymbol{i}$ et $\boldsymbol{i}+\mathbf{1}$.

c) Quartile

Les quartiles sont des percentiles particuliers. Les étapes de calcul des percentiles peuvent être directement appliquées au calcul des quartiles. Il y a trois quartiles :
$\boldsymbol{Q_1}$ = Premier quartile soit 25$^{\text{e}}$ percentile,
$\boldsymbol{Q_2}$ = Deuxième quartile soit 50$^{\text{e}}$ percentile,
$\boldsymbol{Q_3}$ = Troisième quartile soit 75$^{\text{e}}$ percentile.

d) Médiane

La médiane ($\boldsymbol{Me}$) d'une distribution est la valeur de la variable statistique qui partage en deux effectifs égaux les individus de la population rangés selon la valeur croissante du caractère. C'est le cas où $\boldsymbol{p} = \mathbf{50}$.

Exercice 2.3
Les données sur les salaires mensuels initiaux (en euros) des employés d'une agence de voyage sont : 2850 2950 3050 2880 2755 2710 2890 3130 2940 3325 2920 2880. Déterminer le 10$^{\text{e}}$ percentile ainsi que les quartiles $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$ et $\boldsymbol{Q_3}$.

e) Moyenne arithmétique

La moyenne arithmétique d'une variable statistique est la somme, pondérée par les fréquences, des valeurs.

$$\bar{\boldsymbol{x}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{f_i x_i} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{n_i x_i}.$$

f) Moyenne de sous-populations

La moyenne $\bar{\boldsymbol{x}}$ d'une population $\boldsymbol{P}$ composée de $\boldsymbol{p}$ sous-populations $\boldsymbol{P_j}$ peut être exprimée en fonction des moyennes $\bar{\boldsymbol{x}}_{\boldsymbol{j}}$ des sous-populations :

$$\bar{\boldsymbol{x}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{f_j}\bar{\boldsymbol{x}}_{\boldsymbol{j}}, \quad \text{où} \quad \bar{\boldsymbol{x}}_{\boldsymbol{j}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_j}} \boldsymbol{f_{ij}x_{ij}}.$$

La moyenne $\bar{\boldsymbol{x}}$ est donc la moyenne des moyennes des sous-populations.

2.3.3 Caractéristiques de dispersion

a) Variance et écart-type

La variance $\boldsymbol{V(X)}$ se calcule par la formule

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

L'écart-type $\boldsymbol{\sigma_X}$ est la racine carrée de la variance :

$$\boldsymbol{\sigma_X} = \sqrt{\boldsymbol{V(X)}}.$$

Pour une population composée de sous-populations :
La variance d'une population $\boldsymbol{P}$ composée de $\boldsymbol{p}$ sous-populations $\boldsymbol{P_j}$ s'exprime en fonction des variances $\boldsymbol{V_j(X)}$ et des moyennes $\bar{\boldsymbol{x}}_{\boldsymbol{j}}$ des sous-populations.

$$\boldsymbol{V(X)} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{f_j V_j(X)} + \left(\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{f_j}\bar{\boldsymbol{x}}_{\boldsymbol{j}}^{\boldsymbol{2}} - \bar{\boldsymbol{x}}^{\boldsymbol{2}} \right) = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{n_j V_j(X)} + \left(\frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{n_j}\bar{\boldsymbol{x}}_{\boldsymbol{j}}^{\boldsymbol{2}} - \bar{\boldsymbol{x}}^{\boldsymbol{2}} \right).$$

La variance totale est donc la somme de la moyenne arithmétique des variances et de la variance des moyennes.

La moyenne des variances, $\dfrac{\boldsymbol{1}}{\boldsymbol{n}} \displaystyle\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{n_j V_j(X)}$ est appelée la variance intragroupe.

La variance des moyennes, $\dfrac{\boldsymbol{1}}{\boldsymbol{n}} \displaystyle\sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{n_j}\bar{\boldsymbol{x}}_{\boldsymbol{j}}^{\boldsymbol{2}} - \bar{\boldsymbol{x}}^{\boldsymbol{2}}$ est appelée la variance intergroupe.

b) Coefficient de variation

C'est le rapport de la moyenne arithmétique à l'écart type, défini par :

$$\boldsymbol{CV(X)} = \frac{\boldsymbol{\sigma_X}}{\bar{\boldsymbol{x}}}.$$

Le CV permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations. Il donne une bonne idée du degré d'homogénéité d'une série. Il faut qu'il soit le plus faible possible ($< \mathbf{0.15}$ en pratique).

c) Moment d'ordre k

$$\boldsymbol{m_k} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{n_i x_i^k}.$$

d) Moment centré d'ordre k

$$\mu_k = \frac{1}{n}\sum_{i=1}^{m} n_i(x_i - \bar{x})^k.$$

Exercice 2.4
Lors d'une journée, on a relevé les âges de 20 personnes venant se présenter à l'examen théorique du permis de conduire : 19, 20, 20, 24, 37, 22, 58, 24, 23, 20, 19, 19, 21, 22, 20, 27, 33, 20, 22, 21. (a) Préciser la population, l'échantillon et le caractère étudiés. Quelle est la nature de ce caractère ? (b) Déterminer la moyenne arithmétique de cette série. (c) Déterminer la médiane, le mode, la variance, l'écart-type, le coefficient de variation et l'écart inter-quartile de cette distribution d'âges. (d) La distribution est-elle homogène ? Justifier.

