

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abou Bekr Belkaid Tlemcen



Statistiques descriptives et exercices



Rappels de cours et exercices corrigés sur la statistique descriptive

Abdennasser Chekroun

Courriels : abdennasser.chekroun@gmail.com / chekroun@math.univ-lyon1.fr

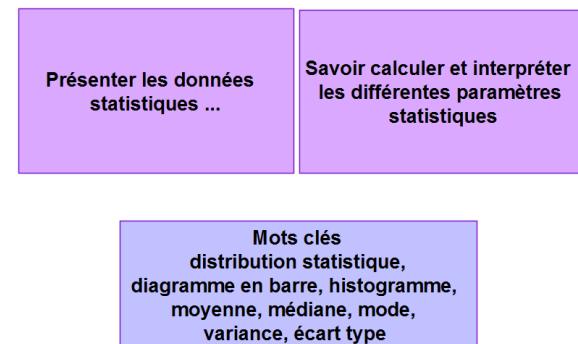
2017 - 2018

Préambule

Le cours a pour but d'initier les étudiants aux principes de base de la statistique. Il s'adresse aux étudiants de deuxième année de Licence des sciences et techniques, ainsi qu'aux étudiants des autres filières. Il est le fruit d'un enseignement de mathématiques pour les ingénieurs dispensé à la faculté de technologie de l'université de Tlemcen. Ce cours vise principalement à introduire et faire méditer les concepts fondamentaux et méthodes élémentaires de la statistique pour permettre un apprentissage autonome ultérieur de méthodes complémentaires.

On veut développer le sens critique nécessaire lors de la mise en œuvre et de l'interprétation d'un traitement statistique. Pour cela, on introduira et utilisera un cadre mathématique rigoureux. Nous fournirons autant d'exemples et de figures nécessaires afin d'obtenir une meilleure compréhension du cours.

La statistique descriptive a pour but d'étudier un phénomène à partir de données. Cette description se fait à travers la présentation des données (la plus synthétique possible), leur représentation graphique et le calcul de résumés numériques.



La place de ce cours dans le future métier des étudiants :

- Analyse des données (outils scientifiques permettant de résumer un ensemble de données afin de mettre en évidence l'information).
- Simulations (processus stochastique - variable temporelle)
- Prédiction et décisions (probabilités de risque ou d'occurrence)

Table des matières

1 Généralités sur la statistique	1
1.1 Vocabulaire	1
1.1.1 Épreuve statistique	2
1.1.2 Population	2
1.1.3 Individu (unité statistique)	3
1.1.4 Caractère (variable statistique)	4
1.1.5 Modalités	4
1.2 Types des caractères	5
1.2.1 Caractère qualitatif	5
1.2.2 Caractère quantitatif	6
1.3 Exercices corrigés	7
1.4 Exercices supplémentaires	8
2 Étude d'une variable statistique discrète	11
2.1 Effectif partiel - effectif cumulé	12
2.1.1 Effectif partiel (fréquence absolue)	12
2.1.2 Effectif cumulé	13
2.2 Fréquence partielle - Fréquence cumulée	13
2.2.1 Fréquence partielle (fréquence relative)	13
2.2.2 Fréquence cumulée	15
2.3 Représentation graphique des séries statistiques	16
2.3.1 Distribution à caractère qualitatif	16
2.3.2 Distribution à caractère quantitatif discret	18
2.3.3 Représentation sous forme de courbe et fonction de répartition	18
2.4 Paramètres de position	20
2.5 Paramètres de dispersion (variabilité)	22
2.6 Exercices corrigés	24
2.7 Exercices supplémentaires	29
3 Étude d'une variable statistique continue	33
3.1 Caractère continu	33

3.1.1	Classe de valeurs	34
3.1.2	Nombre de classes	34
3.1.3	Effectif et fréquence d'une classe	36
3.2	Représentation graphique d'un caractère continu	37
3.2.1	Histogramme des fréquences (ou effectifs)	37
3.2.2	Fonction de répartition	38
3.3	Paramètres de tendance centrale	39
3.4	Paramètres de dispersion	42
3.5	Exercices corrigés	43
3.6	Exercices supplémentaires	48
4	Étude d'une variable statistique à deux dimensions	53
4.1	Représentation des séries statistiques à deux variables	54
4.2	Description numérique	60
4.2.1	Caractéristique des séries marginales	60
4.2.2	Série conditionnelle	61
4.2.3	Notion de covariance	62
4.3	Ajustement linéaire	64
4.3.1	Coefficient de corrélation	64
4.3.2	Droite de régression	66
4.4	Exercices corrigés	68
4.5	Exercices supplémentaires	73
5	Annexe historique	81
Bibliographie		83

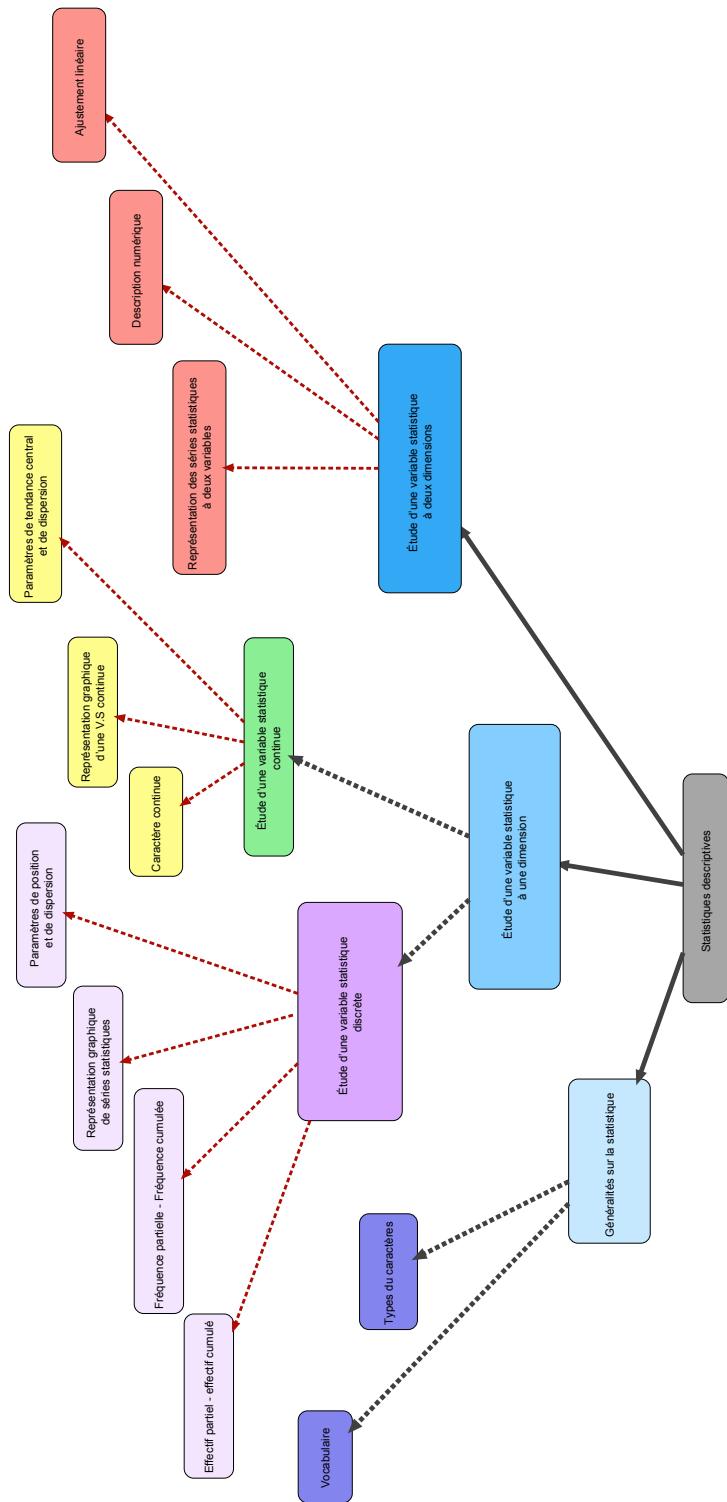


Table des figures

2.1	Le nombre d'individus (effectif)	12
2.2	Quelques caractéristiques du graphique	16
2.3	Tuyaux d'orgues	17
2.4	Diagramme par secteur	17
2.5	Diagramme à bâtons	18
2.6	Représentation d'une variable quantitative discrète par la courbe cumulative.	19
2.7	La dispersion d'une série statistique autour de sa moyenne	24
2.8	A gauche "Tuyaux d'orgue" et à droite "Diagramme en secteur"	25
2.9	Diagramme à bâtons	26
3.1	Une représentation de la distribution des valeurs à l'intérieur d'une classe.	35
3.2	Le nombre d'individus (effectif) - cas continu	36
3.3	Histogramme des fréquences ou des effectifs.	38
3.4	Le calcul de $F_x(x)$ par extrapolation.	39
3.5	La courbe des fréquences cumulées.	39
3.6	Le centre de la classe.	40
3.7	Représentation ou détermination graphique du mode (cas continu).	41
3.8	Le calcul de la médiane par extrapolation.	42
3.9	Les quartiles.	43
4.1	Représentation sous forme de nuage de points.	55
4.2	Le nombre d'individus (effectif)	56
4.3	La covariance et la variabilité.	62
4.4	Le coefficient de corrélation	65
4.5	Exemples de diagrammes de dispersion	65
4.6	La corrélation reflète la non-linéarité et la direction	66
4.7	La méthode des moindres carrés et la droite de régression	66
4.8	Acceptation ou refus de l'ajustement linaire	68

Symboles et Notations

Symbole	Signification
$[\]$	La partie entière.
$\text{Card}(\Omega)$	Le cardinal : nombre d'éléments de l'ensemble Ω .
$:=$	Est défini comme étant (symbole d'affectation).
\mathbb{N}	Ensemble des nombres entiers naturels.
\mathbb{Z}	Ensemble des nombres entiers relatifs.
\mathbb{R}	Ensemble des nombres réels.
\mathbb{R}^2	Ensemble des couples de nombres réels.
$\sum_{i=1}^n$	La somme pour i variant de 1 à n .
$V.S$	La variable statistique
Me	La médiane.
Me^+	Me par valeur supérieure.
Me^-	Me par valeur inférieure.
M_0	Le mode.
\bar{x}	La moyenne d'une série statistique X .
σ_X	L'écart-type de X .
$\text{Var}(X)$	La variance de X .
$\text{Cov}(X,Y)$	La covariance entre les variables X et Y .
ρ_{XY}	Le coefficient de corrélation entre les variables X et Y .
F_x	La fonction de répartition du caractère X

Chapitre 1

Généralités sur la statistique

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

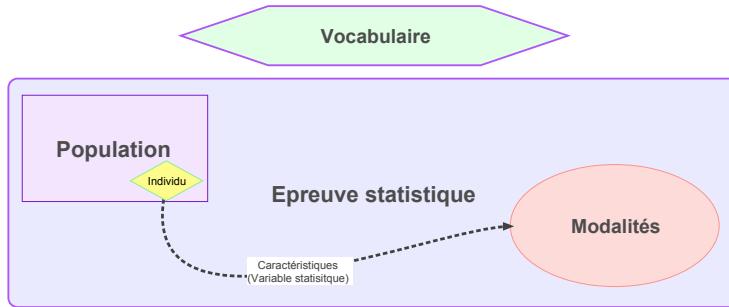
L'analyse des données est utilisée pour décrire les phénomènes étudiés, faire des prévisions et prendre des décisions à leur sujet. En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie et bien sûr les sciences de l'ingénieur. La statistique consiste à :

- Recueillir des données.
- Présenter et résumer ces données.
- Tirer des conclusions sur la population étudiée et d'aider à la prise de décision.
- En présence de données dépendant du temps, nous essayons de faire de la prévision.

1.1 Vocabulaire

Les statistiques consistent en diverses méthodes de classement des données tels que les tableaux, les histogrammes et les graphiques, permettant d'organiser un grand nombre de données. Les statistiques se sont développées dans la deuxième moitié du XIX^e siècle dans le domaine des sciences humaines (sociologie, économie, anthropologie, ...). Elles se sont dotées d'un vocabulaire particulier.



1.1.1 Épreuve statistique

Les statistiques descriptives visent à étudier les caractéristiques d'un ensemble d'observations comme les mesures obtenues lors d'une expérience. L'expérience est l'étape préliminaire à toute étude statistique. Il s'agit de prendre "contact" avec les observations. De manière générale, la méthode statistique est basée sur le concept suivant.

Définition 1

L'épreuve statistique est une expérience que l'on provoque.

Exemple 1 (La durée de vie des lampes)

Imaginons le cas suivant : un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament sur la durée de vie des ampoules fabriquées. Pour ce faire, il va faire fabriquer 4 échantillons d'ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus.

1.1.2 Population

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population. Mais, en statistique, le terme de population s'applique à tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques. Nous définissons la notion de population.

Définition 2

On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω .

Exemple 2

- *On considère l'ensemble des étudiants de la section A. On s'intéresse aux nombres de frères et sœurs de chaque étudiant. Dans ce cas*

$$\Omega = \text{ensemble des étudiants.}$$

- *Si l'on s'intéresse maintenant à la circulation automobile dans une ville, la population est alors constituée de l'ensemble des véhicules susceptibles de circuler dans cette ville à une date donnée. Dans ce cas*

$$\Omega = \text{ensemble des véhicules.}$$

1.1.3 Individu (unité statistique)

Une population est composée d'individus. Les individus qui composent une population statistique sont appelés unités statistiques.

Définition 3

On appelle individu tout élément de la population Ω , il est noté ω (ω dans Ω).

Remarque 1

L'ensemble Ω peut être un ensemble de personnes, de choses ou d'animaux...

L'unité statistique est un objet pour lequel nous sommes intéressés à recueillir de l'information.

Exemple 3

- *Dans l'exemple indiqué ci-dessus, un individu est tout étudiant de la section.*
- *Si on étudie la production annuelle d'une usine de boîtes de boisson en métal (canettes). La population est l'ensemble des boîtes produites durant l'année et*

une boîte constitue un individu.

1.1.4 Caractère (variable statistique)

La statistique « descriptive », comme son nom l'indique cherche à décrire une population donnée. Nous nous intéressons au caractéristique des unités qui peuvent prendre différentes valeurs.

Définition 4

On appelle caractère (ou variable statistique, dénotée V.S) toute application

$$X : \Omega \rightarrow C.$$

L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus)

Exemple 4

Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle,...

Remarque 2

Soit Ω un ensemble. On appelle et on note $\text{Card}(\Omega)$, le nombre d'éléments de Ω .

$$\text{Card}(\Omega) := \text{nombre d'éléments de } \Omega = N.$$

1.1.5 Modalités

Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci.

Exemple 5

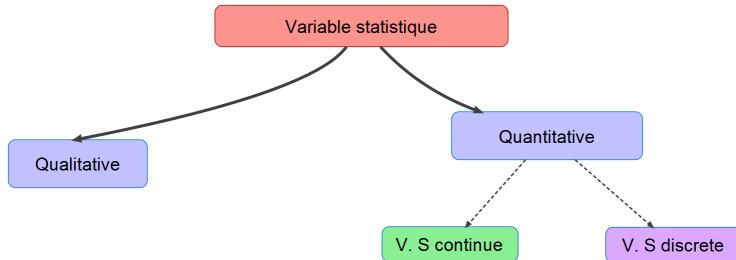
- Variable est " situation familiale "
- Modalités sont " célibataire, marié, divorcé "

- Variable est " statut d'interrupteur "
- Modalités sont " 0 et 1 ".
- Variable est " catégories socio-professionnelles "
- Modalités sont " Employés, ouvriers, retraités,... "

Les modalités sont les différentes situations dans lesquelles les individus peuvent se trouver à l'égard du caractère considéré.

1.2 Types des caractères

Nous distinguons deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.



1.2.1 Caractère qualitatif

Les caractères qualitatifs sont ceux dont les modalités ne peuvent pas être ordonnées, c'est-à-dire que si l'on considère deux caractères pris au hasard, on ne peut pas dire de l'un des caractères qu'il est inférieur ou égal à l'autre. Plus précisément, nous avons la définition suivante.

Définition 5

Les éléments de C sont représentés par autre chose que des chiffres.

Exemple 6

L'état d'une maison : on peut considérer les modalités suivantes

- Ancienne.
- Dégradée.
- Nouvelle.

- Rénovée.

1.2.2 Caractère quantitatif

Les caractères quantitatifs sont des caractères dont les modalités peuvent être ordonnées. Ainsi, l'âge, la taille de vie ou le salaire d'un individu sont des caractères quantitatifs. Donc, nous avons la définition suivante.

