# MA272: Statistiques et Probabilités pour les étudiants de deuxième année en Chimie

Dr LOUMNGAM

Semestre 2, 2024-2025

## 0.1 Informations Générales

- **Intitulé du cours :** Statistiques et Probabilité appliquées à la Chimie
- **Niveau :** Licence 2 (Chimie)
- **Langue :** Français
- **Format :** Cours Magistraux (CM) et Travaux Dirigés (TD) en grand groupe (TP supprimés en raison du grand effectif)
- **Prérequis :** Algèbre, notions élémentaires en calcul différentiel et introduction à la chimie analytique.

## 0.2 Objectifs du Cours

- Comprendre les concepts fondamentaux de la statistique descriptive et de la théorie des probabilités.
- Appliquer les méthodes statistiques à l'analyse des données expérimentales en chimie.
- Développer la capacité à interpréter, quantifier et propager l'incertitude sur les résultats expérimentaux.
- Acquérir des bases solides en inférence statistique, incluant estimation, intervalles de confiance et tests d'hypothèses.
- Aborder l'analyse de régression et la corrélation dans une perspective d'optimisation des expériences chimiques.

## 0.3 Contenu et Programme

Le cours s'étend sur 12 semaines à raison de 3 heures par semaine.

### Module 0 : Introduction à la Statistique (Semaines 1)

- Introduction aux statistiques en chimie, importance des données expérimentales.
- Terminologie
- La démarche statistique

### Module 1 : Statistique Descriptive (Semaines 1 et 3)

- Introduction à la statistique descriptive.
- Organisation des données pour une et deux variables : tableaux, représentations graphiques.
- Mesures de tendance centrale : moyenne, médiane, mode.
- Mesures de dispersion : variance, écart-type, étendue, intervalle interquartile.
- Autres mesures : forme et concentration
- Régression et corrélation statistique

## Module 2 : Fondamentaux de la Probabilité (Semaines 4 à 6)

- Éléments de la théorie des ensembles et combinatoire
- Définitions et axiomes de la probabilité, espaces probabilisés.
- Probabilités conditionnelles, théorème de Bayes et ses applications en chimie analytique, indépendance des événements.
- Variables aléatoires discrètes : définition, fonction de masse de probabilité, espérance mathématique, variance.
- Distributions discrètes courantes : uniforme, Bernoulli, binomiale, géométrique, hypergéométrique, Poisson (et leurs applications en cinétique chimique et comptage d'événements).
- Variables aléatoires continues : définition, fonction de densité de probabilité, espérance mathématique, variance.
- Distribution normale (gaussienne) : propriétés, importance dans l'analyse des erreurs.
- Théorème central limite et son importance pour l'inférence statistique.
- Autres distributions continues importantes : t de Student, chi carré, F de Fisher (introduction).

## Module 3 : Inférence Statistique (Semaines 7 à 9)

- Échantillonnage, distributions d'échantillonnage, théorème central limite.
- Estimation paramétrique : estimateurs ponctuels et intervalles de confiance.
- Intervalles de confiance pour la moyenne et la variance.
- Tests d'hypothèses : formulation, erreurs de type I et II, puissance d'un test.
- Tests statistiques courants : test t de Student (pour une et deux moyennes), test du chi carré (pour l'indépendance et l'ajustement), ANOVA (introduction).

## Module 4 : Régression et Corrélation (Semaines 10 à 11)

- Régression linéaire simple : modèle, estimation des paramètres, interprétation
- Analyse de corrélation : coefficient de corrélation, coefficient de détermination et leurs interprétations.
- Analyse des résidus et évaluation de la qualité du modèle.
- Introduction à la régression multiple (si le temps le permet).

## Module 5 : Analyse des Erreurs et Conception Expérimentale (Semaines 12 à 13)

(si le temps le permet)

- Types d'erreurs : aléatoires et systématiques.
- Précision et exactitude.
- Analyse des erreurs expérimentales, propagation des incertitudes.
- Introduction aux plans d'expériences et optimisation expérimentale.
- Calcul des erreurs et présentation des résultats expérimentaux.

## 0.4 Méthodologie Pédagogique

- **Cours Magistraux (CM) :** Présentation théorique des concepts fondamentaux.
- **Travaux Dirigés (TD) :** Exercices pratiques et interprétation de résultats.