2.4 Étude d'une variable quantitative continue

2.4.1 Classe

a) Définition des classes

Pour les variables continues, il est nécessaire que leurs valeurs soient regroupées en classes avant tout traitement pour les besoins d'analyse. Le nombre de classes à retenir dépend de la précision des mesures et de l'effectif de la population étudiée.

b) Amplitude de classe

Par définition, l'amplitude a_i de la classe $[x_i; x_{i+1}[$ est donné par $a_i = x_{i+1} - x_i$.

c) Centre de classe

Pour mener des calculs statistiques sur des séries classées, les classes sont réduites à une seule donnée, à savoir, le centre de classe. Cela revient à considérer que tous les individus peuvent être décrits par ce centre de classe. Par définition, le centre c_i de la classe $[x_i; x_{i+1}[$ est donné par $c_i = \dfrac{x_i + x_{i+1}}{2}$.

2.4.2 Tableau statistique

Classe	Centre c_i	Effectif n_i	Fréquence f_i	Pourcentage p_i	FCC F_i
$[x_1; x_2[$	c_1	n_1	f_1	p_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_i; x_{i+1}[$	c_i	n_i	$f_i = \dfrac{n_i}{n}$	$p_i = \dfrac{n_i}{n} \times 100$	$F_i = \sum_{k=1}^{i} f_k$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_m; x_{m+1}[$	c_m	n_m	f_m	p_m	$F_m = 1$
Total		$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

2.4.3 Représentation graphique

a) Histogramme

Soit la distribution $([\boldsymbol{x_i}; \boldsymbol{x_{i+1}}[, \boldsymbol{n_i})$ d'une variable statistique continue $\boldsymbol{X}$. Pour chaque classe $[\boldsymbol{x_i}; \boldsymbol{x_{i+1}}[$, l'histogramme associe un rectangle de largeur $\boldsymbol{a_i} = \boldsymbol{x_{i+1}} - \boldsymbol{x_i}$ (amplitude da la classe) et de hauteur $\boldsymbol{h_i} = \dfrac{\boldsymbol{f_i}}{\boldsymbol{a_i}}$.

b) Polygone des fréquences

Il lisse l'histogramme de façon à éliminer les ruptures qui dépendent du choix du découpage en classe. Il respecte la compensation des aires ; la surface incluse par la courbe est identique à celle de l'histogramme.

c) Courbe des fréquences cumulées croissantes

Elle représente graphiquement la fonction cumulative ou fonction de répartition définie par $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{F_i}$. En abscisse se trouvent les bornes supérieures des classes et en ordonnée, les fréquences cumulés croissantes.

Exercice 2.5
Compléter le tableau statistique ci-après puis représenter l'histogramme des fréquences, le polygone des fréquences et la courbe cumulative des fréquences croissantes.

Classe	*Effectif* $\boldsymbol{n_i}$
[15 ;20[	*67*
[20 ;30[	*1942*
[30 ;35[	*1364*
[35 ;45[	*2814*
[45 ;55[	*2540*
[55 ;70[	*710*

2.4.4 Caractéristiques de tendance centrale

a) Mode

Pour une variable continue on définit la classe modale. C'est celle dont la fréquence par unité d'amplitude $\boldsymbol{h_i} = \boldsymbol{f_i}/\boldsymbol{a_i}$ est la plus élevée. Après la définition de la classe modale, on déduit la valeur du mode par la formule suivante :

$$\boldsymbol{Mo} = \boldsymbol{x_i} + \frac{|\boldsymbol{\Delta_i}|}{|\boldsymbol{\Delta_i}| + |\boldsymbol{\Delta_{i+1}}|} \times \boldsymbol{a_i}$$

avec $\boldsymbol{Mo}$ le mode, $\boldsymbol{x_i}$ la borne inférieure de la classe modale, $\boldsymbol{a_i}$ l'amplitude de la classe modale, $\boldsymbol{\Delta_i} = \boldsymbol{n_i} - \boldsymbol{n_{i-1}}$, différence entre l'effectif de la classe modale et l'effectif de la classe précédant la classe modale, $\boldsymbol{\Delta_{i+1}} = \boldsymbol{n_{i+1}} - \boldsymbol{n_i}$, différence entre l'effectif de la classe suivant la classe modale et l'effectif de la classe modale.

b) Percentile

Pour déterminer le p-ième percentile Pe dans le cas d'une variable continue, on détermine d'abord l'intervalle auquel appartient ledit percentile : $Pe \in [x_i; x_{i+1}[$ et $F(Pe) = p/100 = \tilde{p}$ avec $F_{i-1} < \tilde{p} \leqslant F_i$. Par la formule de l'interpolation linéaire, on obtient alors :

$$Pe = x_i + a_i \times \frac{\tilde{p} - F_{i-1}}{F_i - F_{i-1}} = x_i + a_i \times \frac{\tilde{p} - F_{i-1}}{f_i}.$$

La **médiane**, Me, est le **50**-ième percentile.

c) Moyenne arithmétique

Toutes ces moyennes se calculent comme dans le cas d'une variable discrète, en y remplaçant les x_i par les c_i. Pour la moyenne arithmétique par exemple, on a :

$$\bar{x} = \sum_{i=1}^{m} f_i c_i = \frac{1}{n} \sum_{i=1}^{m} n_i c_i.$$

Il en est de même pour **la moyenne des sous-populations**.