Définition 6

L'ensemble des valeurs est représenté par des chiffres. De même, il est partagé en deux sortes de caractères, discret et continu (voir l'exemple).

Exemple 7

- Le salaire d'employés d'une usine.

Modalités : 10000DA , 20000DA...

Type : Discret.

- La rigidité des ressorts.

Modalités : [10, 20] N/m

Type : continu.

En général, la variable quantitative discrète est une variable ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible. Citons, par exemple, le nombre de maisons par quartier d'une ville. Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises, mais des intervalles. C'est le cas lorsque nous avons un grand nombre d'observations distinctes.

La statistique descriptive a pour objectif de synthétiser l'information contenue dans les jeux de données au moyen de tableaux, figures ou résumés numériques. Les variables statistiques sont analysées différemment selon leur nature (quantitative, qualitative).

1.3 Exercices corrigés

Exercice 1

- La variable statistique "couleur de maisons d'un quartier" est-elle :

- qualitative quantitative
- discrète continue

La variable statistique "revenu brut" est-elle :

- qualitative quantitative
- discrète continue

La variable statistique "nombre de maisons vendues par ville" est-elle :

- qualitative quantitative
- discrète continue

Solution : Pour le premier cas, la variable statistique est qualitative. Pour le deuxième cas, la variable statistique est quantitative continue. Pour le troisième cas, la variable statistique est quantitative discrète.

Exercice 2

- Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses.

1. On appelle variable, une caractéristique que l'on étudie.
2. La tâche de la statistique descriptive est de recueillir des données.
3. La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.
4. En Statistique, on classe les variables selon différents types.
5. Les valeurs des variables sont aussi appelées modalités.
6. Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.
7. Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative.
8. La variable quantitative poids d'automobile peut être reclassée en compacte, intermédiaire et grosse.

9. En pratique, lorsqu'une variable quantitative discrète prend un grand nombre de valeurs distinctes, on la traite comme continue.

Solution : le corrigé en ordre est donné par

1. VRAI
2. FAUX
3. VRAI
4. VRAI
5. VRAI
6. VRAI
7. VRAI
8. VRAI
9. VRAI

Exercice 3

- Proposer des exemples de variable quantitative transformée en variable qualitative. Préciser les modalités de cette dernière.

Solution : Les variables quantitatives dans le tableau ci-dessous peuvent être transformées en variable qualitative. Les modalités de cette dernière sont précisées dans la seconde colonne.

Variable quantitative	Modalités envisageables
Hauteur	Petit, Moyen, Grand
Poids	Très léger, Léger, Moyen, Lourd, Très lourd
Rendement	Faible, Moyen, Elevé
Chiffre d'affaire	Modéré, Moyen, Important, Très important
Cylindrée	Petite, Moyenne, Grosse

1.4 Exercices supplémentaires

Exercice 4

- Pour chacune des variables suivantes, préciser si elle est qualitative, quantitative discrète ou quantitative continue,

- | | |
|---------------------------|-------------------------|
| <i>(a) Revenu annuel.</i> | <i>(b) Citoyenneté.</i> |
| <i>(c) Distance.</i> | <i>(d) Taille.</i> |

- (e) Lieu de résidence.
(g) Couleur des yeux.

- (f) Âge.
(h) Nombre de langues parlées.

Exercice 5

- Pour les sujets d'étude qui suivent, spécifier : l'unité statistique, la variable statistique et son type,

1. Étude du temps de validité des lampes électriques.
2. Étude de l'absentéisme des ouvriers, en jours, dans une usine.
3. Répartition des étudiants d'une promotion selon la mention obtenue sur le diplôme du Bac.
4. On cherche à modéliser¹ le nombre de collisions impliquant deux voitures sur un ensemble de 100 intersections routières choisies au hasard dans une ville. Les données sont collectées sur une période d'un an et le nombre d'accidents pour chaque intersection est ainsi mesuré.

Exercice 6

- Quelles formes de présentation de données correspondent ces propriétés ?
1. Il donne une bonne idée des données, mais on lui préfère en général les graphiques.
 2. Il n'est pas nécessaire de lire des nombres. D'un simple coup d'œil, on a une vision d'ensemble des données.

1. Fabriquer un modèle à partir des informations disponibles, généralement dans un but précis comme mieux appréhender un projet de construction ou se rendre compte de l'ampleur d'un sujet à partir d'une échelle réduite.

Chapitre 2

Étude d'une variable statistique discrète

Le caractère statistique (ou variable statistique, dénotée V.S) peut prendre un nombre fini raisonnable de valeurs (note, nombre d'enfants, nombre de pièces, ...). Dans ce cas, le caractère statistique étudié est alors appelé un caractère discret.

Dans toute la suite du chapitre, nous considérons la situation suivante :

$$X : \Omega \rightarrow \{x_1, x_2, \dots, x_n\},$$

avec $\text{Card}(\Omega) := N$ est le nombre d'individus dans notre étude.

Nous allons utiliser souvent l'exemple ci-dessous pour illustrer les énoncés de ce chapitre.

Exemple 8

Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille.

On note X le nombre d'enfants, les résultats sont données par ce tableau :

x_i	0	1	2	3	4	5	6
n_i (Effectif)	18	32	66	41	32	9	2

Nous avons

- Ω ensemble des familles.
- ω une famille.
- X nombre d'enfants par famille

$$X : \omega \rightarrow X(\omega).$$

On lit, à la famille ω , on associe $X(\omega) = \text{le nombre d'enfants de cette famille.}$

2.1 Effectif partiel - effectif cumulé

On étudie ici un caractère statistique numérique représenté par une suite x_i décrivant la valeur du caractère avec i varie de 1 à k .

2.1.1 Effectif partiel (fréquence absolue)

Définition 7

Pour chaque valeur x_i , on pose par définition

$$n_i = \text{Card}\{\omega \in \Omega : X(\omega) = x_i\}.$$

n_i : le nombre d'individus qui ont la même x_i , ça s'appelle effectif partiel de x_i .

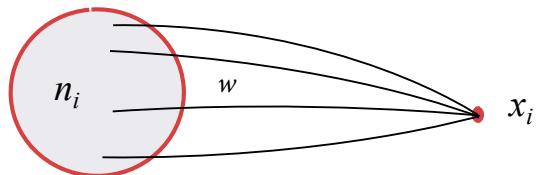


FIGURE 2.1: Le nombre d'individus qui prennent la valeur x_i .

Exemple 9

Dans l'exemple 8, on a 66 est le nombre de familles qui ont 2 enfants.

x_i	...	2	...
n_i (<i>Effectif</i>)	...	66	...

2.1.2 Effectif cumulé

Définition 8

Pour chaque valeur x_i , on pose par définition

$$N_i = n_1 + n_2 + \dots + n_i.$$

L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précédent.

Exemple 10

Dans l'exemple 8, on a 50 est le nombre de familles qui ont un nombre d'enfants inférieur à 1. Nous le regardons dans le tableau suivant :

x_i	0	1	2	3	4	5	6
N_i	18	50	116	157	189	198	200

Interprétation : N_i est le nombre d'individus dont la valeur du caractère est inférieur ou égale à x_i . De ce fait, l'effectif total est donné par

$$N = \text{card}\{\Omega\} = \sum_{i=1}^n n_i.$$

Dans notre exemple précédent, nous avons $N = 200$.

2.2 Fréquence partielle - Fréquence cumulée

Typiquement les effectifs n_i sont grands et il est intéressant de calculer des grandeurs permettant de résumer la série.

2.2.1 Fréquence partielle (fréquence relative)

Définition 9

Pour chaque valeur x_i , on pose par définition

$$f_i := \frac{n_i}{N}.$$

f_i s'appelle la fréquence partielle de x_i . La fréquence d'une valeur est le rapport de l'effectif de cette valeur par l'effectif total.

Remarque 3

On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage.

Interprétation : f_i = est le pourcentage des ω tel que $X(\omega) = x_i$.

Exemple 11

Dans l'exemple précédent, $0,33 :=$ il y a 33% de familles dont le nombre d'enfants égale à 2. Ce pourcentage est calculé de la façon suivante ($N = 200$) :

x_i	...	2	...
n_i (Effectif)	...	66	...
f_i (fréquence)	...	$\frac{66}{200} = 0.33$...

Nous pouvons conclure la propriété suivante.

Proposition 1

Soit f_i défini comme précédemment. Alors,

$$\sum_{i=1}^n f_i = 1.$$

Démonstration. Rappelons que

$$\sum_{i=1}^n n_i = N.$$

Ce qui implique que

$$\sum_{i=1}^n f_i = \sum_{i=1}^n \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^n n_i = 1.$$

2.2.2 Fréquence cumulée

Définition 10

Pour chaque valeur x_i , on pose par définition

$$F_i = f_1 + f_2 + \dots + f_i.$$

La quantité F_i s'appelle la fréquence cumulée de x_i .

Interprétation : F_i est le pourcentage des ω tel que la valeur $X(\omega)$ est inférieure ou égale à x_i .

Exemple 12

- Dans l'exemple précédent, 0.785 représente 78.5% de familles dont le nombre d'enfants est inférieur ou égale à 3.
- Dans un deuxième exemple, nous nous intéressons aux nombres d'erreurs d'assemblage sur un ensemble d'appareils,

Nombre d'erreurs	Nombre d'appareils	Fréquences cumulées
0	101	0.26
1	140	0.61
2	92	0.84
3	42	0.94
4	18	0.99
5	3	1

Nous avons 94% des appareils qui ont un nombre d'erreurs d'assemblage inférieur ou égale à 3.

Nous avons vu que les tableaux sont un moyen souvent indispensable, en tous cas très utile, de classification et de présentation des unités d'une population statistique. Dans le paragraphe suivant, nous allons voir comment on traduit ses tableaux en graphique permettant aussi de résumer d'une manière visuelle les données.

2.3 Représentation graphique des séries statistiques

On distingue les méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe).

Le graphique est un support visuel qui permet :

La synthèse : visualiser d'un seul coup d'œil les principales caractéristiques (mais on perd une quantité d'informations), voir Figure 2.2.

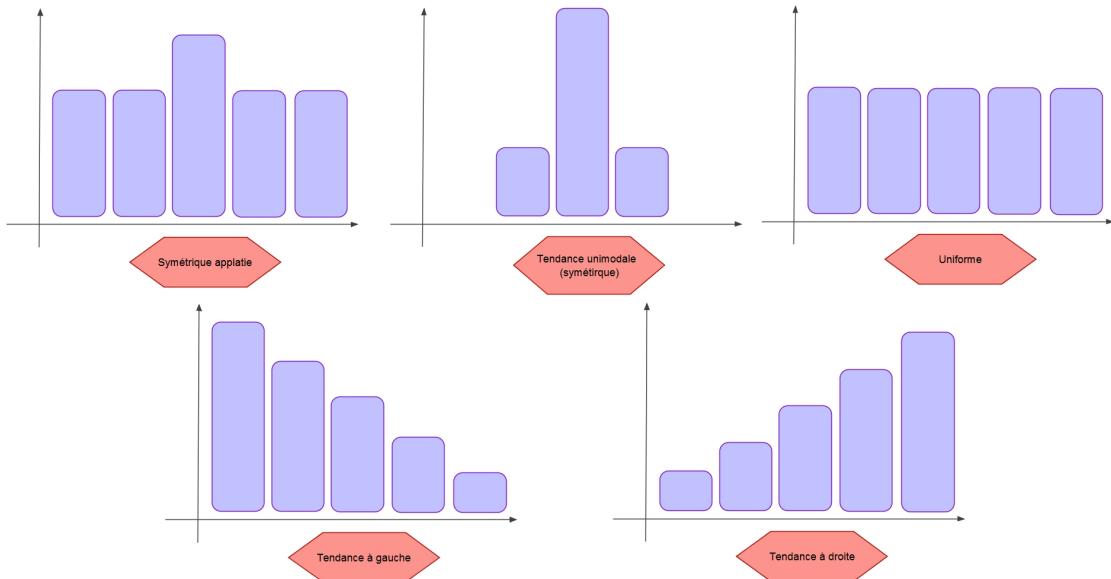


FIGURE 2.2: Quelques caractéristiques du graphique

La découverte : met en évidence les tendances.

Le contrôle : on aperçoit mieux les anomalies sur un graphique que dans un tableau.

La recherche des régularités : régularité dans le mouvement, répétition du phénomène.

2.3.1 Distribution à caractère qualitatif

A partir de l'observation d'une variable qualitative, deux diagrammes permettent de représenter cette variable : le diagramme en bandes (dit tuyaux d'orgue) et le diagramme à secteurs angulaires (dit camembert).

Tuyaux d'orgues

Nous portons en abscisses les modalités, de façon arbitraire. Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité (voir Figure 2.3).

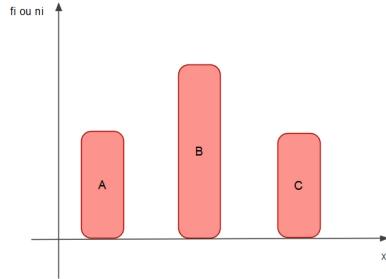


FIGURE 2.3: Tuyaux d'orgues

Diagramme par secteur (diagramme circulaire)

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité (voir Figure 2.4).

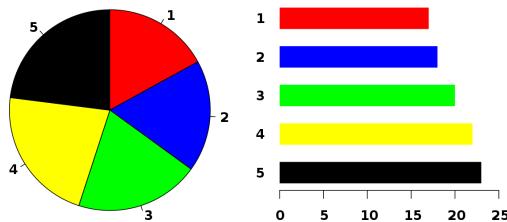


FIGURE 2.4: Diagramme par secteur

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

$$N \longrightarrow 360^\circ$$

$$n_i \longrightarrow d_i \quad (\text{degré de la modalité } i).$$

Donc,

$$d_i = \frac{n_i \times 360}{N}.$$

2.3.2 Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir ci-dessous).

Pour l'illustration, nous prenons l'exemple précédent de départ (nombre d'enfants par famille). Nous rappelons le tableau statistique associé.

x_i	0	1	2	3	4	5	6
n_i	18	32	66	41	32	9	2

Diagramme à bâtons

On veut représenter cette répartition sous la forme d'un diagramme en bâtons. À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés (voir Figure 2.5).

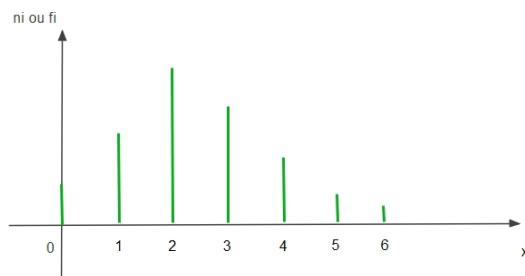


FIGURE 2.5: Diagramme à bâtons

2.3.3 Représentation sous forme de courbe et fonction de répartition

Nous avons déjà abordé les distributions cumulées d'une variable statistique. Nous allons dans cette partie exploiter ses valeurs cumulées pour introduire la notion de la fonction de répartition. Cette notion ne concerne que les variables quantitatives.

Soit la fonction $F_x : \mathbb{R} \rightarrow [0, 1]$ définie par

$F_x(x) :=$ pourcentage des individus dont la valeur du caractère est $\leq x$.

Cette fonction s'appelle la fonction de répartition du caractère X .

Remarque 4

Pour tout $i \in \{1, \dots, n\}$, on a

$$F_x(x_i) = F_i.$$

La courbe de F_x passe par les points $(x_1, F_1), (x_2, F_2), \dots$ et (x_n, F_n) .

En se basant sur notre exemple, la courbe de F_x est représentée ci-dessous (Figure 2.6) sur

$$\mathbb{R} =]-\infty, 0[\cup [0, 1[\cup \dots \cup [6, +\infty[.$$

Dans ce cas, nous avons

- Si $x < 0$, alors $F_x(x) = 0$.
- Si $x \in [0, 1[$, alors $F_x(x) = 0.09$.
- ...
- Si $x \geq 6$, alors $F_x(x) = 1$.

Cette courbe s'appelle "la courbe cumulative des fréquences". La courbe cumulative est une courbe en escalier représentant les fréquences cumulées relatives.

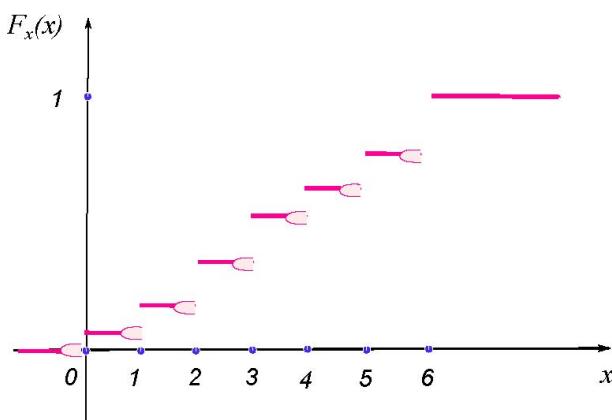


FIGURE 2.6: Représentation d'une variable quantitative discrète par la courbe cumulative.