## 0.5 Modalités d'Évaluation

- **Contrôle Continu (30%) :** Exercices en classe et devoirs surveillés.
- **Examen Final (70%) :** Épreuve écrite de synthèse.

## 0.6 Bibliographie et Ressources

### Ouvrages de Référence


- Montgomery, D. C., *Applied Statistics and Probability for Engineers*.
- Sarig, *Introduction to Probability and Statistics : Part I - Probability*.
- Harris, *Exploring Chemical Analysis* (applications en chimie analytique).


### Ressources Complémentaires

- Supports en ligne via les plateformes universitaires (MIT OpenCourseWare, FUN-MOOC).
- Documents PDF de l'universités comme l'Université de Montpellier, Université Hébraïque de Jérusalem.

# Chapitre 1

# Introduction à la Statistique

## 1.1 Définition et domaines d'application de la statistique

### 1.1.1 Définition

La statistique est la science qui vise à collecter, traiter et analyser des données issues de l'observation de phénomènes aléatoires, c'est-à-dire où le hasard intervient. L'analyse des données permet de :

— Décrire les phénomènes étudiés

— Faire des prévisions

— Prendre des décisions à leur sujet

En cela, la statistique est un outil essentiel pour la compréhension et la gestion des phénomènes complexes.

### 1.1.2 Introduction aux Statistiques en Chimie

Les statistiques jouent un rôle central en chimie, notamment pour l'interprétation et la validation des données expérimentales. Dans un contexte où la précision et la fiabilité des mesures sont essentielles, une bonne maîtrise des outils statistiques permet d'évaluer l'incertitude associée aux expériences chimiques et de distinguer les tendances réelles du bruit expérimental. En chimie, elle vise plusieurs objectifs :

— Décrire les résultats expérimentaux à l'aide d'indicateurs numériques (moyenne, écart-type, médiane, etc.).

— Évaluer la dispersion des mesures et la qualité des données expérimentales.

— Modéliser les phénomènes chimiques à partir des données recueillies.

— Comparer différentes séries de mesures pour identifier des tendances ou des relations.

### 1.1.3 Rôle des Statistiques en Chimie

Les statistiques sont particulièrement utiles dans plusieurs branches de la chimie :

— **Chimie analytique** : Validation des méthodes analytiques par l'étude des erreurs et des incertitudes.

— **Chimie physique** : Analyse des relations entre variables et modélisation des comportements thermodynamiques ou cinétiques.

— **Chimie organique et inorganique** : Suivi des réactions chimiques et de leur reproductibilité.

— **Chimie environnementale** : Traitement des données issues des mesures de pollution et d'impact écologique.

### 1.1.4 Importance des Données Expérimentales

En chimie, les données expérimentales sont la base de toute avancée scientifique. Leur qualité et leur interprétation correcte sont primordiales pour garantir des résultats fiables et reproductibles.

#### 1.1.4.1 Types de Données Expérimentales

Les données expérimentales en chimie peuvent être classées en différentes catégories :

— **Données quantitatives** : Résultats mesurés numériquement (concentrations, températures, pressions, etc.).

— **Données qualitatives** : Observations non numériques (couleur d'un précipité, état physique d'un échantillon).

— **Données continues et discrètes** : Une grandeur mesurée peut prendre toutes les valeurs possibles dans un intervalle (continue) ou être limitée à des valeurs spécifiques (discrète).

— **Données univariées et multivariées** : Une seule variable mesurée (ex : température) ou plusieurs en interaction (ex : pression et volume dans une réaction gazeuse).

#### 1.1.4.2 Précision, Exactitude et Fiabilité des Mesures

L'interprétation correcte des résultats expérimentaux repose sur trois notions fondamentales :

— **Précision** : Capacité d'une série de mesures à donner des résultats proches les uns des autres.

— **Exactitude** : Capacité d'une mesure à se rapprocher de la valeur réelle.

— **Fiabilité** : Stabilité et reproductibilité des résultats dans des conditions identiques.

Les erreurs expérimentales peuvent être :

— **Erreurs systématiques** : Biais dans la méthode de mesure, par exemple un instrument mal calibré.

— **Erreurs aléatoires** : Variabilité inhérente aux conditions expérimentales, réduites par l'augmentation du nombre de mesures.