2.4.5 Caractéristiques de dispersion

a) Etendue

L'étendue ou amplitude de la distribution est la différence entre la plus grande et la plus petite valeur observée : $E = x_{\max} - x_{\min}$.

b) Ecart inter-quartile et rapport inter-quartile

L'écart inter-quartile est donné par $I_Q = Q_3 - Q_1$ et le rapport inter-quartile par $\frac{Q_3}{Q_1}$.

c) Coefficient de dispersion

Le coefficient de dispersion C_{dis} est défini par le rapport de l'écart inter-quartile à la médiane, i.e.

$$C_{\text{dis}} = \frac{Q_3 - Q_1}{Me}$$

Exercice 2.6
On donne la série unidimensionnelle suivante, correspondant à la répartition des entreprises du secteur automobile en fonction de leur chiffre d'affaire en millions d'euros.

Chiffres d'affaires	*Nombre d'entreprises*
moins de 0,25	*137*
[0,25 ;0,5[	*106*
[0,5 ;1[	*112*
[1 ;2,5[	*154*
[2,5 ;5[	*100*
[5 ;10[	*33*

1. *Calculer le chiffre d'affaire moyen et l'écart-type de la série.*
2. *Calculer le mode, les intervalles inter-quartile et inter-décile de la série.*
3. *Calculer la médiane et le coefficient de dispersion de la distribution.*

2.5 Exercices

Exercice 2.7
On estime que la série statistique suivante représente les résultats du concours d'entrée à une grande école.

Note	[0 ;2[	[2 ;4[	[4 ;6[	[6 ;8[	[8 ;10[	[10 ;12[	[12 ;14[	[14 ;16[	[16 ;18[	[18 ;20[
Effectif(n_i)	8	40	75	114	212	286	182	67	15	1

On donne les valeurs numériques suivantes

$$\sum n_i = 1000 \qquad \sum n_i x_i = 10000 \quad \sum n_i x_i^2 = 110136$$

Parmi les propositions suivantes, quelle est **la combinaison de toutes les propositions exactes** ?

1. La moyenne de cette série statistique vaut 10.
2. La médiane vaut 10,0 à $\mathbf{10^{-1}}$ près.
3. Le premier quartile $\boldsymbol{Q_1}$ de cette série statistique est la valeur qui permet de partager l'échantillon en deux parties e telle sorte que 25% de l'échantillon soit supérieur à $\boldsymbol{Q_1}$ et que 75% soit inférieur à $\boldsymbol{Q_1}$. Dans cette série, $\boldsymbol{Q_1}$ appartient à la classe [8 ;10[
4. Le troisième quartile $\boldsymbol{Q_3}$ de cette série statistique appartient à la classe [12 ;14[
5. Dans cette série statistique, la variance est égale à 110,136
6. Le coefficient de variation de cette série statistique est égal à l'écart-type divisé par la variance.

A : 1+2+3+4+5	B : 1+2+4	C : 2+3+5
D : 1+4+5+6	E : 1+2+3+4	F : Aucune des propositions précédentes

Exercice 2.8
On teste un nouveau médicament contre l'épilepsie sur 100 malades. Le traitement est prescrit pendant six mois, et à la fin, on mesure pour chaque patient le nombre de crises d'épilepsie survenues pendant la durée du traitement. Les résultats sont détaillés dans le tableau suivant :

Nombre d'épilepsie pendant les six mois	Effectif	Fréquence cumulée
0	5	5%
1	10	15%
2	30	45%
3	40	85%
4	15	100%

Parmi les caractéristiques de la distribution de la variable « nombre de crises d'épilepsie pendant six mois » , quelle est **la combinaison de toutes les propositions exactes** ?

1. La variable « nombre de crises d'épilepsie pendant six mois » est une variable quantitative ordinale.
2. La moyenne de cette distribution est égale à 2,5.

3. La La médiane de cette distribution est égale à 2.
4. En cas de représentation graphique de la distribution sous forme de diagramme circulaire, l'angle du secteur en degré pour la classe « 2 crises d'épilepsie pendant six mois » est égale à 30°.
5. L'étendue inter quartile de la distribution « nombre de crises d'épilepsie pendant six mois » est égale à 4.
6. En cas de représentation graphique de la distribution sous forme de diagramme en barre, les proportions sont ordonnées.

A : 2+3+6	B : 1+2+5	C : 1+2+4+6
D : 2+3+4+6	E : 2+6	F : Aucune des propositions précédentes

Exercice 2.9
La distribution de la glycémie chez les sujets malades suit une loi normale de moyenne 1,6g/l et d'écart-type 0,2g/l. Chez les sujets sains, la glycémie suit une loi normale de moyenne 1g/l et d'écart-type 0,1g/l. La fréquence de la maladie est de 10%.
Quelle est la **combinaison des propositions exactes** ?