Proposition 2

La fonction de répartition satisfait, pour $i \in \{1, \dots, n\}$,

- l'égalité, $F_x(x_i) = F_i$,

$$\text{– l'expression, } F_x(x) = \begin{cases} 0, & \text{si } x < x_1, \\ F_1, & \text{si } x_1 \leq x < x_2, \\ F_i, & \text{si } x_i \leq x < x_{i+1}, \\ 1, & \text{si } x \geq x_n. \end{cases} .$$

2.4 Paramètres de position (caractéristique de tendance centrale)

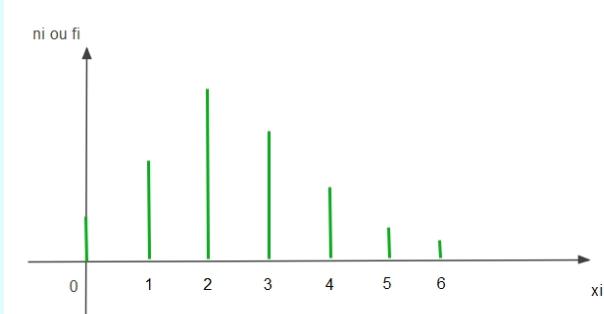
Les indicateurs statistiques de tendance centrale (dits aussi de position) considérés fréquemment sont la moyenne, la médiane et le mode.

Le mode

Le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_o .

Exemple 13

Dans l'exemple 8, le mode est égal à 2 qui correspondant au plus grand effectif.



Remarque 5

On peut avoir plus d'un mode ou rien.

La médiane

On appelle médiane la valeur Me de la V.S X qui vérifie la relation suivante :

$$F_x(Me^-) < 0.5 \leq F_x(Me^+) = F_x(Me).$$

La médiane partage la série statistique en deux groupes de même effectif.

Exemple 14

Dans l'exemple 8, la relation

$$F_x(0) = 0 < 0.5 \leq F_x(0^+) = 0.09$$

n'est pas satisfaite. Donc, la médiane est différente de 0. Par contre, nous avons

$$F_x(2^-) = 0.25 < 0.5 \leq F_x(2^+) = F(2) = 0.58.$$

Donc, $Me = 2$.

La moyenne

On appelle moyenne de X , la quantité

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n n_i x_i = \sum_{i=1}^n f_i x_i,$$

avec $N = \text{Card}(\Omega)$. On peut donc exprimer et calculer la moyenne dite "arithmétique" avec des effectifs ou avec des fréquences.

Exemple 15

Si $\bar{x} = 2.46$, alors nous avons au moyenne une famille de quartier a 2.46 d'enfants.

La valeur de la moyenne est abstraite. Comme dans l'exemple précédent, $\bar{x} = 2.46$ est un chiffre qui ne correspond pas à un fait concret.

La moyenne arithmétique dont on vient d'indiquer la formule est dite moyenne pondérée ; cela signifie que chaque valeur de la variable est multipliée (pondérée) par un coefficient, ici par l'effectif n_i qui lui correspond. Dans ce cas, chaque valeur x_i de la variable intervient dans le calcul de la moyenne autant de fois qu'elle a été observée. On parle de moyenne arithmétique simple quand on n'effectue pas de pondération. Par exemple, si 5 étudiants ont pour âge respectif 18, 19, 20, 21 et 22 ans, leur âge moyen est donné par $(18 + 19 + 20 + 21 + 22)/5 = 20$ ans.

Remarque 6

Nous mentionnons qu'il existe d'autres moyennes que la moyenne arithmétique

2.5 Paramètres de dispersion (variabilité)

Les indicateurs statistiques de dispersion usuels sont l'étendue, la variance et l'écart-type.

L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$e = x_{\max} - x_{\min},$$

s'appelle l'étendue de la V.S X . Le calcul de l'étendue est très simple. Il donne une première idée de la dispersion des observations. C'est un indicateur très rudimentaire et il existe des indicateurs de dispersion plus élaborés (voir ci-dessous).

La variance

On appelle variance de cette série statistique X , le nombre

$$Var(X) = \sum_{i=1}^n f_i(\bar{x} - x_i)^2$$

On dit que la variance est la moyenne des carrés des écarts à la moyenne \bar{x} . Les « écarts à la moyenne » sont les $(\bar{x} - x_i)$, les « carrés des écarts à la moyenne » sont donc les $(\bar{x} - x_i)^2$. En faisant la moyenne de ces écarts, on trouve la variance.

Le théorème suivant (Théorème de König-Huygens) donne une identité remarquable reliant la variance et la moyenne, parfois plus pratique dans le calcul de la variance.

Théorème 1

Soit (x_i, n_i) une série statistique de moyenne \bar{x} et de variance $Var(X)$. Alors,

$$Var(X) = \sum_{i=1}^n f_i x_i^2 - \bar{x}^2.$$

Démonstration. Par définition, nous avons

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

Donc,

$$Var(X) = \frac{\sum_{i=1}^n n_i (\bar{x} - x_i)^2}{\sum_{i=1}^n n_i} = \frac{\sum_{i=1}^n n_i (\bar{x}^2 + x_i^2 - 2\bar{x}x_i)}{\sum_{i=1}^n n_i}.$$

Par égalité, nous avons

$$Var(X) = \frac{\sum_{i=1}^n n_i \bar{x}^2}{\sum_{i=1}^n n_i} + \frac{\sum_{i=1}^n n_i x_i^2}{\sum_{i=1}^n n_i} - \frac{\sum_{i=1}^n 2n_i \bar{x}x_i}{\sum_{i=1}^n n_i}.$$

Ce qui implique que

$$Var(X) = \bar{x}^2 + \frac{\sum_{i=1}^n n_i x_i^2}{\sum_{i=1}^n n_i} - 2\bar{x}\bar{x} = -\bar{x}^2 + \frac{1}{N} \sum_{i=1}^n n_i x_i^2.$$

Remarque 7

Dans l'utilisation de la formule du théorème précédent, il faut veiller à remplacer \bar{x} par sa valeur approchée la plus précise possible.

L'écart type

La quantité

$$\sigma_X = \sqrt{Var(x)}$$

s'appelle l'écart type de la V.S X .

Remarque 8

Le paramètre σ_x mesure la distance moyenne entre \bar{x} et les valeurs de X (voir Figure 2.7). Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.

- Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).
- Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).

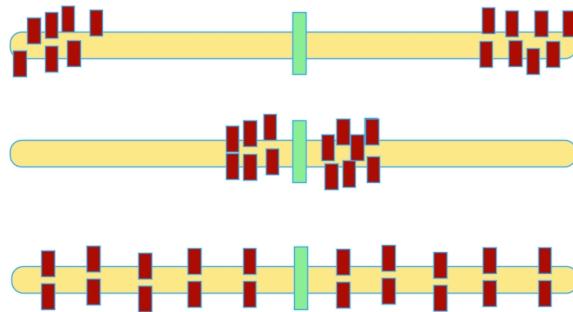


FIGURE 2.7: La dispersion d'une série statistique autour de sa moyenne

2.6 Exercices corrigés

Exercice 7

- Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population,

Groupes sanguins	A	B	AB	O
L'effectif	20	10	n_3	5

1. Déterminer la variable statistique et son type.
2. Déterminer l'effectif des personnes ayant un groupe sanguin AB.
3. Donner toutes les représentations graphiques possibles de cette distribution.

Solution 1 - La population dans cette étude est les 40 personnes. Donc $N = 40$. La variable statistique est le groupe sanguin des individus et elle est qualitative.

2 - L'effectif total est égal à 40. Par conséquent,

$$N = 40 = \sum_{i=1}^4 n_i.$$

Alors,

$$20 + 10 + n_3 + 5 = 40.$$

Ce qui implique que $n_3 = 5$.

3- Nous avons deux représentations possibles "Tuyaux d'orgue" et "Diagramme en secteur".

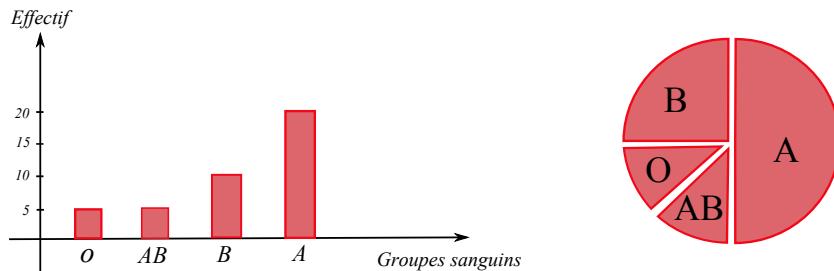


FIGURE 2.8: A gauche "Tuyaux d'orgue" et à droite "Diagramme en secteur"

Les angles dans la figure 2.8 se calcule en utilisant la règle de trois.

Exercice 8

- Le gérant d'un magasin vendant des articles de consommation courante a relevé pour un article particulier qui semble connaître une très forte popularité, le nombre d'articles vendus par jour. Son relevé a porté sur les ventes des mois de Mars et Avril, ce qui correspond à 52 jours de vente. Le relevé des observations se présente comme suit :

7 13 8 10 9 12 10 8 9 10 6 14 7 15 9 11 12 11 12 11 12 5 14 11 8 10 14 12 8
5 7 13 12 16 11 9 11 11 12 12 15 14 5 14 9 9 14 13 11 10 11 12 9 15.

1. Quel type est la variable statistique étudiée.
2. Déterminer le tableau statistique en fonction des effectifs, des fréquences, des effectifs cumulés et des fréquences cumulées.
3. Tracer le diagramme des bâtonnés associé à la variable X .
4. Soit F_x la fonction de répartition. Déterminer F_x .
5. Calculer le mode M_o et la moyenne arithmétique \bar{x} .

6. Déterminer à partir du tableau puis à partir du graphe, la valeur de la médiane M_e .

7. Calculer la variance et l'écart-type.

Solution 1 - La population est les 52 jours et la variable statistique étudiée est le nombre d'articles vendus par jour. Son type est bien évidemment quantitatif discret (nombre).

2 - Le tableau statistique est donné par

x_i	5	6	7	8	9	10	11	12	13
n_i	3	1	3	4	7	5	8	8	3
f_i	3/52	1/52	3/52	4/52	7/52	5/52	8/52	8/52	3/52
N_i	3	4	7	11	18	23	31	39	42
F_i	3/52	4/52	7/52	11/52	18/52	23/52	31/52	39/52	42/52

14	15	16	Σ
6	3	1	$N = 52$
6 /52	3/52	1/52	1
48	51	52	\emptyset
48/52	51/52	1	\emptyset

3 - L'élaboration du diagramme des bâtonnets de X ,

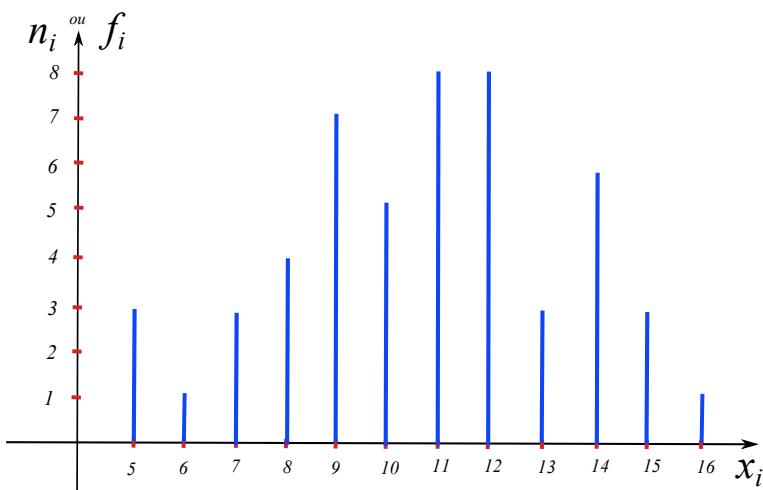


FIGURE 2.9: Diagramme à bâtons

4 - La fonction de répartition est donnée par

$$F_x(x) = \begin{cases} 0, & \text{si } x < 5, \\ 3/52, & \text{si } 5 \leq x < 6, \\ 4/52, & \text{si } 6 \leq x < 7, \\ 7/52, & \text{si } 7 \leq x < 8, \\ \ddots & \ddots \\ 1, & \text{si } x \geq 16. \end{cases}$$

5 - Le mode est la valeur de la variable qui a le plus grand effectif, c'est à dire, $n_i = 8$. Donc,

$$M_o = 11 \quad \text{et} \quad M_o = 12.$$

La moyenne arithmétique est donnée par ;

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{12} n_i x_i = \sum_{i=1}^{12} f_i x_i.$$

Par conséquent,

$$\bar{x} = \frac{1}{52} (3 \times 5 + 1 \times 6 + 3 \times 7 + \dots + 1 \times 16) = \frac{555}{52} = 10.67.$$

6 - La médiane est la valeur de la variable qui divise la population de la série statistique en deux parties égales. Nous avons,

$$F_x(11^-) = \frac{23}{52} < 0.5 \leq F_x(11^+) = F(Me) = \frac{31}{52}.$$

Donc, $Me = 11$.

7 - Nous commençons par la variance,

$$Var(X) = \frac{1}{N} \sum_{i=1}^n n_i x_i^2 - \bar{x}^2.$$

Après calcule, on trouve

$$Var(X) = 7.64.$$

Par conséquent, l'écart type est calculé à partir de

$$\sigma_X = \sqrt{Var(x)} = 2.76.$$

Exercice 9

- On considère deux groupes d'étudiants. Nous relevons leurs notes d'examens dans les deux tableaux suivants :

Note (groupe A)	8	9	10	11
Effectif	2	2	1	1

Note (groupe B)	6	8	9	13	14
Effectif	2	2	2	1	1

Calculer la moyenne et l'écart type de chaque groupe. Comparer les deux groupes.

Solution Dans un premier temps, nous remarquons que l'effectif total du groupe A est égal à 6 et celui du groupe B est égal à 8.

En utilisant la formule de la moyenne, nous obtenons

$$\bar{x}_A = 9.2 \quad \text{et} \quad \bar{x}_B = 9.1.$$

On remarque que les moyennes sont très proches. Peut-on pour autant conclure que ces deux groupes ont des niveaux identiques ?

Nous répondons à cette question après le calcul des écarts type. Ils sont donnés par

$$\sigma_X^A = 1.11 \quad \text{et} \quad \sigma_X^B = 2.8.$$

Nous remarquons que même si les deux groupes ont des moyennes quasiment identiques, le groupe B est beaucoup plus dispersé que le groupe A car $\sigma_X^B > \sigma_X^A$. Les étudiants de ce groupe ont des notes plus irrégulières. On peut dire donc que le groupe B est moins homogènes que le groupe A. En observant les valeurs du tableau, on voit que c'est cohérent.

Exercice 10

- Un quartier résidentiel comprend 99 unités d'habitation ayant une valeur locative moyenne de 10000 Da. Deux nouvelles unités d'habitation sont construites dans le quartier : l'une a une valeur locative de 7000 Da et l'autre, une villa luxueuse, a une valeur locative de 114000 Da.

- Quelle est la nouvelle moyenne de valeur locative pour le quartier ?
- Pouvait-on s'attendre à de tel résultat ?

Solution - Le nouveau total des mesures de valeur locative est

$$(99 \times 10000) + 7000 + 114000 = 1111000.$$

Le nouveau total d'individus statistiques est $99 + 2 = 101$. La nouvelle moyenne est donc

$$\frac{1111000}{101} = 110000.$$

- On pouvait s'attendre à une augmentation de la moyenne car l'une des deux nouvelles valeurs est très nettement au dessus de la moyenne initiale.

2.7 Exercices supplémentaires

Exercice 11

- Pour déterminer le type de logement (F_2, F_3, \dots) à construire, on étudie 20 familles selon leur nombre d'enfants. Durant l'expérience, on note les résultats suivants :

1, 3, 5, 5, 3, 2, 4, 4, 7, 0, 2, 4, 3, 7, 0, 5, 4, 2, 3, 2

- Déterminer, la population, l'unité (individu), la variable statistique et les modalités.
- Déterminer le tableau statistique avec x_i , n_i , f_i et F_i .

Modalité	0	1	2	3	4	5	6	7	\sum
n_i									
N_i									
f_i									
F_i									

Exercice 12

- Voici le tableau des pourcentages obtenu pour la variable "Mode de logement" :

x_i	"Cité U"	"Studio"	"Résidence"	"Maison"	"Autre"	TOTAL
%	4.8	16.5	38.6	28.6	11.6	100

Sachant que la taille de l'échantillon $N = 189$, retrouver les effectifs pour chaque modalité.