Il y a donc intervention du hasard et des probabilités. L'objectif essentiel de la statistique est de maîtriser au mieux cette incertitude pour extraire des informations utiles des données, par l'intermédiaire de l'analyse des variations dans les observations.

#### 1.1.4.3 Méthodes de Collecte et Traitement des Données

Le processus de collecte des données suit une méthodologie stricte :

1. **Définition des objectifs expérimentaux** : Identifier les variables à mesurer et les conditions expérimentales.
2. **Sélection des instruments de mesure** : Adapter les outils aux besoins spécifiques de l'expérience.
3. **Réplication des mesures** : Effectuer plusieurs mesures pour réduire les incertitudes.
4. **Analyse statistique** : Utilisation de méthodes statistiques (moyenne, écart-type, tests d'hypothèses) pour interpréter les données.

### 1.1.5 Application des Statistiques à l'Analyse des Données Expérimentales

L'analyse des données expérimentales repose sur plusieurs techniques statistiques :

— **Statistique descriptive** : Moyenne, médiane, mode, variance et écart-type pour résumer les données.

— **Régression et corrélation** : Étude des relations entre les variables expérimentales.

— **Tests statistiques** : Tests de normalité, tests d'hypothèses (t-test, ANOVA) pour valider les résultats.

### 1.1.6 Les deux branches de la statistique

— Statistique descriptive : vise à résumer l'information contenue dans les données de façon synthétique et efficace. Elle utilise des graphiques, tableaux et indicateurs numériques (moyennes, etc.) pour dégager les caractéristiques essentielles du phénomène étudié et suggérer des hypothèses pour des études ultérieures.

— Statistique inférentielle : va au-delà de la simple description des données. Elle vise à faire des prévisions et prendre des décisions au vu des observations. Cela nécessite de proposer des modèles probabilistes du phénomène et de gérer les risques d'erreurs.

### 1.1.7 Statistique et informatique

L'informatique et la statistique sont deux éléments complémentaires du traitement de l'information : l'informatique acquiert et traite l'information tandis que la statistique l'analyse. L'augmentation de la puissance des ordinateurs et et la facilité de transmission des données par internet ont rendu possible l'analyse de très grandes masses de données (big data). Les logiciels comme **Excel, R et Python** (bibliothèques NumPy et SciPy) sont souvent utilisés pour l'analyse des données expérimentales.

### 1.1.8 La science des données

La science des données ou data science désigne l'ensemble des méthodes permettant d'extraire des informations utiles de ces grandes masses de données et de les traiter pour résoudre des problèmes complexes dans divers domaines.

## 1.2 La démarche statistique

La statistique et les probabilités sont deux aspects complémentaires de l'étude des phénomènes aléatoires. Elles sont cependant de natures bien différentes.

### 1.2.1 Les probabilités

Les probabilités peuvent être envisagées comme une branche des mathématiques pures, basée sur la théorie de la mesure, abstraite et complètement déconnectée de la réalité.

### 1.2.2 Les probabilités appliquées

Les probabilités appliquées proposent des modèles probabilistes du déroulement de phénomènes aléatoires concrets. On peut alors, *avant toute expérience*, faire des prévisions sur ce qui va se produire.

#### 1.2.2.1 Exemple : Durée de vie d'une ampoule

Il est usuel de modéliser la durée de vie d'une ampoule électrique par une variable aléatoire $X$ de loi exponentielle de paramètre $\lambda$.

#### 1.2.2.2 Propriétés du modèle

— La probabilité que l'ampoule ne soit pas encore tombée en panne à la date $t$ est $P(X > t) = e^{-\lambda t}$.

— La durée de vie moyenne est $E[X] = \frac{1}{\lambda}$.

— Si $n$ ampoules identiques sont mises en fonctionnement en même temps, et qu'elles fonctionnent indépendamment les unes des autres, le nombre $N_t$ d'ampoules qui tomberont en panne avant un instant $t$ est une variable aléatoire de loi binomiale $B(n, P(X \leq t)) = B\left(n, 1 - e^{-\lambda t}\right)$.

#### 1.2.2.3 Intérêt du modèle

L'utilisateur de ces ampoules souhaite avoir une évaluation de leur durée de vie, de la probabilité qu'elles fonctionnent correctement pendant plus d'un mois, un an, etc.