1. Le pourcentage des sujets malades ayant une glycémie comprise entre 1,2 et 1,4g/l est de 32%
2. Le pourcentage des sujets malades ayant une glycémie comprise entre 1,2 et 1,4g/l est de 13,6%
3. Le pourcentage des sujets sains ayant une glycémie strictement supérieure à 1,2g/l est 5%
4. La valeur médiane de la glycémie chez les sujets malades est de 1,6g/l%.
5. Chez les sujets sains, l'étendue inter quartile correspond aux valeurs comprises entre 0,9g/l et 1,1g/l
6. Si une variable aléatoire suit une loi normale de moyenne $\boldsymbol{\mu}$ et d'écart-type $\boldsymbol{\sigma}$ alors la probabilité d'observer une valeur entre $\boldsymbol{\mu - 2\sigma}$ et $\boldsymbol{\mu + \sigma}$ est égale à 81,9% à $10^{-}\mathbf{1}$ près.

A : 2+4+6	B : 1+3+4+5	C : 1+4+5+6
D : 2+4+5+6	E : 3+5	F : Aucune des propositions précédentes

Exercice 2.10
On donne à 100 malades diabétiques un nouveau traitement destiné à diminuer le taux de sucre dans le sang (taux de glycémie) et améliorer leur qualité de vie (exécrable, médiocre, convenable) chez tous les patients. Les taux de glycémie mesurés chez les 100 patients sont présentés dans le tableau suivant :

Taux de glycémie (en mmol/l)	[1 ;10[	[10 ;20[	[20 ;30[	[30 ;40[	[40 ;50[
Nombre de sujets	10	5	20	40	25

Quelle est la **combinaison de toutes les propositions exactes** ?

1. La variable « qualité de la vie » est une variable qualitative ordinale
2. La variable « taux de glycémie » est une variable quantitative discrète
3. En cas de discrétisation d'une variable quantitative, les classes sont toujours égales

4. La classe du $\mathbf{2}^{e}$ quartile est la classe [30 ;40[
5. L'étendue inter quartile correspond à 75% des observations
6. En cas de distribution symétrique du taux de glycémie, la médiane et la moyenne sont identiques et le mode est différent.

A : 2+3+4	B : 1+2+5	C : 3+5+6
D : 1+4	E : 1+2+4	F : Aucune des propositions précédentes

Exercice 2.11
On a effectué un dosage biologique $\boldsymbol{X}$ sur sur un échantillon de 100 sujets normaux avec les résultats suivants :

Valeur de $\boldsymbol{X}$	[5 ;10[	[10 ;15[	[15 ;20[	[20 ;25[	[25 ;30[
Nombre de sujets	5	10	30	40	15

Quelle est la **combinaison de toutes les propositions exactes** ?

1. Le mode de la distribution de $\boldsymbol{X}$ appartient à la classe [20 ;25[
2. La médiane de la distribution de $\boldsymbol{X}$ est égale à 17,5
3. L'étendue de la distribution de $\boldsymbol{X}$ est égale à 20
4. Le pourcentage des sujets ayant une valeur inférieure à 25 et supérieure à 10 est égal à 85%
5. Si l'on représente les résultats en diagramme sectoriel, l'angle du secteur (en degrés) pour la classe [20 ;25[est égal à 140°
6. Dans un diagramme en barre, les proportions sont ordonnées

A : 1	B : 2+3	C : 1+5+6
D : 1+6	E : 2+3+5	F : Aucune des propositions précédentes

Exercice 2.12
Dans une enquête portant sur les bronchites infectieuses, on recueille les informations pour chacun des 150 enfants participant à l'étude : l'âge (en année), le sexe, la profession du père, le nombre de cigarettes fumées par le père (en nombre de cigarettes par jour), l'antécédent familial d'allergie (oui, non), le nombre de bronchites infectieuse dans l'année précédente :
Quelle est la **combinaison de toutes les propositions exactes** ?

1. L'âge est une variable qualitative discrète
2. Le sexe est une variable qualitative binaire
3. La profession du père est une variable qualitative ordinale
4. e nombre de cigarettes fumées par le père est une variable qualitative ordinale
5. l'antécédent familial d'allergie est une variable qualitative nominale
6. le nombre de bronchites infectieuse dans l'année précédente est une variable quantitative discrète

A : 1+2+4	B : 2+4+6	C : 1+2+3+6
D : 2+6	E : 3+4+5	F : Aucune des propositions précédentes

Exercice 2.13

On teste un nouveau médicament contre la migraine chez 100 malades. Le traitement est prescrit pendant un mois et à la fin on mesure pour chaque patient le nombre de crises de migraines survenues pendant le mois. Les résultats sont détaillés dans le tableau suivant :

Nombre de crises de migraines pendant le mois	Effectif	Fréquence cumulée
0	5	5%
1	10	15%
2	30	45%
3	40	85%
4	15	100%

Parmi les caractéristiques de la distribution de la variable « nombre de crises de migraines survenues pendant le mois », quelle est la **combinaison de toutes les propositions exactes** ?