Exercice 13

- Au poste de péage, on compte le nombre de voitures se présentant sur une période de 5 min. Sur 100 observations de 5 min, on obtient les résultats suivants :

<i>Nombre de voitures</i>	1	2	3	4	5	6	7	8	9	10	11	12
<i>Nombre d'observations</i>	2	8	14	20	19	15	9	6	2	3	1	1

1. Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.
2. Calculer la moyenne et l'écart-type de cette série.
3. Déterminer la médiane.

Exercice 14

- Dans une petite localité, on a relevé le nombre de pièces par appartement :

<i>Nombre de pièces</i>	1	2	3	4	5	6	7
<i>Nombre d'appartements</i>	48	72	96	64	39	25	3

Le « nombre de pièces par appartement » est à considérer comme une variable aléatoire discrète à valeurs entières. (A l'interprétation, il faudra préciser que les « demi pièces » ne sont pas comptabilisées).

- Déterminer le tableau statistique.
- Tracer le diagramme des bâtonnés et la courbe des fréquences cumulées associés à la variable statistique.
- Calculer la moyenne et l'écart-type de cette série.
- Déterminer la médiane.

Exercice 15

- Une machine coupe des barres de 12 cm. Mais malheureusement, elle n'est pas bien réglée et les longueurs varient autour de la valeur attendue. Une étude sur 185 barres donnent les résultats suivants :

<i>Longueurs en cm</i>	11.5	11.6	11.7	11.8	11.9	12	12.1	12.2	12.3
<i>Effectif</i>	3	15	16	16	18	20	25	25	28

1. Quel type est la variable statistique étudiée.
2. Déterminer le tableau statistique.
3. Tracer le diagramme des bâtonnés et la courbe des fréquences cumulées associés à la variable statistique.

4. Calculer la moyenne et l'écart type.

Exercice 16

- La répartition en 2016 du nombre de pièces des résidences principales en Algérie est

Nombre de pièces	1	2	3	4	5	6 et plus
Effectif ($\times 10^3$)	1200	2700	4700	5500	3500	2500

- Donner l'individu de la population statistique étudiée.
- Quelle est la nature de la variable étudiée ?
- Quel outil graphique peut-on utiliser pour représenter cette variable ?

Exercice 17

- On observe le nombre d'arrivées des clients à un bureau de poste pendant un intervalle de temps donné (disant 10 minutes). En répétant 100 fois cette observation, on obtient les résultats suivants :

Nombre d'arrivées	1	2	3	4	5	6	Total
Nombre d'observations	15	25	26	20	7	7	100

- Représenter graphiquement ces résultats (pour les effectifs et pour les fréquences cumulées).
- Calculer la valeur de la moyenne arithmétique, de la médiane, de la variance et de l'écart type des résultats.

Exercice 18

- On mesure les diamètres de troncs d'arbres d'une même espèce. On étudie 400 échantillons. On obtient les résultats suivants :

Diamètre en cm	25	26	27	28	29	30
Pourcentage	10%	15%	30%	35%	5%	5%

On donne :

$$(25 \times 0.1) + (26 \times 0.15) + (27 \times 0.3) + (28 \times 0.35) + (29 \times 0.05) + (30 \times 0.05) = 27.25.$$

$$(25^2 \times 0.1) + (26^2 \times 0.15) + (27^2 \times 0.3) + (28^2 \times 0.35) + (29^2 \times 0.05) + (30^2 \times 0.05) = 744.05.$$

1. Établir le tableau statistique en fonction des effectifs et des fréquences relatives.
2. Quel est le diamètre moyen de ces troncs d'arbres ?

3. Déterminer la variance puis l'écart-type de la série statistique résumée dans le tableau ci-dessus.
4. Représenter graphiquement ces résultats (juste pour les effectifs).
5. Déterminer le mode et donner son interprétation.

Chapitre 3

Étude d'une variable statistique continue

Nous rappelons qu'une variable statistique (V.S) quantitative concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité et sur lesquelles les opérations arithmétiques (addition, multiplication, etc,...) ont un sens. Nous allons dans ce chapitre se focaliser sur la V.S quantitative continue.

3.1 Caractère continu

Définition 11

On appelle V.S continue (ou caractère continu) toute application de Ω et à valeurs réelles et qui prend un nombre "important" de valeurs (Les caractères continus sont ceux qui ont une infinité de modalités).

Exemple 16

Soit Ω l'ensemble des nouveaux nés au C.H.U d'une ville pendant les 3 premiers mois de 2017. Nous désignons par X le poids des nouveaux nés. On suppose que

$$x_{\min} = 2.701 \quad \text{et} \quad x_{\max} = 5.001.$$

Remarque 9

Comment étudier ce caractère ?

Réponse : Partager les valeurs prises par X en classes de valeurs.

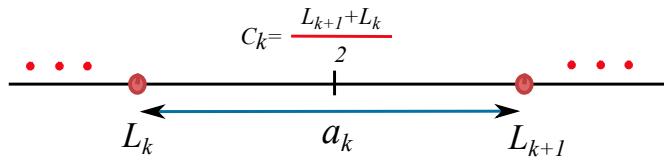
3.1.1 Classe de valeurs

Définition 12

On appelle classe de valeurs de X un intervalle de type $[a, b[$ tel que $X \in [a, b[$ si et seulement si $a \leq X(w) < b$, c'est à dire, que les valeurs du caractère sont dans la classe $[a, b[$.

Dès qu'un caractère est identifié en tant que continu, ces modalités $C_k = [L_k, L_{k+1}[$ sont des intervalles avec

- L_k : borne inférieure.
- L_{k+1} : borne supérieure.
- $a_k = L_{k+1} - L_k$: son amplitude, son pas ou sa longueur.
- $C_k = x_k = (L_{k+1} + L_k)/2$: son centre (nous la notons encore C_k).



Remarque 10

On supposera dans tous les cas étudiés que la distribution à l'intérieur des classes est uniforme (voir Figure 3.1). Cette hypothèse permet de justifier le fait qu'on choisisse le centre des classes comme représentant.

3.1.2 Nombre de classes

En combien de classes partageons-nous les valeurs ? la réponse n'est pas unique. Soit N l'effectif total. Nous pouvons considérer dans ce cours trois réponses à titre d'exemple.

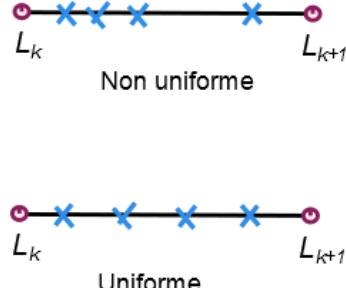


FIGURE 3.1: Une représentation de la distribution des valeurs à l'intérieur d'une classe.

1. Une réponse : \sqrt{N} , $[\sqrt{N}]$ (partie entière) ou $[\sqrt{N}] + 1$. Donc, le nombre de classes

$$k \simeq \sqrt{N}.$$

Exemple 17

Considérons 30 valeurs entre 56.5 cm et 97.8 cm. Dans ce cas, $k = \sqrt{30}$ et on prend $k = 6$.

2. Une réponse : la formule de Sturge

$$k = 1 + 3.3 \log_{10}(N).$$

3. Une réponse : la formule de Yule

$$k = 2.5 \sqrt[4]{N}.$$

Remarque 11

De ce fait, on peut avoir plusieurs tableaux statistiques selon le nombre de classes.

Exemple 18

Si on prend $N = 30$, alors le nombre de classes est donné, par exemple, par

- soit la formule de Sturge $k = 1 + 3.3 \log_{10}(30) \simeq 6$,

- soit la formule de Yule $k = 2.5 \sqrt[4]{30} \simeq 6$.

Nous mentionnons que les deux formules sont presque pareils si $N \ll 200$.

Nous rappelons maintenant la définition de l'étendu. De plus, dans le cas continu nous parlons aussi du pas ou de la longueur de la classe.

Définition 13

Le nombre

$$e = x_{\max} - x_{\min}$$

s'appelle étendu de X . Dans ce cas, on peut définir le pas par

$$a_i := \frac{\text{étendu}}{\text{nombre de classes}} = \frac{x_{\max} - x_{\min}}{k}.$$

3.1.3 Effectif et fréquence d'une classe

Définition 14

La quantité

$$n_i := \text{Card}\{w \in \Omega : X(w) \in C_i\}$$

s'appelle effectif partiel de C_i .

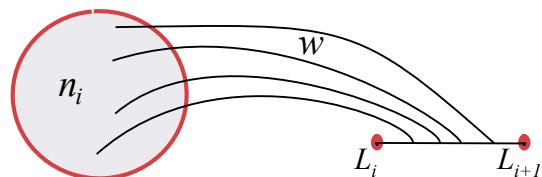


FIGURE 3.2: Le nombre d'individus qui prennent des valeurs x_i dans C_i .

Définition 15

Le nombre

$$f_i := \frac{n_i}{N}$$

est appelé la fréquence partielle de C_i .

Définition 16

On appelle l'*effectif cumulé* de C_i la quantité

$$N_i := \sum_{j=1}^i n_j.$$

Définition 17

On appelle la *fréquence cumulée* de C_i la quantité

$$F_i := \sum_{j=1}^i f_j.$$

Remarque 12

Nous avons, comme dans le chapitre précédent, les interprétations suivantes :

- n_i : est le nombre d'individus dont les valeurs des caractères sont dans la classe C_i ,
- f_i : est le pourcentage des w tel que $X(w) \in C_i$,
- N_i : est égale au $\text{Card}\{w : X(w) \in C_1 \cup C_2 \cup \dots \cup C_i\}$,
- F_i : est le pourcentage des w tel que

$$X(w) \in C_1 \cup \dots \cup C_i.$$

3.2 Représentation graphique d'un caractère continu

3.2.1 Histogramme des fréquences (ou effectifs)

Nous pouvons représenter le tableau statistique par un histogramme. Nous reportons les classes sur l'axe des abscisses et, au-dessus de chacune d'elles, nous traçons un rectangle dont l'aire est proportionnelle à la fréquence f_i (ou l'effectif n_i) associée. Ce graphique est appelé l'*histogramme des fréquences* (voir Figure 3.3).

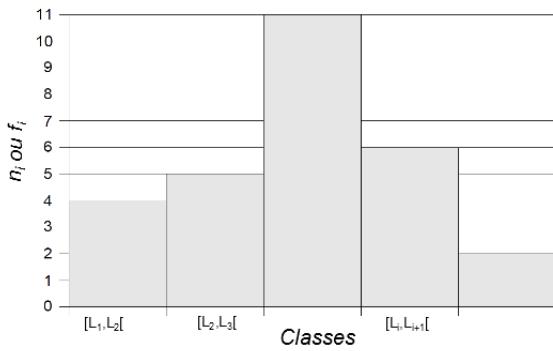


FIGURE 3.3: Histogramme des fréquences ou des effectifs.

3.2.2 Fonction de répartition

Notation : Nous allons noter par

$$C_i = [x_{\min} = a_0, x_{\min+1} = a_1[.$$

Définition 18

La fonction $F_x : \mathbb{R} \rightarrow [0, 1]$ définie par $F_x(x)$ représente le pourcentage des individus tel que la valeur de leur caractère est inférieure ou égale à x . Elle est donnée par

$$F_x(x) = \begin{cases} 0, & \text{si } x < a_0, \\ \frac{f_1}{h}(x - a_0), & \text{si } a_0 \leq x < a_1, \\ F_i + \frac{f_{i+1}}{h}(x - a_i), & \text{si } a_i \leq x < a_{i+1}, \\ 1, & \text{si } x \geq a_n, \end{cases}$$

et elle s'appelle la fonction de répartition de X .

Nous expliquons cette formulation de la fonction de répartition dans cette remarque.

Remarque 13

Nous calculons $F_x(x)$ par extrapolation (voir Figure 3.4). Nous avons déjà $F(L_i) = F_i$.

De plus,

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{F(x) - F(L_i)}{x - L_i}.$$

Ce qui implique la formule de la fonction de répartition

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

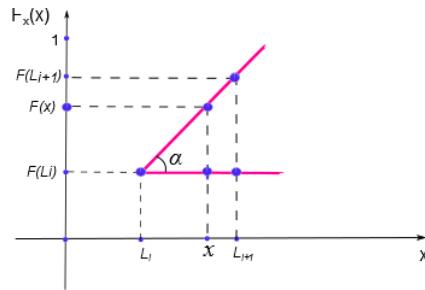


FIGURE 3.4: Le calcul de $F_x(x)$ par extrapolation.

La courbe de F_x est nulle avant a_0 , constante égale à 1 après a_n et joint les points $(a_0, 0)$, $(a_1, F_1), \dots, (a_n, 1)$ par des segments de droites (voir Figure 3.5).

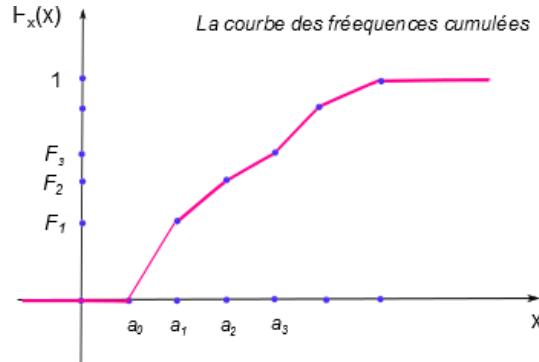


FIGURE 3.5: La courbe des fréquences cumulées.

3.3 Paramètres de tendance centrale

On note par C_i le centre de la classe C_i et nous considérons f_i la fréquence partielle de C_i .



FIGURE 3.6: Le centre de la classe.

La moyenne

Définition 19

La quantité

$$\bar{x} = \sum_{i=1}^n f_i C_i$$

s'appelle la moyenne de X.

Le mode

La définition suivante permet de comprendre la démarche à suivre pour calculer le mode d'une manière exacte et qui se trouve dans une des classes appelée "classe modale".

Définition 20

Nous définissons la classe modale comme étant la classe des valeurs de X qui a le plus grand effectif partiel (ou la plus grande fréquence partielle). La quantité

$$M_o = L_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_i$$

s'appelle le mode avec (voir Figure 3.7)

- L_i : la borne inférieure de la classe modale.
- a_i : le pas de la classe modale.
- $\Delta_1 = n_0 - n_1$, $\Delta_2 = n_0 - n_2$ ou bien $\Delta_1 = f_0 - f_1$, $\Delta_2 = f_0 - f_2$.
- n_0 et f_0 sont l'effectif et la fréquence associés à la classe modale.
- n_1 et f_1 sont l'effectif et la fréquence de la classe qui précède la classe modale.
- n_2 et f_2 sont l'effectif et la fréquence de la classe qui suit la classe modale.

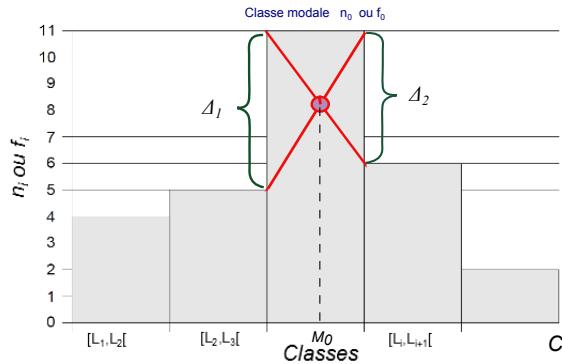


FIGURE 3.7: Représentation ou détermination graphique du mode (cas continu).

Remarque 14

L'expression du mode donnée ci-dessus est déterminée à partir de l'intersection des deux segments représentés dans la Figure 3.7. Cette notion n'est pas unique.

La médiane

Définition 21

La médiane est la valeur Me telle que $F(Me) = 0.5$. Cette valeur est unique.

Nous pouvons la déterminer graphiquement ou par calcul.

1. **Première méthode :** Graphiquement à partir de la formule

$$\tan(\alpha) = \frac{F(L_{i+1}) - F(L_i)}{L_{i+1} - L_i} = \frac{0.5 - F(L_i)}{Me - L_i}.$$

Plus précisément, dans la figure 3.8, nous mettons $F(x) = 0.5$ et $x = Me$.

2. **Deuxième méthode :** En utilisant directement la fonction de répartition donnée par

$$F(x) = \frac{f_{i+1}}{h}(x - L_i) + F_i.$$

Nous retrouvons donc

$$0.5 = \frac{f_{i+1}}{h}(Me - L_i) + F_i.$$

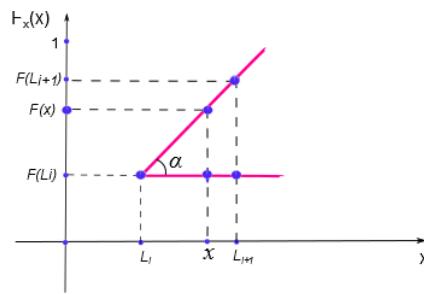


FIGURE 3.8: Le calcul de la médiane par extrapolation.