#### 1.2.2.4 Limites du modèle

Pour utiliser les résultats théoriques, il faut :

1. *Choisir un bon modèle*, c'est-à-dire s'assurer que la durée de vie des ampoules est bien une variable aléatoire de loi exponentielle.
2. *Calculer la valeur du paramètre* $\lambda$.

#### 1.2.2.5 Rôle de la statistique

La statistique permet de résoudre ces problèmes en :

— *Effectuant une expérimentation*

— *Recueillant des données*

— *Les analysant*

## 1.3 Terminologie de base

Précisons le sens de certains termes fondamentaux pour une étude statistique

### 1.3.1 Unites statistique

**Definition 1** *On appelle Population ou population statistique ou univers statistique l'ensemble des personnes, d'objets ou des élements équivalents sur lesquels porte l'étude. On parle parfois de champs de l'étude.*

**Definition 2** *On appelle individu ou unité statistique tout élement de la population.*

**Definition 3** *On appelle échantillon, un ensemble d'élements tirés au hasard de la population sur lequel on éffectue une étude exhaustive pour ensuite porter certaines conclusions sur l'ensemble de la population. C'est simplement un sous ensemble de la population.*

**Definition 4** *La taille d'une population (resp d'un échantillon) est le cardinal (le nombre d'élement) de la population (resp de l'échantillon). Elle est généralement notée N (resp n.)*

**Definition 5** *Un recensement est une étude de tous les individus d'une population. Difficile en pratique lorsque les populations sont grandes pour des questions de coût et de temps.*

Le recensement est différent d'un sondage.

**Definition 6** *Un sondage est un recueil d'une partie de la population (échantillon).*

Le recueil d'un échantillon à partir de la population initiale se fait par des techniques statistiques, appelées *méthodes d'échantillonnage.*

### 1.3.2 Caractères

Chaque individus de la population peut être considéré selon un ou plusieurs caractères

**Definition 7** *Un caractère ou une variable statistique est un critère étudié dans la population. C'est une caractèristique relative à chacun des individus de la population et sur laquelle on veut faire porter l'étude.*

**Definition 8** *Une observation est le résultat de la mesure d'un caractère.*

Chacun des caractères étudiés peut présenter deux ou plusieurs modalités

**Definition 9** *Les modalités sont les différentes situations où les individus peuvent se trouver à l'égard du caractère considéré. Le nombre de modalités varie selon le niveau de détails de l'information disponible*

Les modalités d'un caractère sont à la fois incompatibles et exhaustive, c'est à dire un individus de la population ne doit posséder qu'une et une seule modalité.

### 1.3.3 Les types de caractères

On classe les caractères en deux catégories : le caractère qualitatif et le caractère quantitatif. Parmi ces derniers, on distingue les caractères quatitatifs discret et le caractère quantitatif continu.

**Definition 10** *Un caractère est dit qualitatif si ses diverses modalités ne sont pas mesurables. On parle également d'attributs ou de variables catégorielle.*

Dans la littérature, on fait la différence entre les variables qualitatives nominales et les variables qualitatives ordinales. Une variable qualitative ordinale est une variable sur lequel on peut y établir un ordre. Ce qui n'est pas le cas pour les variables qualitatives nominale.

**Definition 11** *Un caractère est dit quantitatif si ses différentes modalités sont mésurables ou répérables. Chaque modalité correspond à un nombre. ce nombre varie d'une modalité à une autre. Un caractère quantitatif est ausi appélé variable statistique.*

Une variable quantitative peut être mise sous forme qualitative ordinale en constituant des classes d'appartenance.

La création des amplitudes des classes est un problème délicat, qui nécessite un arbitrage entre information et simplification.

**Definition 12** *Une variable statistique est dite discrète lorsque ses valeurs possibles sont des nombres isolés. les cas les plus généralement rencontrés sont ceux où les valeurs possibles sont les entiers.*

**Definition 13** *Une variable statistique est dite continue lorsque ses valeurs possibles sont à priori en nombre infini et quelconques dans un intervalle de valeurs.*

### 1.3.4 Le caractère quantitatif continu

Les observations d'une variable statistique continue sont généralement regroupées en intervalles disjoints successifs et contigues (deux à deux disjoints) appélés classes. Le regroupement en classe permet de condenser les données et de les rendre plus commode à étudier.