1. La variable « nombre de crises de migraines survenues pendant le mois » est une variable quantitative ordinale
2. La moyenne de cette distribution est égale à 2,5
3. La médiane de cette distribution est égale à 2
4. Si l'on représente les résultats sous la forme d'un diagramme circulaire, l'angle du secteur (en degrés) pour la variable « 2 crises de migraines survenues pendant le mois » est égal à 30°.
5. L'étendue inter quartile de cette distribution est égale à 4
6. Si l'on représente les résultats sous la forme d'un diagramme en barre, les proportions sont ordonnées

A : 2+3+6	B : 1+2+5	C : 1+2+4+6
D : 2+3+4+6	E : 2+6	F : Aucune des propositions précédentes

Chapitre 3

Étude d'une série statistique à deux variables

3.1 Tableaux de contingence, distributions marginales et conditionnelles

3.1.1 Tableau de contingence : croisement de deux variables

Il est courant d'étudier une population à l'aide de plusieurs caractères. Pour deux variables $\boldsymbol{X}$ et $\boldsymbol{Y}$, on a le tableau suivant, qui est appelé tableau de contingence.

$\boldsymbol{X}$ \ $\boldsymbol{Y}$	$\boldsymbol{y_1}$	$\cdots$	$\boldsymbol{y_j}$	$\cdots$	$\boldsymbol{y_p}$	Eff. marg. de $\boldsymbol{X}$
$\boldsymbol{x_1}$	$\boldsymbol{n_{11}}$	$\cdots$	$\boldsymbol{n_{1j}}$	$\cdots$	$\boldsymbol{n_{1p}}$	$\boldsymbol{n_{1\cdot}} = \sum_{j=1}^{p} \boldsymbol{n_{1j}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x_i}$	$\boldsymbol{n_{i1}}$	$\cdots$	$\boldsymbol{n_{ij}}$	$\cdots$	$\boldsymbol{n_{ip}}$	$\boldsymbol{n_{i\cdot}} = \sum_{j=1}^{p} \boldsymbol{n_{ij}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x_m}$	$\boldsymbol{n_{m1}}$	$\cdots$	$\boldsymbol{n_{mj}}$	$\cdots$	$\boldsymbol{n_{mp}}$	$\boldsymbol{n_{m\cdot}} = \sum_{j=1}^{p} \boldsymbol{n_{mj}}$
Eff. marg. de $\boldsymbol{Y}$	$\boldsymbol{n_{\cdot 1}} = \sum_{i=1}^{m} \boldsymbol{n_{i1}}$	$\cdots$	$\boldsymbol{n_{\cdot j}} = \sum_{i=1}^{m} \boldsymbol{n_{ij}}$	$\cdots$	$\boldsymbol{n_{\cdot p}} = \sum_{i=1}^{m} \boldsymbol{n_{ip}}$	$\boldsymbol{n} = \sum_{i=1}^{m} \boldsymbol{n_{i\cdot}} = \sum_{j=1}^{p} \boldsymbol{n_{\cdot j}}$

L'effectif $\boldsymbol{n_{ij}}$ de la classe $(\boldsymbol{i}, \boldsymbol{j})$ est le nombre d'individus de la population qui présentent simultanément la modalité $\boldsymbol{x_i}$ de la variable $\boldsymbol{X}$ et la modalité $\boldsymbol{y_j}$ de la variable $\boldsymbol{Y}$. La distribution s'écrit $(\boldsymbol{x_i}, \boldsymbol{y_j}, \boldsymbol{n_{ij}})$. Tous les individus présentant ces deux modalités sont comme équivalents. Le total des lignes et le total des colonnes définissent les distributions marginales. Une ligne ou une colonne constitue une distribution conditionnelle.

3.1.2 Fréquence conjointe

On appelle fréquence conjointe de x_i et y_j, la proportion f_{ij} d'individus qui présentent simultanément les modalités x_i et y_j. Elle s'obtient par :

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

3.1.3 Distribution marginale : profile ligne et profile colonne

La série $(x_i, n_{i.})$ définit le profile ligne. La fréquence marginale des individus présentant la modalité x_i est

$$f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{n}.$$

De même, la série $(y_j, n_{.j})$ constitue le profile colonne. La fréquence marginale des individus présentant la modalité y_j est

$$f_{.j} = \sum_{i=1}^{m} f_{ij} = \frac{n_{.j}}{n}.$$

On a la relation suivante : $\sum_{i=1}^{m}\sum_{j=1}^{p} f_{ij} = \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{p} f_{.j} = 1$.

3.1.4 Distribution conditionnelle

La fréquence conditionnelle de la modalité x_i sachant la modalité y_j est

$$f(x_i \mid y_j) = \frac{n_{ij}}{n_{.j}}.$$

La fréquence conditionnelle de la modalité y_j sachant la modalité x_i est

$$f(y_j \mid x_i) = \frac{n_{ij}}{n_{i.}}.$$

La relation entre fréquences marginales et fréquences conditionnelles est

$$f_{ij} = f(x_i \mid y_j) \times f_{.j} = f(y_j \mid x_i) \times f_{i.}$$

Exercice 3.1

La tableau suivant présente les revenus des ménages (RM) en fonction du niveau d'étude (NE) du chef de ménage.

RM / *NE*	*[0 ;25[*	*[25 ;50[*	*[50 ;75[*	*[75 ;100[*	*[100 ;125[*	*Total*
Secondaire sans diplôme	*9285*	*4093*	*1589*	*541*	*354*	*15862*
Secondaire avec diplôme	*10150*	*9821*	*6050*	*2737*	*2028*	*30786*
Université sans diplôme	*5011*	*9221*	*5813*	*3215*	*3120*	*26380*
Licence	*2138*	*3985*	*3952*	*2698*	*4748*	*17521*
Master et plus	*813*	*1497*	*1815*	*1589*	*3765*	*9479*
Total	*27397*	*28617*	*19219*	*10780*	*14015*	*100028*

(a) Calculer les fréquences marginales. (b) Quel est le pourcentage des chefs de ménages diplômés d'université et ayant un revenu compris entre 50 et 75 mille francs ? (c) Calculer le revenu moyen des ménages, le revenu le plus fréquent, la médiane, et l'écart-type des revenus.