3.4 Paramètres de dispersion

Définition 22

La variance est la quantité

$$Var(x) = \sum_{i=1}^n f_i (\bar{x} - C_i)^2.$$

Remarque 15

Pour le calcul, on utilise (voir Chapitre 2, Théorème 1)

$$Var(x) = \sum_{i=1}^n f_i C_i^2 - \bar{x}^2.$$

Définition 23

La quantité

$$\sigma_x = \sqrt{Var(x)}$$

s'appelle l'écart type de la V.S X.

Nous généralisons la notion de la médiane dans la définition suivante.

Définition 24

Pour $i \in \{1, 2, 3\}$, la quantité Q_i tel que $F(Q_i) = \frac{i}{4}$ s'appelle le i^{em} quartile.

Exemple 19

Pour $i = 2$, Q_2 tel que $F(Q_2) = \frac{2}{4} = 0.5$. Donc, $Q_2 = Me$.

La détermination ou le calcul de Q_i se fait exactement comme le calcul de la médiane (graphiquement ou analytiquement).

Interprétation : Il y a 25 % d'individus dont la valeur du caractère est dans l'intervalle $[a_0, Q_1]$. De même pour les autres quartiles. Ces intervalles s'appellent "intervalles interquartiles".

$$Q_1 \rightarrow 25\%,$$

$$Q_2 \rightarrow 50\%,$$

$$Q_3 \rightarrow 75\%.$$

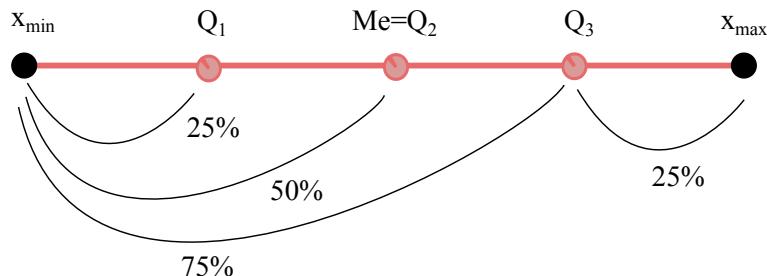


FIGURE 3.9: Les quartiles.

3.5 Exercices corrigés

Exercice 19

- Classer ces statistiques selon leurs natures (indicateur de position ou de dispersion)

	Minimum	Moyenne	Écart-type	Mode	Médiane	Premier quartile
Position						
Dispersion						

Solution Les natures des statistiques sont classées dans ce tableau,

Position	Minimum, Moyenne, Médiane, Mode, Premier quartile
Dispersion	Écart-type

Exercice 20

- Chez un fabricant de tubes de plastique, on a prélevé un échantillon de 100 tubes dont on a mesuré le diamètre en décimètre.

1.94	2.20	2.33	2.39	2.45	2.50	2.54	2.61	2.66	2.85
1.96	2.21	2.33	2.40	2.46	2.51	2.54	2.62	2.68	2.87
2.07	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.90
2.09	2.26	2.34	2.40	2.47	2.52	2.55	2.62	2.68	2.91
2.09	2.28	2.35	2.40	2.48	2.52	2.56	2.62	2.71	2.94
2.12	2.29	2.36	2.41	2.49	2.52	2.56	2.63	2.73	2.95
2.13	2.30	2.37	2.42	2.49	2.53	2.57	2.63	2.75	2.99
2.14	2.31	2.38	2.42	2.49	2.53	2.57	2.65	2.76	2.99
2.19	2.31	2.38	2.42	2.49	2.53	2.59	2.66	2.77	3.09
2.19	2.31	2.38	2.42	2.50	2.54	2.59	2.66	2.78	3.12

1. Identifier la population, les individus, le caractère et son type.
2. En utilisant la méthode de Yule puis de Sturge, établir le tableau statistique (Faites débuter la première classe par la valeur 1.94).
3. Tracer l'histogramme de cette variable statistique.
4. Déterminer par le calcul la valeur du diamètre au-dessous de laquelle se trouvent 50% des tubes de plastique. Que représente cette valeur.
5. Déterminer par le calcul le pourcentage de tubes ayant un diamètre inférieur à 2.58.

Solution 1 - Identification de cet épreuve statistique,

- Population : les tubes.

- Individus : le tube.
- Caractère : le diamètre.
- Type : quantitative continue.
- Modalités : 1.94,..., 3.12.

2 - Par la méthode de Yule, nous avons

$$k = 2.5 \sqrt[4]{N} = 2.5 \sqrt[4]{100} = 7.9 \simeq 8.$$

Par la méthode de Sturge, nous avons

$$k = 1 + 3.3 \log_{10}(N) = 1 + 3.3 \log_{10}(100) = 7.6 \simeq 8.$$

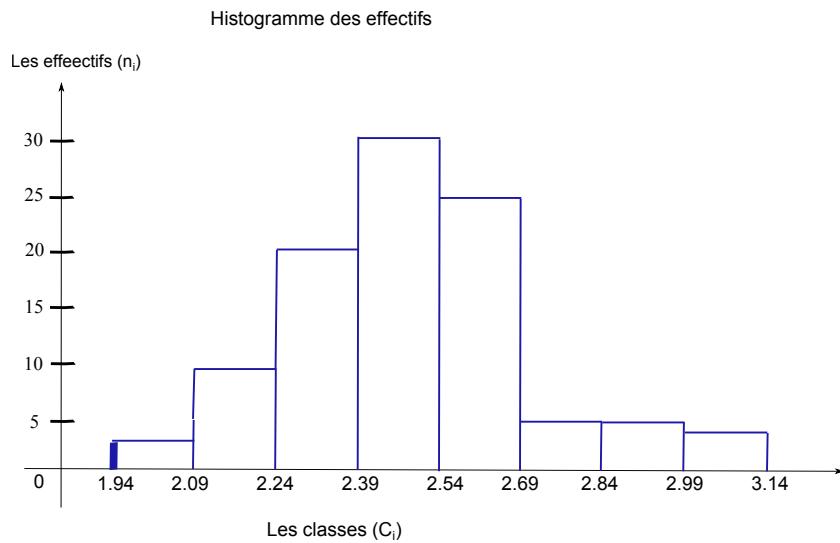
Nous avons donc l'amplitude qui égale

$$a_i = \frac{x_{\max} - x_{\min}}{k} \simeq 0.15.$$

Nous obtenons le tableau statistique suivant,

X	n_i	f_i	N_i	F_i
[1.94, 2.09[3	0.03	3	0.03
[2.09, 2.24[9	0.09	12	0.012
[2.24, 2.39[18	0.18	30	0.3
[2.39, 2.54[29	0.29	59	0.59
[2.54, 2.69[25	0.25	84	0.84
[2.69, 2.84[6	0.06	90	0.90
[2.84, 2.99[6	0.06	96	0.96
[2.99, 3.14[4	0.04	100	1
\sum	100	1	\	\

3 - Nous dessinons l'histogramme de cette variable,



4 - Cette valeur représente la médiane. Le calcul se fait par extrapolation

$$\tan(\alpha) = \frac{0.59 - 0.3}{2.54 - 2.39} = \frac{0.5 - 0.3}{Me - 2.39}.$$

Nous trouvons $Me = 2.5$.

5 - Le calcul du pourcentage de tubes ayant un diamètre inférieur à 2.58 se fait de la même manière et nous avons

$$\tan(\alpha) = \frac{0.84 - 0.59}{2.69 - 2.54} = \frac{x - 0.59}{2.58 - 2.54}.$$

Nous trouvons que la valeur cherché est égale à 0.66 (66%).

Exercice 21

- Une étude sur le budget consacré aux vacances d'été auprès de ménages a donné les résultats suivants

Budget X	Fréquence cumulée	Fréquences
[800, 1000[0.08	
[1000, 1400[0.18	
[1400, 1600[0.34	
[1600, β [0.64	
$[\beta, 2400[$	0.73	
$[2400, \alpha[$	1	

Le travail demandé :

- Certaines données sont manquantes. Calculer la borne manquante α sachant que l'étendue de la série est égale à 3200.
- Calculer les fréquences dans le tableau.
- Calculer la borne manquante β dans les deux cas suivants :
 1. Le budget moyen est égal à 1995.
 2. Le budget médian est égal à 1920.

Solution - On sait que l'étendue est égale au maximum moins le minimum. Ainsi,

$$3200 = x_{\max} - x_{\min} = \alpha - 800,$$

et donc $\alpha = 4000$.

- Nous complétons le tableau comme suit

Budget X	Fréquence cumulée	Fréquences
[800, 1000[0.08	0.08
[1000, 1400[0.18	0.1
[1400, 1600[0.34	0.16
[1600, β [0.64	0.3
$[\beta, 2400[$	0.73	0.09
$[2400, \alpha[$	1	0.27

- Le calcul la borne manquante β dans le cas où le budget moyen est égal à 1995, c'est à dire, $\bar{x} = 1995$ se fait comme suit

$$\bar{x} = 1995 = 0.08 \times 900 + 0.1 \times 1200 + 0.16 \times 1500 + 0.3 \times \frac{1600 + \beta}{2} + 0.09 \times \frac{\beta + 2400}{2} + 0.27 \times 3200.$$

Ce qui implique que

$$1644 + 0.195 \times \beta = 1995,$$

et on trouve $\beta = 1800$.

- Le calcul de la borne manquante β dans le cas où le budget médian est égal à 1920, c'est à dire, $Me = 1920$ se fait comme suit : il faut raisonner par interpolation linéaire sur

l'intervalle $[1600 - \beta]$. On pose le rapport des distances suivant,

$$\frac{1920 - 1600}{\beta - 1600} = \frac{0.5 - 0.34}{0.64 - 0.34},$$

et on trouve $\beta = 2200$.

3.6 Exercices supplémentaires

Exercice 22

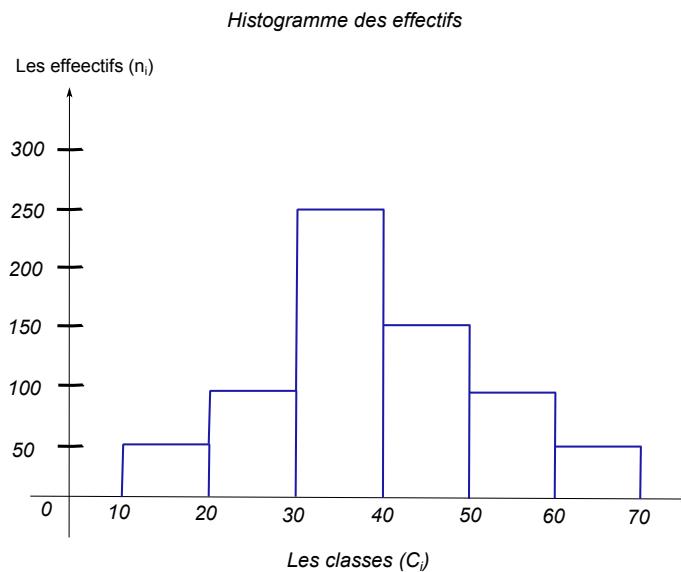
- On considère la variable "temps vécu dans le logement" pour laquelle on a obtenu le tableau d'effectifs suivants :

x_i	$[0, 1[$	$[1, 2[$	$[2, 3[$	$[3, 5[$	$[5, 11[$	$[11, 16[$	$[16, 21[$	$[21, 26[$
n_i	35	36	32	25	20	18	16	7

1. Quel est le type de cette variable ?
2. Déterminer la médiane ainsi que les 1^{er} et 3^{ème} quartiles. Interpréter ces différents indices de position.
3. A cause d'une erreur de saisie, la borne supérieure 26 a été remplacée par 66, cela a-t-il un impact sur la détermination de la médiane ?

Exercice 23

- Dans une gare routière, on évalue le temps d'attente des voyageurs en minutes. Voici l'histogramme des fréquences absolues de cette variable.

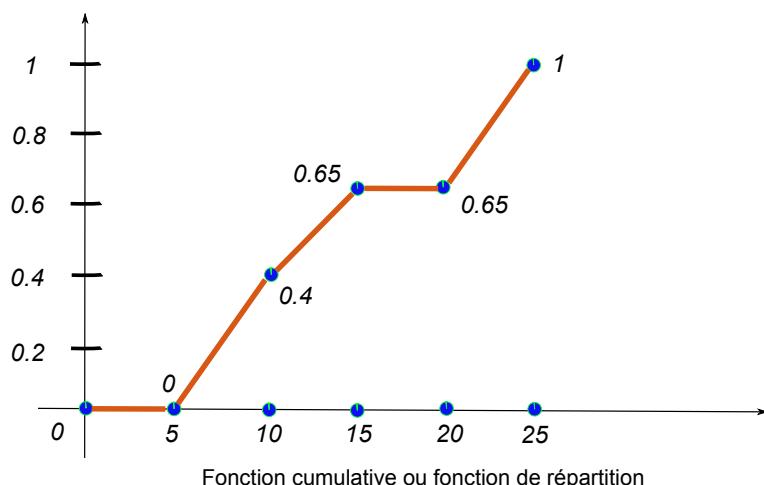


1. Déterminer la variable statistique X et son type et sa population.
2. Déterminer le nombre de voyageurs.
3. Depuis le graphe, déterminer le tableau statistique.
4. Tracer la fonction cumulative.
5. Déterminer le mode graphiquement et dire ce que représente cette valeur par rapport à notre étude.
6. Calculer la médiane à partir du graphe de la fonction cumulative.
7. Calculer la moyenne et l'écart type.

Exercice 24

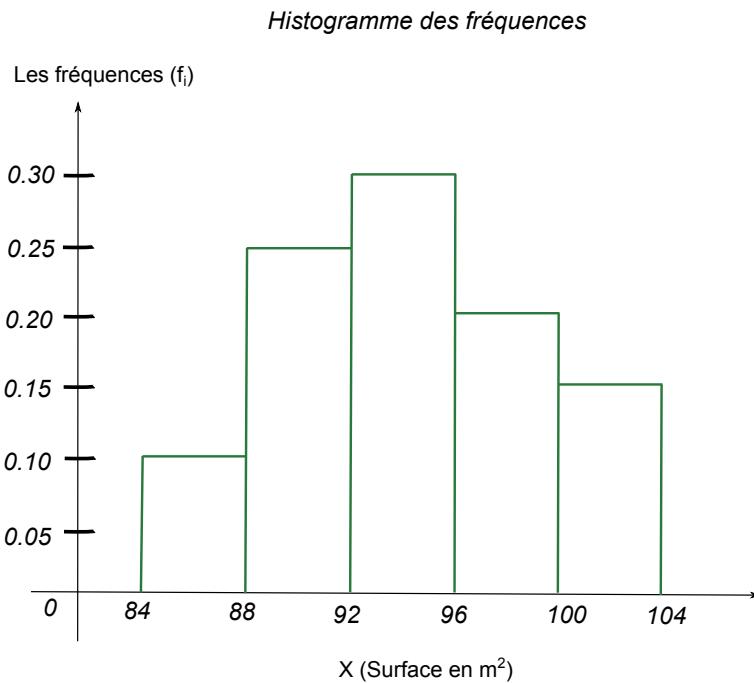
- Le traitement de l'information sur un caractère X a permis de dresser sa fonction cumulative (fonction de répartition dans la figure ci-dessous).

1. Dresser le tableau statistique du caractère X .
2. Tracer l'histogramme du caractère X .
3. Calculer la moyenne et l'écart type.
4. Déduire graphiquement la médiane.
5. Déduire graphiquement le mode.



Exercice 25

- Soit X la surface d'une maison mesurée en m^2 . Le traitement de l'information relatif à 100 maisons a permis de dresser l'histogramme de la variable statistique X (Voir la figure suivante).



1. Calculer la moyenne de la variable statistique X .
2. Déterminer l'écart type de la variable statistique X .
3. Tracer la fonction cumulative et déduire graphiquement la médiane.
4. Donner la définition du mode et trouver le graphiquement.