**Definition 14** *On appelle extremités ou limites de la classe, les nombres entre lesquels sont comprises les valeurs rangées dans une classe. On la note souvent* $]e_{i-1}, e_i]$. $e_{i-1}$ *est l'extremité inferieure(ou initiale) de la classe* $]e_{i-1}, e_i]$. $e_i$ *est l'extremité supérieure (ou finale) de la classe* $]e_{i-1}, e_i]$.

**Remark 15** *Un individu de la population doit être dans une classe et une seule.*

On est souvent aménè à recalculer les limites réelles des classes surtout lorsque les extremités des classes ne sont pas contigus. Ces limites doivent être calculées de manière à conserver les centres des classes, les effectifs des classes et avoir des classes adjacentes.

**Definition 16** *La largeur de la classe ou la longueur de l'intervalle est l'amplitude de la classe* $a_i = e_i - e_{i-1}$

### 1.3.5 Série statistique

**Definition 17** *On appelle série ou distribution statistique associé à un caractère, l'ensemble des modalités du caractère avec en regard de chaque modalité les frequences correspondantes. Elle est généralement retranscrite dans un tableau de données.*

**Definition 18** *L'effectif ou la fréquence absolue d'une modalité est le nombre ou la proportion d'individus presentant cette modalité. Elle est généralement notée $x_i$ ou $n_i$.*

**Definition 19** *La fréquence relative d'une modalité est l'effectif ou la fréquence absolue rapportée à la taille de la population. Elle est généralement notée $f_i = \frac{n_i}{N}$ ou $f_i = \frac{x_i}{N}$.*

**Definition 20** *La proportion ou le porcentage d'une modalité est la fréquence relative exprimée en pourcentage.*

**Definition 21** *La hauteur ou la densité de fréquence de la classe est l'effectif de la classe rapportée à son amplitude : $h_i = \frac{n_i}{a_i}$.*

**Definition 22** *L'étendue d'une serie statistique d'un caractère quantitatif notée e est la différence entre la plus grande modalité et la plus petite modalité de la série.*

# EXERCICE

Pour chacun des exemples suivants d'études statistiques, identifier :

1. la population étudiée
2. l'échantillon prélevé pour effectuer cette étude (s'il y a lieu)
3. le caractère à l'étude
4. le type de caractère

## 1. Dosage d'un principe actif dans un médicament

Un laboratoire pharmaceutique souhaite connaître la concentration moyenne en principe actif d'un nouveau médicament. Pour cela, il prélève 30 comprimés au hasard dans un lot de production et mesure la teneur en principe actif de chaque comprimé.

## 2. Analyse de la composition d'un échantillon de gaz

Des chimistes analysent la composition d'un échantillon de gaz prélevé dans une cheminée d'usine. Ils mesurent la concentration en $CO_2$, en $SO_2$ et en NOx du gaz.

## 3. Détermination du point de fusion d'un composé

Un chimiste souhaite déterminer le point de fusion d'un nouveau composé qu'il a synthétisé. Il utilise un appareil de mesure du point de fusion et effectue plusieurs mesures sur une petite quantité de poudre du composé.

## 4. Étude de la résistance d'un matériau

Des ingénieurs testent la résistance d'un nouveau matériau pour la construction d'avions. Ils fabriquent 5 éprouvettes de forme et de dimensions identiques et soumettent chacune d'elles à un test de traction jusqu'à la rupture.

## 5. Comparaison de deux méthodes d'analyse

Des chercheurs comparent deux méthodes d'analyse d'un polluant dans l'eau. Ils analysent 20 solutions identiques par les deux méthodes et comparent les résultats obtenus.

## 6. Contrôle de la qualité d'un produit alimentaire

Un inspecteur de la DGCCRF contrôle la qualité d'un lot de yaourts. Il prélève 10 yaourts au hasard dans le lot et mesure la teneur en sucre de chaque yaourt.

## 7. Étude de l'impact d'un polluant sur la croissance des plantes

Des biologistes étudient l'impact d'un polluant sur la croissance des plantes. Ils divisent 40 plants de la même espèce en deux groupes : un groupe exposé au polluant et un groupe non exposé. Ils mesurent ensuite la hauteur de chaque plante après un certain temps.