3.2 Ajustement linéaire

L'ajustement linéaire consiste à trouver une relation de la forme $\boldsymbol{Y = aX + b}$ entre la variable $\boldsymbol{X}$ appelée variable explicative ou encore variable indépendante et la variable $\boldsymbol{Y}$ appelée variable expliquée ou encore variable dépendante. La représentation graphique de $\boldsymbol{Y}$ en fonction de $\boldsymbol{X}$, ou plus précisément du nuage des points $(\boldsymbol{x_i, y_i})$ permet d'avoir une idée sur le type de relation entre $\boldsymbol{X}$ et $\boldsymbol{Y}$.

3.2.1 Covariance entre deux variables

La covariance entre la variable $\boldsymbol{X}$ et la variable $\boldsymbol{Y}$ est le nombre noté $\mathbf{cov}(\boldsymbol{X, Y})$ ou encore $\boldsymbol{\sigma_{XY}}$ et défini par $\mathbf{cov}(\boldsymbol{X, Y}) = \boldsymbol{\sigma_{XY}} = \dfrac{\mathbf{1}}{\boldsymbol{n}} \displaystyle\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} \boldsymbol{n_k}(\boldsymbol{x_k} - \bar{\boldsymbol{x}})(\boldsymbol{y_k} - \bar{\boldsymbol{y}})$ dans le cas de la série $(\boldsymbol{x_k, y_k, n_k})$ et par $\mathbf{cov}(\boldsymbol{X, Y}) = \boldsymbol{\sigma_{XY}} = \dfrac{\mathbf{1}}{\boldsymbol{n}} \displaystyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{n_{ij}}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_j} - \bar{\boldsymbol{y}})$ dans le cas de la série $(\boldsymbol{x_i, y_j, n_{ij}})$.

3.2.2 Méthode des moindres carrées ordinaires

Cette méthode consiste à déterminer la fonction affine $\boldsymbol{y = ax + b}$ telle que les carrés des erreurs entre les valeurs observées $\boldsymbol{y_i}$ et les valeurs estimées $\boldsymbol{\hat{y}_i = ax_i + b}$ soient les plus petites possibles. En clair, il s'agit de trouver les valeurs de $\boldsymbol{a}$ et $\boldsymbol{b}$ qui minimisent la fonction $\boldsymbol{g(a, b)} = \displaystyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\varepsilon_i^2} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i - ax_i - b})^{\mathbf{2}}$.

On démontre que les valeurs $\boldsymbol{a}$ et $\boldsymbol{b}$ qui minimisent la fonction de deux variables $\boldsymbol{g(a, b)}$ sont :

$$\boldsymbol{a = \frac{\sigma_{XY}}{\sigma_X^2}} \quad \text{et} \quad \boldsymbol{b = \bar{y} - a\bar{x}.}$$

La droite d'équation $\boldsymbol{y = ax + b}$ pour les valeurs de $\boldsymbol{a}$ et $\boldsymbol{b}$ précédemment trouvées s'appelle la droite de régression linéaire de $\boldsymbol{Y}$ sur $\boldsymbol{X}$ et se note $\boldsymbol{D_{Y/X}}$.

De la même façon, on définit la droite de régression linéaire de $\boldsymbol{X}$ sur $\boldsymbol{Y}$ notée $\boldsymbol{D_{X/Y}}$ et ayant pour équation $\boldsymbol{x = a'y + b'}$, avec :

$$\boldsymbol{a' = \frac{\sigma_{XY}}{\sigma_Y^2}} \quad \text{et} \quad \boldsymbol{b' = \bar{x} - a'\bar{y}.}$$

3.2.3 Coefficient de corrélation linéaire

Le coefficient de corrélation linéaire entre la variable $\boldsymbol{X}$ et la variable $\boldsymbol{Y}$ est le nombre

$$\boldsymbol{r = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.}$$

La corrélation linéaire entre $\boldsymbol{X}$ et $\boldsymbol{Y}$ est d'autant plus importante que $\boldsymbol{D = r^2}$ (appelé coefficient de détermination) est proche de $\mathbf{1}$.

3.2.4 Propriétés importantes

1. $\boldsymbol{\rho_{XY} = \sqrt{aa'}}$.
2. La variance de $\boldsymbol{Y}$ se décompose de la façon suivante (formule de décomposition de la variance) :
$$\mathbf{Var}(\boldsymbol{Y}) = \mathbf{Var}(\hat{\boldsymbol{Y}}) + \mathbf{Var}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$
avec $\mathbf{Var}(\boldsymbol{Y})$ = variance totale, $\mathbf{Var}(\hat{\boldsymbol{Y}})$ = variance expliquée et $\mathbf{Var}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$ = variance résiduelle. La proportion de variance expliquée par le modèle linéaire est $\boldsymbol{R^2} = \dfrac{\mathbf{Var}(\hat{\boldsymbol{Y}})}{\mathbf{Var}(\boldsymbol{Y})}$.
3. $\boldsymbol{R^2 = r^2}$.