Exercice 26

- Le tableau statistique suivant représente les montants, en Dinars, des ventes réalisées au mois de Janvier par des employés d'une société :

Modalités C_i	Effectifs n_i	Fréquences f_i	Effectifs cumulés ?	Fréquences cumulées ?
$[0, 1000[$	20	?	?	?
$[1000, 2000[$?	0.35	?	?
$[2000, ?[$?	?	?	?
$[?, ?[$	15	?	100	?
Total	?	?	/	/

1. Compléter le tableau statistique précédent.
2. Déterminer la variable statistique étudiée et son type.
3. Déterminer le pourcentage des ventes qui ont réalisé un montant entre 0 et 2000 Dinars.
4. Représenter graphiquement les effectifs et les fréquences cumulées.
5. Calculer la moyenne arithmétique et donner sa signification.
6. Trouver le mode et la médiane graphiquement, puis par le calcul, et donner leurs significations.
7. Calculer la variance et l'écart-type de cette série statistique.

Chapitre 4

Étude d'une variable statistique à deux dimensions

Dans les chapitres précédents, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salariés en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables). Donc, soit Ω une population et

$$\begin{aligned} Z: \Omega &\rightarrow \mathbb{R}^2, \\ w &\mapsto Z(w) = (X(w), Y(w)), \end{aligned}$$

ou directement

$$\begin{aligned} (X, Y): \Omega &\rightarrow \mathbb{R}^2, \\ w &\mapsto (X(w), Y(w)). \end{aligned}$$

Dans ce cas, Z est dite variable statistique à deux dimensions avec $\text{Card}(\Omega) = N$, avec N un entier fini. Le couple (X, Y) est appelé le couple de la variable statistique.

Exemple 20

- On observe simultanément sur un échantillon de 200 foyers, le nombre d'enfants X et le nombre de chambres Y .
- On observe sur un échantillon de 20 foyers, le revenu mensuel X en Da et les dépenses mensuelles Y .
- Au près des étudiants pris au hasard parmi une section de L2 génie civil, on

observe les notes de Math3 X et de Statistique Y .

- *Une entreprise mène une étude sur la liaison entre les dépenses mensuelles en publicité X et le volume des ventes Y qu'elle réalise.*

4.1 Représentation des séries statistiques à deux variables

Les séries statistiques à deux variables peuvent être présentées de deux façons.

Présentation 1

A chaque w_i , on associe (x_i, y_i) , c'est à dire,

$$w_i \longrightarrow (x_i, y_i).$$

On rassemblera les données comme dans le tableau suivant

w_i	w_1	w_2	...	w_N
Variable X	$X(w_1)$	$X(w_2)$...	$X(w_N)$
Variable Y	$Y(w_1)$	$Y(w_2)$...	$Y(w_N)$

Cette représentation on la notera "présentation 1". Nous allons utiliser toujours les notations suivantes :

$$x_i := X(w_i)$$

et $y_i := Y(w_i)$.

Exemple 21

Soit Ω l'ensemble de 8 étudiants. Nous avons le tableau suivant

w_i	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$X(w)$	8	2	6	6	11	10	7	2
$Y(w)$	9	10	11	7	14	16	12	5

avec X représente le nombre d'heures passées à préparer l'examen de statistique par étudiant et Y représente la note sur 20 obtenue à l'examen par l'étudiant.

Lors de cette représentation, nous pouvons traduire le tableau associé dans une figure appelée "le nuage de points" ou "diagramme de dispersion" (voir Figure 4.1). Cette représentation est obtenue en mettant dans un repère cartésien chaque couple d'observation (x_i, y_j) par un point.

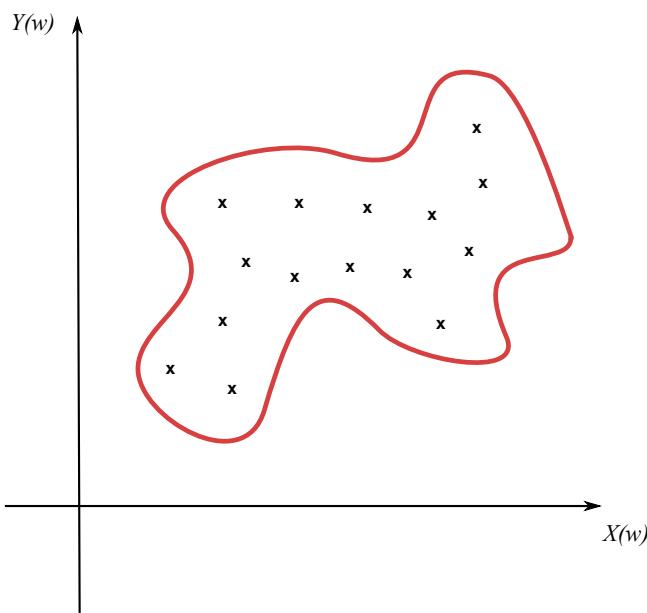


FIGURE 4.1: Représentation sous forme de nuage de points.

Présentation 2

Soit la variable statistique Z donnée par le couple (X, Y) . Soient x_1, \dots, x_k et y_1, \dots, y_l les valeurs prises respectivement par X et Y . Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à k et pour j allant de 1 à l ,

$$z_{ij} := (x_i, y_j).$$

La variable statistique Z prend $k \times l$ valeurs. Lors de cette étude, nous avons le tableau à double entrée (ou tableau de contingence) suivant (discrète ou continue)

$\mathbf{X} \setminus \mathbf{Y}$	$C'_1 = [L'_1, L'_2[\text{ ou } y_1]$...	$C'_l = [L'_l, L'_{l+1}[\text{ ou } y_l]$	Marginale % à \mathbf{X}
$C_1 = [L_1, L_2[\text{ ou } x_1]$	$n_{11} \text{ ou } f_{11}$...	$n_{1l} \text{ ou } f_{1l}$	$n_{1\bullet} \text{ ou } f_{1\bullet}$
$C_2 = [L_2, L_3[\text{ ou } x_2]$	$n_{21} \text{ ou } f_{21}$...	$n_{2l} \text{ ou } f_{2l}$	$n_{2\bullet} \text{ ou } f_{2\bullet}$
$C_3 = [L_3, L_4[\text{ ou } x_3]$	$n_{31} \text{ ou } f_{31}$...	$n_{3l} \text{ ou } f_{3l}$	$n_{3\bullet} \text{ ou } f_{3\bullet}$
...
$C_k = [L_k, L_{k+1}[\text{ ou } x_k]$	$n_{k1} \text{ ou } f_{k1}$...	$n_{kl} \text{ ou } f_{kl}$	$n_{k\bullet} \text{ ou } f_{k\bullet}$
Marginale % à \mathbf{Y}	$n_{\bullet 1} \text{ ou } f_{\bullet 1}$...	$n_{\bullet l} \text{ ou } f_{\bullet l}$	N

Cette représentation on la notera "présentation 2". A chaque couple (x_i, y_i) , on a n_{ij} est l'effectif qui représente le nombre d'individus qui prennent en même temps la valeur x_i et y_i , c'est à dire,

$$n_{ij} := \text{Card}\{w \in \Omega : Z(w) = z_{ij}\}.$$

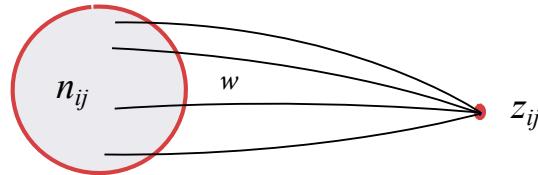


FIGURE 4.2: Le nombre d'individus qui prennent en même temps la valeur x_i et y_i .

Nous notons par f_{ij} la fréquence du couple (x_i, y_i) . Cette fréquence est donnée par

$$f_{ij} := \frac{n_{ij}}{N},$$

avec

$$N = \text{Card}(\Omega),$$

$$\begin{aligned} &= \sum_{j=1}^l \sum_{i=1}^k n_{ij}, \\ &= \sum_{i=1}^k \sum_{j=1}^l n_{ij}. \end{aligned}$$

Le calcul ou le développement de cette double série est donné par

$$\begin{aligned}
 \sum_{i=1}^k \sum_{j=1}^l n_{ij} &= n_{11} + n_{12} + n_{13} + \dots + n_{1l} \\
 &\quad + n_{21} + n_{22} + n_{23} + \dots + n_{2l} \\
 &\quad \ddots \quad \ddots \quad \ddots \quad \ddots \\
 &\quad + n_{k1} + n_{k2} + n_{k3} + \dots + n_{kl}.
 \end{aligned}$$

Remarque 16

Nous avons la propriété suivante,

$$\sum_{i=1}^k \sum_{j=1}^l f_{ij} = 1.$$

Lois marginales

Sur la marge du tableau de contingence, on peut extraire les données seulement par rapport à X et seulement par rapport à Y (voir le tableau de contingence établi auparavant).

1. Effectifs et fréquences marginale par rapport à Y : nous avons, pour $j = 1 \dots l$,

$$n_{\bullet j} := \sum_{i=1}^k n_{ij},$$

et

$$f_{\bullet j} := \frac{n_{\bullet j}}{N} = \sum_{i=1}^k f_{ij}.$$

2. Effectifs et fréquences marginale par rapport à X : nous avons, pour $i = 1 \dots k$,

$$n_{i\bullet} := \sum_{j=1}^l n_{ij},$$

et

$$f_{i\bullet} := \frac{n_{i\bullet}}{N} = \sum_{j=1}^l f_{ij}.$$

Remarque 17

Nous avons les propriétés suivantes

$$\sum_{i=1}^k n_{i\bullet} = \sum_{j=1}^l n_{\bullet j} = N \quad \text{et} \quad \sum_{i=1}^k f_{i\bullet} = \sum_{j=1}^l f_{\bullet j} = 1.$$

Exercice 27

Nous considérons 10 salariés qui sont observés à l'aide de deux variables "âge" et "salaire". Les informations brutes (pas encore traitées ou façonnées) sont données dans le tableau suivant,

Salaire	6000	7400	7500	8200	8207	8900	9100	9900	9950	10750
Age	15	26	20	43	47	37	52	34	50	44

1. Déterminer le tableau de contingence (X : âge, Y : salaire). Pour l'âge et pour le salaire, former respectivement des classes de pas de 10 ans et de 1000 Da.
2. Calculer f_{21} , f_{12} , f_{45} et f_{33} .
3. Déterminer les effectifs marginaux de X et de Y . Tracer le nuage de points.
4. Déterminer le tableau statistique des deux séries marginales X et Y .

Solution : En utilisant les hypothèses, nous considérons les classes suivantes,

$$[15, 25[, [25, 35[, [35, 45[, [45, 55[,$$

pour l'âge et

$$[6, 7[, [7, 8[, [8, 9[, [9, 10[, [10, 11[,$$

pour le salaire ($\times 1000$). De plus, nous avons

$$\text{Nombre de classe} = \frac{e}{a_{\text{âge}}} = \frac{x_{\max} - x_{\min}}{a_{\text{âge}}} = \frac{52 - 15}{10} = 3.7 \simeq 4 \text{ classes},$$

pour l'âge et

$$\text{Nombre de classe} = \frac{e}{a_{\text{sal}}} = \frac{y_{\max} - y_{\min}}{a_{\text{sal}}} = \frac{10750 - 6000}{1000} = 4.75 \simeq 5 \text{ classes},$$

pour le salaire. Cette série statistique est représentée par le tableau suivant,

Age \ Salaire	[6, 7[[7, 8[[8, 9[[9, 10[[10, 11[$n_{i\bullet}$	$f_{i\bullet}$
[15, 25[1	1	0	0	0	0	0.2
[25, 35[0	1	0	1	0	2	0.2
[35, 45[0	0	2	0	1	3	0.3
[45, 55[0	0	1	2	0	3	0.3
$n_{\bullet j}$	1	2	3	3	1	10	1
$f_{\bullet j}$	0.1	0.2	0.3	0.3	0.1	1	\emptyset

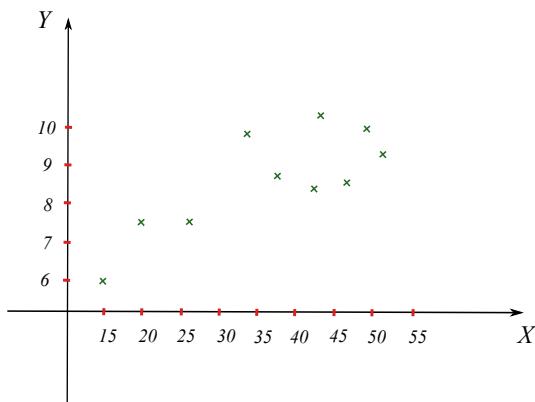
De ce fait, nous avons

$$f_{12} = \frac{n_{12}}{N} = \frac{1}{10} = 0.1, \quad f_{21} = \frac{n_{21}}{N} = \frac{0}{10} = 0,$$

et

$$f_{45} = \frac{n_{45}}{N} = \frac{0}{10} = 0, \quad f_{33} = \frac{n_{33}}{N} = \frac{2}{10} = 0.2.$$

Le nuage de points est tracé, à partir des données brutes, dans la figure suivante.



Enfin, les deux tableaux statistiques de X et de Y sont donnés, respectivement, par

X	$n_{i\bullet}$	$f_{i\bullet}$	x_i le centre
[15, 25[2	0.2	20
[25, 35[2	0.2	30
[35, 45[3	0.3	40
[45, 55[3	0.3	50

Y	$n_{i\bullet}$	$f_{i\bullet}$	y_i le centre
[6, 7[1	0.1	6.5
[7, 8[2	0.2	7.5
[8, 9[3	0.3	8.5
[9, 10[3	0.3	9.5
[10, 11[1	0.1	10.5

4.2 Description numérique

4.2.1 Caractéristique des séries marginales

Dans le cas d'une variable statistique à deux dimensions X et Y , les moyennes sont données respectivement par

$$\bar{x} := \frac{1}{N} \sum_{i=1}^k n_{i\bullet} x_i = \sum_{i=1}^k f_{i\bullet} x_i \quad (\text{moyenne de } X),$$

et

$$\bar{y} := \frac{1}{N} \sum_{j=1}^l n_{\bullet j} y_j = \sum_{j=1}^l f_{\bullet j} y_j \quad (\text{moyenne de } Y).$$

Remarque 18

Dans le cas continu, x_i et y_j représentent respectivement le centre des classes de X et Y , c'est à dire,

$$x_i = \frac{L_{i+1} + L_i}{2} \quad \text{et} \quad y_j = \frac{L_{j+1} + L_j}{2}.$$

Exemple 22

Nous calculons \bar{x} et \bar{y} pour l'exercice traité précédemment. Nous avons la moyenne d'âge

$$\bar{x} = \frac{1}{10}(40 + 60 + 120 + 150) = 37 \text{ ans.}$$

et la moyenne du salaire

$$\bar{y} = \frac{1}{10}(6.5 + 15 + 25.5 + 28.5 + 10.5) \times 100 = 8600 \text{ Da.}$$

Nous définissons maintenant la variance de X et la variance de Y comme suit,

$$Var(X) := \overline{x^2} - (\bar{x})^2, \quad \text{avec} \quad \overline{x^2} := \frac{1}{N} \sum_{i=1}^k n_{i\bullet} x_i^2 = \sum_{i=1}^k f_{i\bullet} x_i^2,$$

et

$$Var(Y) := \overline{y^2} - (\bar{y})^2, \quad \text{avec} \quad \overline{y^2} := \frac{1}{N} \sum_{j=1}^l n_{\bullet j} y_j^2 = \sum_{j=1}^l f_{\bullet j} y_j^2.$$

Les écarts-type de X et de Y sont donnés, respectivement, par

$$\sigma_X := \sqrt{Var(X)} \quad \text{et} \quad \sigma_Y := \sqrt{Var(Y)}.$$

4.2.2 Série conditionnelle

La notion de série conditionnelle est essentielle pour comprendre l'analyse de la régression. Un tableau de contingence se compose en autant de séries conditionnelles suivant chaque ligne et chaque colonnes.

Série conditionnelle par rapport à X

Elle est notée par X/y_j (ou X_j) et on dit que c'est la série conditionnelle de X sachant que $Y = y_j$. Nous calculons dans ce cas la fréquence conditionnelle $f_{i/j}$ (f_i sachant j), pour $i = 1, \dots, k$, par

$$f_{i/j} := \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

Nous avons aussi la moyenne conditionnelle \bar{x}_j , c'est à dire la moyenne des valeurs de X sous la condition y_j , elle est définie par

$$\bar{x}_j := \sum_{i=1}^k f_{i/j} x_i = \frac{1}{n_{\bullet j}} \sum_{i=1}^k n_{ij} x_i.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{X_j} := \sqrt{Var(X_j)}$ avec

$$Var(X_j) := \sum_{i=1}^k f_{i/j} (x_i - \bar{x}_j)^2 = \overline{x^2}_j - (\bar{x}_j)^2.$$

Série conditionnelle par rapport à Y

Elle est notée par Y/x_j (ou Y_j) et on dit que c'est la série conditionnelle de Y sachant que $X = x_i$. Nous calculons aussi dans ce cas la fréquence conditionnelle $f_{j/i}$ (f_j sachant i), pour $j = 1, \dots, l$, par

$$f_{j/i} := \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

Nous avons aussi la moyenne conditionnelle \bar{y}_i , c'est à dire la moyenne des valeurs de Y sous la condition x_i , elle est définie par

$$\bar{y}_i := \sum_{j=1}^l f_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^l n_{ij} y_j.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{Y_i} := \sqrt{Var(Y_i)}$ avec

$$Var(Y_i) := \sum_{j=1}^l f_{j/i} (y_j - \bar{y}_i)^2 = \bar{y}_i^2 - (\bar{y}_i)^2.$$

4.2.3 Notion de covariance

Nous notons par $Cov(X, Y)$ la covariance entre les variables X et Y . La covariance est un paramètre qui donne la variabilité de X par rapport à Y (voir Figure 4.3).