## 8. Enquête sur les habitudes de consommation des étudiants en Chimie

Une association d'étudiants en Chimie souhaite connaître les habitudes de consommation de ses membres. Ils réalisent un sondage auprès de 100 étudiants et leur posent des questions sur leur consommation de produits bio, de produits locaux et de produits fairtrade.

## 9. Suivi de la production d'un médicament

Un laboratoire pharmaceutique souhaite suivre la production d'un médicament en continu. Il mesure la concentration en principe actif de chaque lot de production.

## 10. Contrôle de la qualité d'un process de fabrication

Un ingénieur souhaite contrôler la qualité d'un process de fabrication de pièces métalliques. Il mesure la dimension d'une pièce sur chaque lot de production.

# Chapitre 2

# Statistique Descriptive Unidimensionnelle et Bidimensionnelle

## 2.1 Introduction

La statistique descriptive, aussi appelée statistique exploratoire, constitue une étape fondamentale dans toute étude de données. Elle a pour vocation de **résumer et synthétiser** l'information contenue dans un ensemble de données, souvent issu d'un échantillon, en vue d'en faire ressortir les principales caractéristiques. Elle permet également de **formuler des hypothèses** sur la population d'origine, ce qui prépare le terrain à l'inférence statistique.

Dans ce cadre, on distingue classiquement deux niveaux d'analyse :

— **La statistique univariée**, qui s'intéresse à l'étude d'un seul caractère (quantitatif ou qualitatif) à la fois. Elle vise à décrire la répartition d'un caractère au sein d'une population à l'aide de tableaux de fréquences, de représentations graphiques (histogrammes, diagrammes en bâtons, boxplots, etc.) et d'indicateurs numériques (moyenne, médiane, variance, etc.).

— **La statistique bivariée**, qui s'attache à analyser la relation entre deux caractères statistiques observés simultanément sur une même population. Selon le type des variables (qualitatives ou quantitatives), on fera appel à des outils tels que les tableaux de contingence, les nuages de points, les coefficients de corrélation ou encore les ajustements linéaires (régression linéaire simple).

L'ensemble de ces méthodes constitue le socle de la statistique descriptive classique. À cela s'ajoutent des approches plus modernes, regroupées sous les termes d'*analyse des données* ou de *data mining*, qui visent à explorer des jeux de données complexes et multidimensionnels. Ces techniques comprennent :

— Les méthodes de **classification**, comme le partitionnement et la classification ascendante hiérarchique (CAH), qui consistent à regrouper les individus en classes homogènes selon leurs caractéristiques communes.

— Les méthodes d'**analyse factorielle**, telles que l'analyse en composantes principales (ACP) ou l'analyse factorielle des correspondances multiples (AFCM), qui permettent de représenter les données dans des espaces de dimension réduite tout en conservant l'essentiel de l'information.

Ce chapitre se focalisera dans un premier temps sur l'étude univariée, avant de développer les outils essentiels de l'analyse bivariée, en insistant sur leur interprétation dans un contexte expérimental ou applicatif, notamment en chimie et en sciences expérimentales.

## 2.2 Organisation des données d'une série statistique

Les outils utilisés pour organiser les données dépendent de la nature de la série statistiques (notamment le nombre de caractères étudiés) et de la nature des caractères (quantitatifs discrets, continues ou qualitatifs). On considère une variable statistique $X$, observée sur $k$ individus. On dispose alors d'une série statistique unidimensionnelle $x = (x_1, ..., x_k)$.

### 2.2.1 Tableaux pour une série unidimensionnelle

Les tableaux constituent un outil fondamental de la statistique descriptive unidimensionnelle. Ils permettent de présenter de manière ordonnée les données recueillies, facilitant ainsi leur lecture, leur interprétation et la construction d'indicateurs statistiques. Selon la nature du caractère étudié (qualitatif ou quantitatif, discret ou continu), différentes formes de tableaux peuvent être utilisées.