On note que $\boldsymbol{r}$ indique le signe de la liaison et le $\boldsymbol{R^2}$ explicite la proportion de la variance qui pourrait être expliquée si une relation entre les deux variables existait. Les fluctuations de la variable dépendante sont expliquées et non causées par les mouvements de la variable indépendante.

3.2.5 Ajustement linéaire par la méthode de Mayer

Cette méthode est empirique et ne se repose sur aucun critère d'erreur à minimiser. Elle peut s'avérer efficace en présence de valeurs aberrantes.

Principe : La distribution est partagée en deux groupes comportant un nombre égal de données (si $\boldsymbol{n}$ est pair) et un nombre égal de données à l'unité près (si $\boldsymbol{n}$ est impair). On détermine dans chaque groupe le centre de gravité ou point moyen. Soient $\boldsymbol{G_1(x_I, y_I)}$ et $\boldsymbol{G_2(x_{II}, y_{II})}$ les centres de gravité des deux groupes :

$$\boldsymbol{x_I = \frac{1}{n_1}\sum_{i=1}^{m_1} n_i x_i;} \quad \boldsymbol{y_I = \frac{1}{n_1}\sum_{i=1}^{m_1} n_i y_i;} \quad \boldsymbol{x_{II} = \frac{1}{n_2}\sum_{i=1}^{m_2} n_i x_i} \quad \text{et} \quad \boldsymbol{y_{II} = \frac{1}{n_2}\sum_{i=1}^{m_2} n_i y_i;}$$

avec $\boldsymbol{n_1 = \sum_{i=1}^{m_1} n_i}$, $\boldsymbol{n_2 = \sum_{i=1}^{m_2} n_i}$ et $\boldsymbol{n = n_1 + n_2}$.

La droite d'ajustement linéaire de $\boldsymbol{Y}$ sur $\boldsymbol{X}$ par la méthode de Mayer est la droite $(\boldsymbol{G_1G_2})$ d'équation $\boldsymbol{y = ax + b}$ avec :

$$\boldsymbol{a = \frac{y_I - y_{II}}{x_I - x_{II}},} \quad \text{et} \quad \boldsymbol{b = y_I - ax_I = y_{II} - ax_{II}.}$$

Exercice 3.2
On considère la série suivante :

$\boldsymbol{X}$	*1*	*2*	*3*	*4*	*5*	*6*
$\boldsymbol{Y}$	*2,5*	*3*	*4*	*5*	*5,5*	*7*

(a) Déterminer l'ajustement linéaire de $\boldsymbol{Y}$ sur $\boldsymbol{X}$ par la méthode des moindres carrées ordinaires et par la méthode de Mayer. (b) Pour chaque méthode, donner la décomposition de la variance totale $\mathbf{Var}(\boldsymbol{Y})$. (c) Représenter le nuage de points ainsi que les deux droites de régression.

3.3 Exercices

Exercice 3.3
L'observation des prix et des quantités sur un marché de la tomate a donné les résultats suivants :

Quantités $\boldsymbol{x}$ en kg	10	20	35	50	70	90	110	130
Prix $\boldsymbol{y}$ au kg en kFCFA	5	3.75	2.75	2.25	1.75	1.25	0.8	0.5

Ainsi, une quantité de 35 kg de tomates est vendue au prix de **2750** FCFA le kg.

1. Représenter graphiquement le nuage de points.
2. (a) Déterminer la droite d'ajustement linéaire $\boldsymbol{y = ax + b}$ qui permet d'expliquer le prix au kg par la quantité achetée.
   (b) Calculer le coefficient de corrélation entre $\boldsymbol{x}$ et $\boldsymbol{y}$ et expliquer son signe.
   (c) Prévoir le prix d'un kg de tomates pour un achat de 140 kg.
3. (a) Chercher maintenant un ajustement par une fonction logarithme de la forme $\boldsymbol{y = a\ln(x) + b}$. On pourra poser $\boldsymbol{z = \ln(x)}$ et se ramener à un ajustement linéaire $\boldsymbol{y = az + b}$.
   (b) Calculer le coefficient de corrélation entre $\boldsymbol{z}$ et $\boldsymbol{y}$.
   (c) Prévoir le prix au kg pour un achat de 140 kg.
4. Lequel de ces deux ajustements vous semble le plus judicieux ? Justifier la réponse.

Exercice 3.4
On a relevé la distance de freinage $\boldsymbol{y}$ d'un véhicule (distance parcourue par le véhicule entre le moment où le conducteur commence à freiner et l'arrêt du véhicule) à décélération constante sur une route plane sèche pour différentes vitesses $\boldsymbol{x}$ en km/h. Les données sont consignées dans le tableau suivant :

$\boldsymbol{x}$ (en km/h)	30	50	90	110	130
$\boldsymbol{y}$ (en m)	5.1	14.1	45.5	68	95

Les experts affirment que la distance d'arrêt d'un véhicule est proportionnelle à son énergie cinétique et donc au carré de sa vitesse. On cherche à confirmer cela à l'aide d'un modèle du type $\boldsymbol{y = kx^{\alpha}}$ (appelé modèle de régression puissance) où $\boldsymbol{k}$ et $\boldsymbol{\alpha}$ sont des constantes à déterminer.