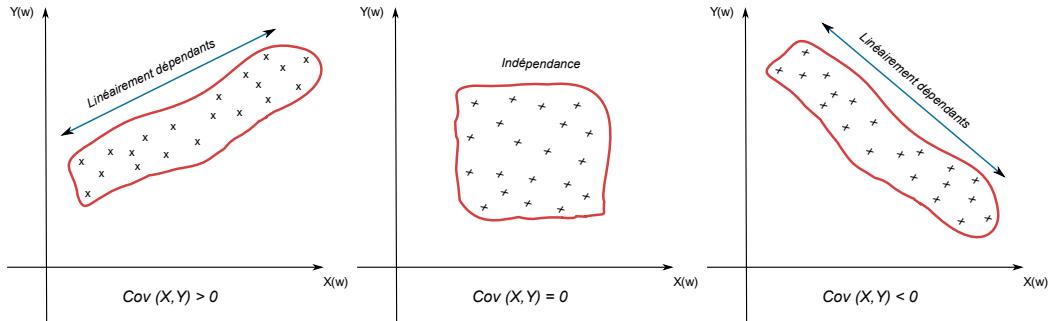


FIGURE 4.3: La covariance et la variabilité.

La covariance se calcule par l'expression suivante

$$Cov(X, Y) = \bar{xy} - \bar{x} \bar{y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^l n_{ij} x_i y_j - \bar{x} \bar{y}.$$

Nous avons aussi cette formule

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^l n_{ij}(x_i - \bar{x})(y_j - \bar{y}).$$

Remarque 19

Dans le cas où nous avons un tableau des données brutes "representation 1" (nous n'avons pas d'effectifs), nous avons les formules suivantes

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i \quad \text{et} \quad \bar{y} = \frac{1}{N} \sum_{i=1}^n y_i.$$

De plus, nous avons

$$\bar{xy} = \frac{1}{N} \sum_{i=1}^n x_i y_i.$$

Remarque 20

La covariance est une notion qui généralise la variance, En effet,

$$Cov(X, X) = Var(X) \quad \text{et} \quad Cov(Y, Y) = Var(Y).$$

Cela provient de la définition, c'est à dire,

$$Cov(X, X) = \bar{xx} - \bar{x} \bar{x} = \bar{x^2} - \bar{x}^2 = Var(X).$$

Définition 25

On dit que deux variables statistiques X et Y sont indépendantes si et seulement si, pour tout i et j ,

$$f_{ij} = f_{i\bullet} \times f_{\bullet j}.$$

Il suffit que cette égalité ne soit pas vérifiée dans une seule cellule pour que les deux variables ne soient pas indépendantes. De manière équivalente, pour tout i et j ,

$$N \times n_{ij} = n_{i\bullet} \times n_{\bullet j}.$$

Dans ce cas, si X et Y sont indépendantes alors (réciproque est fausse) $Cov(X, Y) = 0$.

Cette définition donne une interprétation intéressante d'indépendance ; elle signifie que dans ce cas, les effectifs des modalités conjointes peuvent se calculer uniquement à partir des distributions marginales, supposées « identiques » aux distributions de X et Y dans la population ; en d'autres termes, si X et Y sont indépendantes, les observations séparées de X et de Y donnent la même information qu'une observation conjointe.

4.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

4.3.1 Coefficient de corrélation

Les coefficients de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation permet d'analyser les relations linéaires (voir ci-dessous). Il existe d'autres coefficients pour les relations non-linéaires et non-monotones, mais ils ne seront pas étudiés dans le cadre de ce cours.

Définition 26

La quantité

$$\rho_{XY} := \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

s'appelle le coefficient de corrélation.

Proposition 3

Le coefficient ρ_{XY} est compris entre $[-1, 1]$, ou encore

$$|\rho_{XY}| \leq 1.$$

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y (voir Figure 4.4 et). Nous avons les deux caractéristiques suivantes (voir Figures 4.5 et 4.6)¹ :

1. Source : https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

- Plus le module de ρ_{XY} est proche de 1 plus X et Y sont liées linéairement.
- Plus le module de ρ_{XY} est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y .

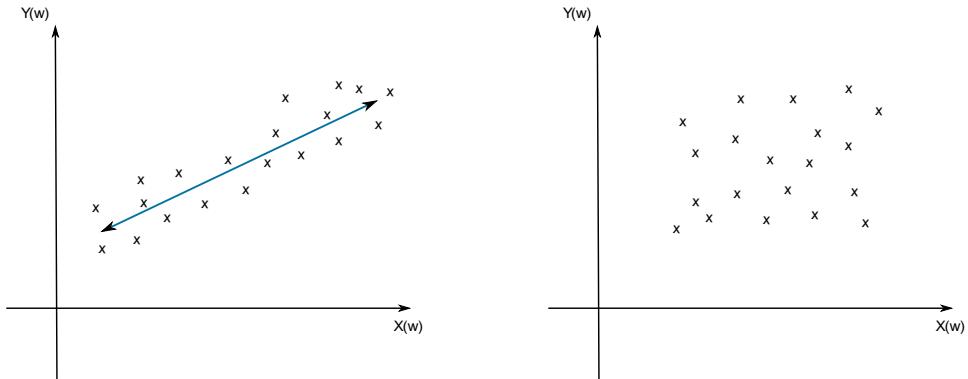


FIGURE 4.4: A gauche, le coefficient de corrélation est proche de 1. A droite, le coefficient de corrélation est proche de 0.

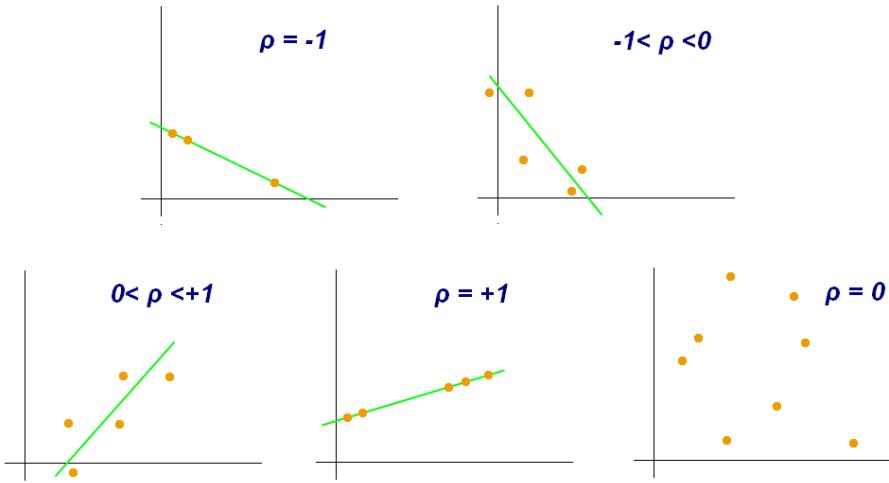


FIGURE 4.5: Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation .

Remarque 21

Par définition, si $\rho_{XY} = 0$, alors $Cov(X, Y) = 0$.

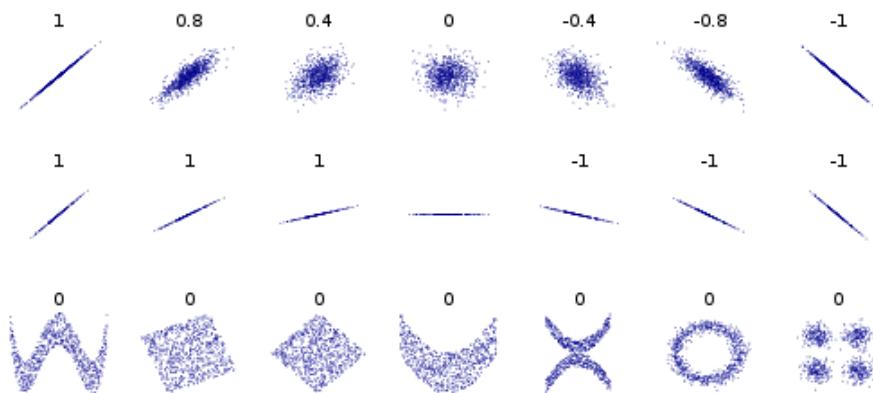


FIGURE 4.6: La corrélation reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (en bas). La figure au centre a une pente de 0, mais dans ce cas, le coefficient de corrélation est indéfini car la variance de Y est nulle. .

4.3.2 Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

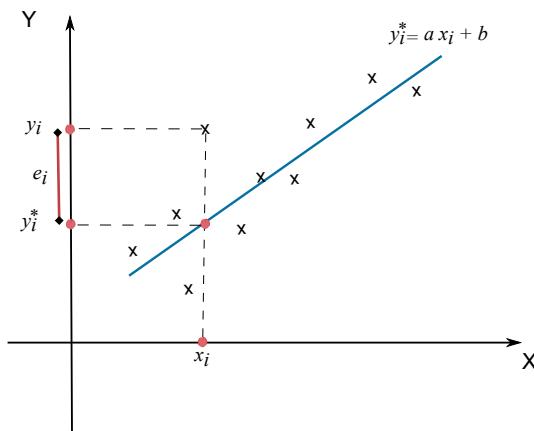


FIGURE 4.7: La droite la plus proche possible de chacun des points.

Pour cela, on utilise la méthode des moindres carrés. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X , c'est à dire,

$$Y = aX + b,$$

telle que la distance entre le nuage de points et la droite soit minimale. Cette distance

matérialise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédict par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des moindres carrés consiste à chercher la valeur des paramètres a et b qui minimise la somme des erreurs élevées au carré.

On pose

$$\sum_{i=1}^n e_i^2 = U(a, b),$$

avec e_i est l'erreur commise sur chaque observation, c'est à dire,

$$|e_i| = |y_i - y_i^\star| = |y_i - ax_i - b|.$$

La méthode des moindres carrés consiste donc à minimiser la fonction U (la somme des erreurs commises). Nous avons la condition de minimisation suivante,

$$\frac{\partial U}{\partial a} = \frac{\partial U}{\partial b} = 0,$$

avec

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

L'équation $\frac{\partial U}{\partial b} = 0$ donne

$$\sum_{i=1}^n -2(y_i - ax_i - b) = 0.$$

Ce qui implique que

$$\left(\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n 1 = 0 \right) \times \frac{1}{N}.$$

Par conséquent, nous obtenons

$$\bar{y} - a\bar{x} - b = 0,$$

c'est à dire,

$$b = \bar{y} - a\bar{x}.$$

De même, après calculer, $\frac{\partial U}{\partial a} = 0$ implique que

$$a = \frac{Cov(X, Y)}{Var(X)}.$$

Donc, la droite de régression, qui rend la distance entre elle et les points minimale, est

donnée par

$$D(Y/X) : Y = aX + b,$$

avec

$$a = \frac{Cov(X, Y)}{Var(X)} \quad \text{et} \quad b = \bar{y} - a\bar{x}.$$

Ou bien

$$D(X/Y) : X = a'Y + b',$$

avec

$$a' = \frac{Cov(X, Y)}{Var(Y)} \quad \text{et} \quad b' = \bar{x} - a'\bar{y}.$$

Remarque 22

Le coefficient de corrélation ρ_{XY} permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir Figure 4.8),

- Si $|\rho_{XY}| < 0.7$, alors l'ajustement linéaire est refusé (droite refusée).
- Si $|\rho_{XY}| \geq 0.7$, alors l'ajustement linéaire est accepté (droite acceptée).

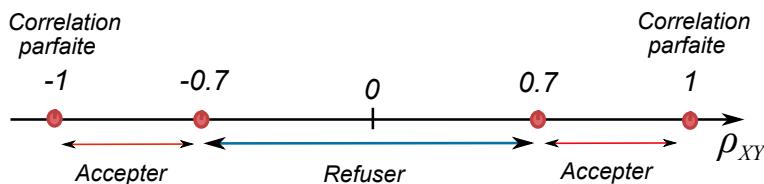


FIGURE 4.8: La zone d'acceptation ou de refus de l'ajustement linéaire.

4.4 Exercices corrigés

Exercice 28

Nous considérons 10 joueurs et soient :

- Y la variable qui représente le nombre de jeux auquel un joueur joue.
- X la variable qui représente le gain ou perte (+1 s'il gagne 10 Da et -1 s'il perd 10 Da et 0 sinon).

Nous avons le tableau de contingence suivant,

$\mathbf{X} \setminus \mathbf{Y}$	1	2	3	4	$n_{i\bullet}$
-1	0	1	2	2	
0	1	1	0	1	
1	0	1	1	0	
$n_{\bullet j}$					

1. Compléter le tableau ci-dessus.

2. Calculer $cov(X, Y)$.

Solution Les lois marginales sont données dans ce tableau,

$\mathbf{X} \setminus \mathbf{Y}$	1	2	3	4	$n_{i\bullet}$
-1	0	1	2	2	5
0	1	1	0	1	3
1	0	1	1	0	2
$n_{\bullet j}$	1	3	3	3	N=10

La covariance est calculée à partir de

$$Cov(X, Y) = \bar{xy} - \bar{x} \bar{y}.$$

Nous avons

$$\bar{x} = \frac{1}{N} \sum_{i=1}^3 n_{i\bullet} x_i = -0.3,$$

et

$$\bar{y} = \frac{1}{N} \sum_{j=1}^4 n_{\bullet j} y_j = 2.8.$$

De plus, nous avons

$$\bar{xy} = \frac{1}{N} \sum_{i=1}^3 \sum_{j=1}^4 n_{ij} x_i y_j = -1.$$

Donc,

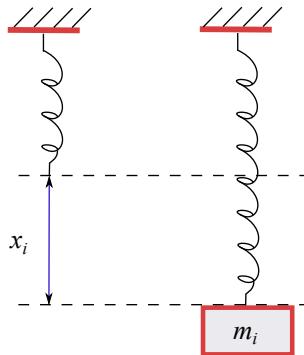
$$Cov(X, Y) = -0.16.$$

Exercice 29

Dans un TP de physique, on a les données suivantes :

x_i	0	0.5	1.1	1.5	1.9
m_i	0	10	20	30	40

La variable m_i représente les différentes masses appliquées comme dans le schéma ci-dessous et la variable x_i les hauteurs induits depuis l'état initial.



1. Déterminer $D(m/x)$.
2. Déterminer $D(x/m)$.
3. Tracer le nuage de points et les deux droites. Représenter le point de coordonnée (\bar{x}, \bar{y}) .
4. Peut-on déterminer x si $m = 51.75 \text{ Kg}$?

Solution Nous déterminons facilement les moyennes $\bar{x} = 1$ et $\bar{m} = 1$. De plus, nous avons

$$\text{Cov}(x, m) = \bar{xm} - \bar{x} \bar{m} = 29.6 - 20 = 9.6.$$

Après calcul, nous avons aussi

$$\text{Var}(X) = \bar{x^2} - (\bar{x})^2 = 0.464 \quad \text{et} \quad \text{Var}(m) = 200.$$

Ce qui implique que

$$\sigma_x = 0.681 \quad \text{et} \quad \sigma_m = 14.14.$$

Dans ce cas, les coefficients de la droite sont donnés par

$$a = \frac{\text{Cov}(x, m)}{\text{Var}(X)} = 20.69 \quad \text{et} \quad b = \bar{m} - a\bar{x} = -0.69.$$

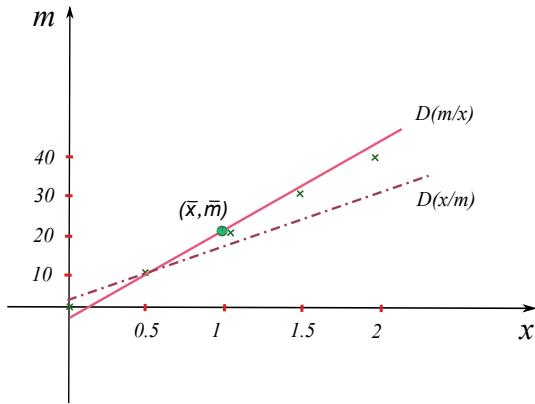
Par conséquent,

$$D(m/x) : m = 20.69x - 0.69.$$

En renversant les axes, nous obtenons

$$D(x/m) : \quad x = 0.048m + 0.04.$$

Nous traçons les deux droites dans la figure ci-dessous ($D(x/m)$ est la symétrie de $D(m/x)$ par rapport à la première bissectrice).