#### 2.2.1.1 Caractère qualitatif

Lorsque la variable statistique $X$ est un caractère *qualitatif*, les données sont organisées dans un tableau de fréquences de la forme :

| Modalité du caractère | $x_1$ | $x_2$ | ... | $x_k$ | Total |
|---|---|---|---|---|---|
| Effectif correspondant | $n_1$ | $n_2$ | ... | $n_k$ | $n$ |

Ce type de tableau permet d'associer à chaque modalité (valeur possible du caractère) l'effectif observé, facilitant ainsi une première synthèse des données.

#### 2.2.1.2 Caractère quantitatif discret

Lorsque $X$ est un caractère *quantitatif discret*, les tableaux utilisés conservent la même structure générale, mais les modalités correspondent alors aux différentes valeurs numériques que peut prendre la variable.

**Tableau des fréquences**

| Valeur de la variable | $x_1$ | $x_2$ | ... | $x_k$ | Total |
|---|---|---|---|---|---|
| Effectif | $n_1$ | $n_2$ | ... | $n_k$ | $n$ |

**Tableaux des fréquences cumulées** Pour analyser la répartition des données selon un ordre croissant ou décroissant, on utilise les tableaux de fréquences cumulées.

**Cumul croissant (par valeurs inférieures)**

| Modalité | $x_1$ | $x_2$ | ... | $x_k$ | |
|---|---|---|---|---|---|
| Fréquence cumulée croissante | $F_1$ | $F_2$ | ... | $F_k$ | 1 |

où $F_i = \frac{1}{n} \sum_{j=1}^{i-1} n_j$ est la fréquence cumulée des valeurs inférieures à $x_i$.

**Cumul décroissant (par valeurs supérieures)**

| Modalité | $x_1$ | $x_2$ | ... | $x_k$ | |
|---|---|---|---|---|---|
| Fréquence cumulée décroissante | $F_1'$ | $F_2'$ | ... | $F_k'$ | 1 |

où $F_i' = \frac{1}{n}\sum_{j=i}^{k} n_j$ est la fréquence cumulée des valeurs supérieures ou égales à $x_i$.

**Remark 23** *Voici quelques remarques importante sur les formes cumulées :*

1. *Les formes cumulées croissantes peuvent se calculer tant pour les effectifs, les frequences relatives ou les pourcentages.*
2. *Les formes cumulées décroissante se déduisent des formes cumulées croissante de la manière suivante : $F_i^{'} = N - F_i$ si on à faire à des effectifs, $F_i^{'} = 1 - F_i$ si on à faire à des fréquences, $F_i^{'} = 100 - F_i$ si on à faire à des pourcentages.*
3. *Les fréquences cumulées dans le tableau sont placées sur les lignes et non à l'interieur des cases corespondantes aux modalités.*

#### 2.2.1.3 Caractère quantitatifs continu

**a) Tableau de fréquences** Lorsque la variable statistique est continue, les modalités du carctère sont les classes des valeurs possibles définies par les extremités des classes. Sa représentation est la suivante :

| X : Classes | $[e_1, e_2[$ | $[e_2, e_3[$ | ... | $[e_i, e_{i+1}[$ | ... | $[e_k, e_{k+1}[$ | Total |
|---|---|---|---|---|---|---|---|
| Fréquence absolue : $n_i$ | $n_1$ | $n_2$ | ... | $n_i$ | ... | $n_k$ | $N$ |

**Remark 24** *Lorsque le nombre d'individu de la série discrète est élevé comme dans l'exemple des 79 étudiants, on peut que également regroupé les modalités en classes, ce regroupement implique nécessairement une perte d'information et donc est délicat à opérer afin de minimiser cette perte. Dans le cas oú l'ensemble des observations est assez symétrique et pas trop dispersé, la* ***formule de Spurge*** *suite permet de déterminer le nombre de classes k si le nombre d'observations N :*

$$k = 1 + log_2(N)$$

**b) Tableau de fréquences cumulées** La fréquence cumulée croissante (resp décroissante) correspondante à la classe $[e_{i-1}, e_i[$ de la variable statistique $X$ notée $F_i$ (resp $F_i^{'}$ ) se calcule de la manière suivante : $F_i = \sum_{j=1}^{i-1} n_j$ (resp $F_i^{'} = \sum_{j=i}^{k} n_j$ ). $F_i$ (resp $F_i^{'}$ ) représente le nombre d'individus dont les modalités du caractère sont inférieures ou égales (resp. supérieures) à $e_{i-1}$. On la représente dans un tableau qui à la forme suivante :