1. Montrer en faisant un changement de variable simple, qu'on peut se ramener à un problème de régression linéaire simple puis déterminer les coefficients $\boldsymbol{k}$ et $\boldsymbol{\alpha}$ par la méthode des moindres carrés.
2. Les résultats obtenus confirment-ils l'affirmation des experts ?
3. Confirmer la qualité du modèle en calculant le coefficient de corrélation linéaire entre $\boldsymbol{y}$ et $\boldsymbol{x^2}$.
4. Estimer la distance de freinage correspondant à une vitesse de 70 km/h.

Exercice 3.5
Dans une expérience de cinétique chimique, l'on a mesuré la concentration $\boldsymbol{x}$ d'une substance chimique en fonction du temps $\boldsymbol{t}$ et obtenu les valeurs numériques suivantes :

$\boldsymbol{t}$	20	30	40	50	60	70	80
$\boldsymbol{x}$	0.05	0.10	0.27	0.52	0.80	0.90	0.95

1. Faire le nuage de points.
2. Au vu de la forme du nuage, l'on se propose d'exprimer la grandeur

$$\boldsymbol{y = \ln\left(\frac{x}{1-x}\right)}$$

comme une fonction linéaire du temps $\boldsymbol{y = A + Kt}$ où la constante $\boldsymbol{K}$ est appelée constante de vitesse de la réaction étudiée. Déterminer les coefficients $\boldsymbol{A}$ et $\boldsymbol{K}$.

3. Donner l'expression de $\boldsymbol{x}$ en fonction de $\boldsymbol{t}$.

Exercice 3.6
La concentration molaire $\boldsymbol{C}$ d'une espèce chimique est mesurée au cours du temps. On obtient les données suivantes :

$\boldsymbol{t}$ (s)	20	40	60	80	100	120
$\boldsymbol{C}$ (mol.l^{-1})	$\boldsymbol{278 \times 10^{-3}}$	$\boldsymbol{192 \times 10^{-3}}$	$\boldsymbol{147 \times 10^{-3}}$	$\boldsymbol{119 \times 10^{-3}}$	$\boldsymbol{100 \times 10^{-3}}$	$\boldsymbol{86 \times 10^{-3}}$

1. Représenter les données par un nuage de points.
2. Considérons le modèle de régression linéaire $\boldsymbol{C = at + b}$. En observant uniquement la variation de la concentration en fonction du temps, dire quel serait le signe de $\boldsymbol{a}$ et que vaudrait la limite de $\boldsymbol{C}$ quand $\boldsymbol{t \to +\infty}$. En déduire que le modèle de régression linéaire n'est pas réaliste.
3. On propose le modèle $\boldsymbol{\frac{1}{C} = at + b}$.
   (a) Calculer la limite de $\boldsymbol{C}$ quand $\boldsymbol{t \to +\infty}$. Ce modèle vous semble-t-il raisonnable ?
   (b) En faisant le changement de variable $\boldsymbol{x = \frac{1}{C}}$, déterminer les coefficients $\boldsymbol{a}$ et $\boldsymbol{b}$.
   (c) Prévoir la valeur de $\boldsymbol{C}$ pour $\boldsymbol{t = 180}$.

Exercice 3.7
Lors d'un exercice de travaux pratiques, un étudiant doit déterminer la valeur $\boldsymbol{R}$ d'une résistance. L'étudiant, qui n'a à sa disposition qu'un ampèremètre et un voltmètre, pense utiliser la loi d'Ohm pour déterminer la valeur de $\boldsymbol{R}$. Il réalise 8 mesures d'intensité et de tension aux bornes de la résistance. Les valeurs obtenues sont les suivantes :

$\boldsymbol{I}$ (A)	0.001	0.005	0.01	0.015	0.02	0.025	0.03	0.035
$\boldsymbol{U}$ (V)	0.48	2.21	4.41	6.95	8.92	11.08	13.6	15.77

1. Faire un nuage de points
2. Déterminer les coefficients de la droite de régression $\boldsymbol{U = aI + b}$ et en déduire la valeur de $\boldsymbol{R}$.

Tables statistiques

Table des valeurs de la fonction de répartition de la loi normale centrée réduite.

$$\boldsymbol{\Phi(x) = \mathbb{P}(X \leqslant x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)\, dt.}$$

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5150	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5754
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7518	0,7549
0,70	0,7580	0,7612	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8079	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9430	0,9441
1,60	0,9452	0,9463	0,9474	0,9485	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9700	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9762	0,9767
2,00	0,9773	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9865	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9980	0,9980	0,9981
2,90	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Grandes valeurs de x

x	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8
$\Phi(x)$	0,99865	0,99903	0,99931	0,99952	0,99966	0,99977	0,99984	0,99989	0,99993

Quantiles usuels de la loi normale centrée réduite.

Pour tout $p \in]0,1[$, le quantile d'ordre p de la loi normale centrée réduite est $\Phi^{-1}(p) = z$ tel que $\mathbb{P}(X \leqslant z) = p$.

p	0,5	0,75	0,8	0,9	0,95	0,975	0,99	0,995
$\Phi^{-1}(p)$	0	0,67	0,84	1,28	1,64	1,96	2,33	2,58