Si on calcule $\rho_{xm} = 0.997 > 0.7$, alors nous avons un ajustement linéaire accepté. Donc, on peut estimer x pour $m = 51.75$ et nous obtenons

$$x = 0.048 \times (51.75) + 0.04 = 2.52.$$

Exercice 30

- Le tableau de contingence suivant est entre le salaire mensuel X et l'ancienneté Y des ouvriers d'une entreprise.

$X(\times 1000) \setminus Y$	[0, 8[[8, 16[[16, 24[[24, 32[Loi marginale
[20, 30[5	6	1	0	
[30, 40[2	4	3	3	
[40, 50[0	2	4	10	
Loi marginale					

1. Etudier les séries marginales.
2. Déterminer si les variables X et Y sont indépendantes.
3. Etudier les séries conditionnelles X/y_3 et Y/x_2 et présenter les résultats pour chaque groupe de séries conditionnelles.

Solution 1- Nous complétons le tableau comme suit

$\mathbf{X} (\times 1000) \setminus \mathbf{Y}$	[0, 8[[8, 16[[16, 24[[24, 32[$n_{i\bullet}$	$f_{i\bullet}$
[20, 30[5	6	1	0	12	0.3
[30, 40[2	4	3	3	12	0.3
[40, 50[0	2	4	10	16	0.4
$n_{\bullet j}$	7	12	8	13	40	1
$f_{\bullet j}$	7/40	12/40	8/40	13/40	1	\emptyset

Les moyennes après le calcul

$$\bar{x} = 36 \ (\times 1000),$$

et

$$\bar{y} = 17.4.$$

La variance et l'écart type de X

$$Var(X) = 69 \quad \text{et} \quad \sigma_X = 8.310.$$

La variance et l'écart type de Y

$$Var(Y) = 78.04 \quad \text{et} \quad \sigma_Y = 8.84.$$

2 - Si on choisit $i = 3$ et $j = 1$, nous obtenons

$$N \times n_{31} = 40 \times 0 = 0,$$

et

$$n_{3\bullet} \times n_{\bullet 1} = 16 \times 7 = 112,$$

qui sont bien évidemment non égaux. Par conséquent, il existe i et j tel que

$$N \times n_{ij} \neq n_{i\bullet} \times n_{\bullet j}.$$

Donc, X et Y ne sont pas indépendants.

3 - La série X/y_3 est la série conditionnelle de X sachant que $Y = y_3$ ($j = 3$). Sa moyenne

est donnée par

$$\bar{x}_3 = 38.75.$$

La série Y/x_2 est la série conditionnelle de Y sachant que $X = x_2$ ($i = 2$). Sa moyenne est donnée par

$$\bar{y}_2 = 16.67.$$

4.5 Exercices supplémentaires

Exercice 31

- Une usine produit des pièces d'une machine. Pour chaque pièce (individu), on dispose du coût de sa production (DA) et du temps nécessaire pour sa réalisation (en heures). Le tableau ci-après (série statistique) donne cette répartition :

<i>Individu</i>	1	2	3	4	5
<i>Temps (X) mesuré en heures</i>	2	3	52	2	4
<i>Coût (Y) mesuré en Dinars</i>	10	16	23	12	18

On donne

- $Cov(X, Y) = \frac{1}{N} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]$. - Coefficient de corrélation = $\frac{Cov(X, Y)}{\sigma(X) \sigma(Y)}$.
- Droite de corrélation linéaire : $Y = \bar{y} - \frac{Cov(X, Y)}{Var(X)}\bar{x} + \frac{Cov(X, Y)}{Var(X)}X$.

1. Calculer la moyenne de la variable statistique X .
2. Calculer la moyenne de la variable statistique Y .
3. Calculer l'écart-type de la variable statistique X .
4. Calculer l'écart-type de la variable statistique Y .
5. Calculer la covariance des variables statistiques X et Y .
6. En supposant qu'il existe une corrélation linéaire entre X et Y , déterminer cette droite de corrélation.
7. Calculer le coefficient de corrélation. Conclusion ?
8. Une nouvelle pièce est réalisée en 6 heures. Estimer le coût de production de cette pièce en utilisant la droite de corrélation établie.

Exercice 32

- Pour les données suivantes

X	1	2	7	4	6
Y	5	4	1	3	2

1. Tracer le nuage de points.
2. Deviner le signe et la valeur du coefficient de corrélation.
3. Calculer le coefficient de corrélation, la pente et l'ordonnée à l'origine de la droite de régression.

Exercice 33

- Soit X et Y deux variables statistiques mesurées sur un même individu. Par exemple, pour l'individu n°3, $X = 2$ et $Y = 8$.

<i>Individu</i>	1	2	3	4	5
X	3	4	2	5	3
Y	12	14	8	19	11

1. Calculer la moyenne de la variable statistique X .
2. Calculer la moyenne de la variable statistique Y .
3. Calculer l'écart-type de la variable statistique X .
4. Calculer l'écart-type de la variable statistique Y .
5. Calculer la covariance des variables statistiques X et Y .
6. En supposant qu'il existe une corrélation linéaire entre X et Y , déterminer cette droite de corrélation.
7. Calculer le coefficient de corrélation. Conclusion ?

Exercice 34

- On vous demande s'il existe une corrélation entre la population de chamois² dans une commune et le nombre de permis de chasse enregistré par l'association de chasse locale.

-
2. Le chamois est une sorte de chèvre des montagnes caractérisé par ses petites cornes en crochets.

Années	1999	2000	2001	2002	2003	2004
Chamois	3200	3650	3430	3890	4200	4350
Permis	202	231	240	225	245	263

Travail à faire :

- Calculer le coefficient de corrélation entre ces deux séries.
- Tracer la droite d'ajustement.

Exercice 35

- Une étude sur le chômage a été faite et qui s'intéresse à l'ancienneté du chômage (X) moins de 24 mois, et l'âge (Y) entre 20 et 35 ans. Les résultats sont donnés par le tableau de contingence suivant :

$X \setminus Y$	[20, 25[[25, 30[[30, 35[
[0, 6[10	8	5
[6, 12[8	9	4
[12, 18[15	11	9
[18, 24[3	6	2

1. Quel est le nombre d'individus qui ont une ancienneté de chômage moins d'un an ?
2. Déterminer les deux distributions marginales.
3. Déterminer la distribution de X conditionnelle à $Y = [25, 30]$, c'est à dire, $X/Y = [25, 30]$.
4. Les variables X et Y sont elles indépendantes ? Justifier.
5. Donner la moyenne arithmétique.
6. Calculer le coefficient de corrélation linéaire. Commenter.
7. Donner l'équation de la droite de régression de Y en fonction de X .
8. Quel sera l'âge d'une personne ayant une ancienneté de chômage de 15 mois.

Exercice 36

- On fait une étude statistique sur 10 sites de commerce électronique, ayant pour but de sonder sur une semaine le nombre de visiteurs et le nombre de commandes. On obtient le tableau suivant :

<i>Le numéro du site (i)</i>	1	2	3	4	5	6	7	8	9	10
<i>Le nombre de connexion (x_i)</i>	80	100	115	110	70	125	105	90	110	95
<i>Le nombre de commandes (y_i)</i>	32	50	62	56	8	80	62	50	62	38

1. Calculer les moyennes arithmétiques de la variable statistique X et de la variable statistique Y .
2. Calculer les écarts-type de la variable statistique X et de la variable statistique Y .
3. Calculer la covariance entre X et Y .
4. Calculer le coefficient de corrélation linéaire entre X et Y . Commenter.
5. Déterminer la droite de corrélation $Y = aX + b$.

Exercice 37

- Nous allons écrire la relation entre la tension U aux bornes d'une batterie de force électromotrice e et de résistance interne R sous la forme $U = -RI + e$. On a procédé à différentes mesures :

<i>Intensité I (Ampère)</i>	0	0.1	0.4	1
<i>Tension U (Volte)</i>	12	11	7	1

1. Calculer la moyenne de la variable I (notée, \bar{I}) puis la moyenne de la variable U (notée, \bar{U}).
2. Calculer l'écart-type de la variable I (noté, $\sigma(I)$) puis l'écart-type de la variable U (noté, $\sigma(U)$).
3. Tracer le nuage de points.
4. Calculer la covariance des variables statistiques I et U ($\text{Cov}(I, U)$).
5. Calculer le coefficient de corrélation linéaire entre I et U . Conclure.
6. Trouver la droite de corrélation $U = -RI + e$.
7. Estimer la tension U si l'intensité $I = 1.05$ Ampère.

Exercice 38

- Une étude a été menée auprès de 12 étudiants afin d'expliquer le score à un examen de mathématiques à partir du temps consacré à la préparation de cet examen. Pour chaque étudiant, on dispose : du temps de révision en heures (variable X) et du score obtenu sur 800 points (variable Y). Les résultats sont :

x_i	4	9	10	14	4	7	12	1	3	8	11	5
y_i	390	580	650	730	410	530	600	350	400	590	640	450

On donne :

$$\sum_{i=1}^{12} x_i^2 = 822, \quad \sum_{i=1}^{12} y_i^2 = 3495600 \quad \text{et} \quad \sum_{i=1}^{12} x_i y_i = 51610.$$

1. Tracer le nuage de points.
2. Calculer le coefficient de corrélation linéaire entre ces deux séries. Commenter.
3. Déterminer la droite de corrélation linéaire $D(Y/X) : Y = aX + b$.

Exercice 39

- On dispose des données suivantes concernant un échantillon de 25 hommes : tailles x_i en mètres et poids y_i en kg (en physique, on parle de masses).

Taille x_i	1.58	1.77	1.81	1.89	1.63	1.58	1.65	1.64	1.64	1.69	1.73	1.75
Poids y_i	55	74	72	73.5	53	58	61	65	58	74	64	72

Taille x_i	1.74	1.80	1.64	1.66	1.72	1.63	1.88	1.59	1.71	1.72	1.82	1.83
Poids y_i	74	64	55	65	61	50	73	48	80	80	81	84

1. Remplir le tableau de contingence suivant :

Taille $X \setminus$ Poids Y	[45, 55[[55, 65[[65, 75[[75, 85[
[1.50, 1.60[
[1.60, 1.70[
[1.70, 1.80[
[1.80, 1.90[

2. Les variables X et Y sont elles indépendantes ? Justifier.

Exercice 40

- Une étude sur le chômage a été faite et qui s'intéresse à l'ancienneté du chômage (X) moins de 24 mois, et l'âge (Y) entre 20 et 35 ans. Les résultats sont donnés par le tableau de contingence suivant :

$X \setminus Y$	[20, 25[[25, 30[[30, 35[
[0, 6[10	8	5
[6, 12[8	9	4
[12, 18[15	11	9
[18, 24[3	6	2

On donne

$$\sum_{i=1}^4 n_{i\bullet} x_i = 1014, \quad \sum_{i=1}^4 n_{i\bullet} x_i^2 = 14634, \quad \sum_{j=1}^3 n_{\bullet j} y_j = 2395, \quad \sum_{j=1}^3 n_{\bullet j} y_j^2 = 65062.5,$$

$$\sum_{i=1}^4 \sum_{j=1}^3 n_{ij} x_i y_j = 27074.5.$$

1. Quel est le nombre d'individus qui ont une ancienneté de chômage moins d'un an ?
2. Déterminer les deux distributions marginales.
3. Déterminer la distribution de X conditionnelle à $Y = [25, 30]$, c'est à dire, $X/Y = [25, 30]$.
4. Les variables X et Y sont elles indépendantes ? Justifier.
5. Donner la moyenne arithmétique.
6. Calculer le coefficient de corrélation linéaire. Commenter.
7. Donner l'équation de la droite de régression de Y en fonction de X .
8. Quel sera l'âge d'une personne ayant une ancienneté de chômage de 15 mois.

Exercice 41

- Une usine produit des pièces d'une machine. Pour chaque pièce (individu), on dispose du coût de sa production (DA) et du temps nécessaire pour sa réalisation (en heures). Le tableau ci-après (série statistique) donne cette répartition :

<i>Individu</i>	1	2	3	4	5
Temps (X) mesurée en heures	2	3	52	2	4
Coût (Y) mesurée en Dinars	10	16	23	12	18

1. Calculer la moyenne de la variable statistique X .

2. Calculer la moyenne de la variable statistique Y .
3. Calculer l'écart-type de la variable statistique X .
4. Calculer l'écart-type de la variable statistique Y .
5. Calculer la covariance des variables statistiques X et Y .
6. En supposant qu'il existe une corrélation linéaire entre X et Y , déterminer cette droite de corrélation.
7. Calculer le coefficient de corrélation. Conclusion ?
8. Une nouvelle pièce est réalisée en 6 heures. Estimer le coût de production de cette pièce en utilisant la droite de corrélation établie.

Chapitre 5

Annexe historique

¹L'histoire de la "statistique" remonte à une époque très ancienne. Les activités statistiques (dénombrements) ont commencé bien avant la création du mot, l'application de la méthode et de l'analyse statistique.

Depuis l'antiquité, les Empereurs, les Rois et les Hommes d'Eglise réalisaient des dénombrements de populations humaines et de terres pour les besoins de la guerre et de l'impôt.

Il y a plus de 4 ou 5000 ans, il existait déjà en Chine des descriptions chiffrées de la population et de l'agriculture.

Les Égyptiens de l'époque des Pharaons procédaient au dénombrement de la population ou du bétail.

A Rome, l'empereur Auguste fit procéder à une vaste enquête en dénombrant les soldats, les navires et les revenus publics.

Jusqu'au moyen âge, les seules "statistiques" existantes étaient les dénombrements faits dans des buts divers : assiettes de l'impôt, répartition des terres, recrutement dans l'armée est effectué avec des méthodes diverses (recensements des personnes, enregistrements de certains actes d'état civil ...).

C'est à partir du XVIII siècle, qu'apparaît le mot "statistique" crée par ACHENWAL en 1749 à partir du mot "STATISTA" (politique). Du simple dénombrement de populations humaines et de terres, la statistique est devenue une science qui a retenu et continue de retenir l'attention, non seulement des empereurs et de rois, mais surtout des personnes de sciences.

L'extension et l'utilisation du calcul des probabilités développé par J. BERNOULLI au 18ème siècle et l'application des études démographiques et sociales ont permis à cette

1. Source : B. Oukacha and M. Benmessaoud, Statistique descriptive et calcul des probabilités, 2013.

science de connaître un essor considérable. Ainsi au 19ème siècle, de la simple statistique descriptive, elle passe au stade de "Statistique Mathématique".

Depuis le 20ème siècle, les travaux de KARL PEARSON (1857-1936), de STUDENT (WILLIAM SEALY GOSSET, 1876-1937) et de RONAL FISHER (1890-1963) ont permis à cette science de connaître un développement considérable et une application vaste et variée. La statistique utilise les techniques et des méthodes de collecte, de présentation, d'étude et d'analyse des données quantitatives.

La statistique n'est pas uniquement utilisée pour décrire, pour mieux connaître un événement survenu dans le passé, mais elle intervient de plus en plus dans les travaux de planification, dans le choix de prises de décisions et d'actions.

Bibliographie

- [1] J. BLARD-LABORDERIE, L'essentiel des outils de statistique descriptive pour aborder des études en sciences humaines et sociales, 2015.
- [2] G. CALOT, Cours de statistique descriptive, Dunod, 1969.
- [3] G. CHAUVAT AND J.-P. REAU, Statistiques descriptives, Armand Colin, 2002.
- [4] M. TENENHAUS, statistique : Méthodes pour décrire, expliquer et prévoir, Dunod, 2006.
- [5] J.-J. DROESBEKE, Éléments de statistiques, Ellipses, 2001.
- [6] L. LEBOUCHER AND M.-J. VOISIN, Introduction à la statistique descriptive, 2013.
- [7] F. MAZEROLLE, Statistique descriptive, 2009.
- [8] B. OUKACHA AND M. BENMESSAOUD, Statistique descriptive et calcul des probabilités, 2013.
- [9] J.-Y. OUVRARD, Probabilités : Tome 1, 2001.
- [10] P. ROGER, Probabilités, statistique et processus stochastiques, Pearson Education,, 2004.
- [11] J. VAILLANT, Eléments de Statistique descriptive, 2015.