| X : Classes | $[e_1, e_2[$ | $[e_2, e_3[$ | ... | $[e_i, e_{i+1}[$ | ... | $[e_k, e_{k+1}[$ | Total |
|---|---|---|---|---|---|---|---|
| Effectifs : $n_i$ | $n_1$ | $n_2$ | ... | $n_i$ | ... | $n_k$ | $N$ |
| Eff. cum. crois. | $0 = F_1$ | $F_2 = n_1$ | ... | $F_i$ | ... | $F_k = N - n_k$ | $F_{k+1} = N$ |
| Eff. cum. décrois. | $N = F_1^{'}$ | $F_2^{'} = N - n_1$ | ... | $F_i^{'}$ | ... | $F_k^{'} = n_k$ | $F_{k+1}^{'} = 0$ |

Ces tableaux permettent d'évaluer la concentration des données dans certaines zones de l'échelle des valeurs et sont particulièrement utiles pour construire des courbes de distribution et des graphiques de type polygone des fréquences ou ogives.

## 2.2.2 Tableaux pour une série bidimensionnelle

On s'interresse à présent à l'étude de deux variables X et Y étudiées sur la même population. Lorsque l'on utilisera les données regroupées en classes, les modalités $x_i$ seront remplacées par les centres de classes.

### 2.2.2.1 Données non groupées

Il s'agit de la donnée de la serie statistique brute sous la forme $(x_i, y_i)$ des modalités des variables $X$ et $Y$ pour chaque individu. Ces données sont généralement representé dans le tableau suivant :

| $i$ | 1 | 2 | ... | $i$ | ... | $n$ |
|---|---|---|---|---|---|---|
| $X$ | $x_1$ | $x_2$ | ... | $x_i$ | ... | $x_n$ |
| $Y$ | $y_1$ | $y_2$ | ... | $y_i$ | ... | $y_n$ |

### 2.2.2.2 Données groupées

C'est le cas le plus rencontrée en pratique. Considérons $X_1, X_2, ..., X_I$ et $Y_1, Y_2, ..., Y_J$ les modalités des variables $X$ et $Y$. Soit $n_{ij}$ l'effectif de la population qui présente à la fois la modalité $X_i$ de $X$ et $Y_j$ de $Y$. Ces données sont souvent representées dans un tableau à double entrées appélé tableau croisé ou tableau de contingence. Ces tableaux ont la forme suivantes :

| | $Y_1$ | $Y_2$ | ... | $Y_j$ | ... | $Y_J$ | Total |
|---|---|---|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1J}$ | $n_{1.}$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2J}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $X_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{iJ}$ | $n_{i.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $X_I$ | $n_{I1}$ | $n_{I2}$ | ... | $n_{Ij}$ | ... | $n_{IJ}$ | $n_{I.}$ |
| Total | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.J}$ | $n_{..}$ |

Où : $n_{.j} = \sum_{i=1}^{I} n_{ij}; n_{i.} = \sum_{j=1}^{J} n_{ij}; n_{..} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}$

**Remark 25** *En divisant toute les valeurs du tableau par l'effectif total, on obtient le tableau de frequences.*

## 2.2.3 Représentations graphiques des séries unidimensionnelles

Bien qu'un tableau statistique renferme toute l'information qu'on a rassemblée, il est très souvent utile de traduire par un graphique pour en réaliser une synthèse visuelle.

### 2.2.3.1 Caractères qualitatifs

On représente habituellement les distributions selon un caractère qualitatifs de quatre façons différentes : Le diagramme en batons, la representation en tuyaux d'orgue, la representation par secteur angulaire et la representation par secteur rectangulaire.

**a) Diagramme en colonnes ou en bâtons** On porte sur un axe d'abscisse les modalités du caractères et on lève à partir de chaque modalité un segment qui est proportionnel à la fréquence de la modalité. Lorsqu'on joint les sommets de ces batons, on obtient le polygone des effectifs ou des frequences. lorsqu'on lisse ce polygone, on obtient la courbe des effectifs ou des fréquences.

**Exemple 26** *Considérons une étude portant sur les religions pratiquées par une population. Le diagramme en bâtons correspondant est donné par la figure 2.1 ci-dessous.*

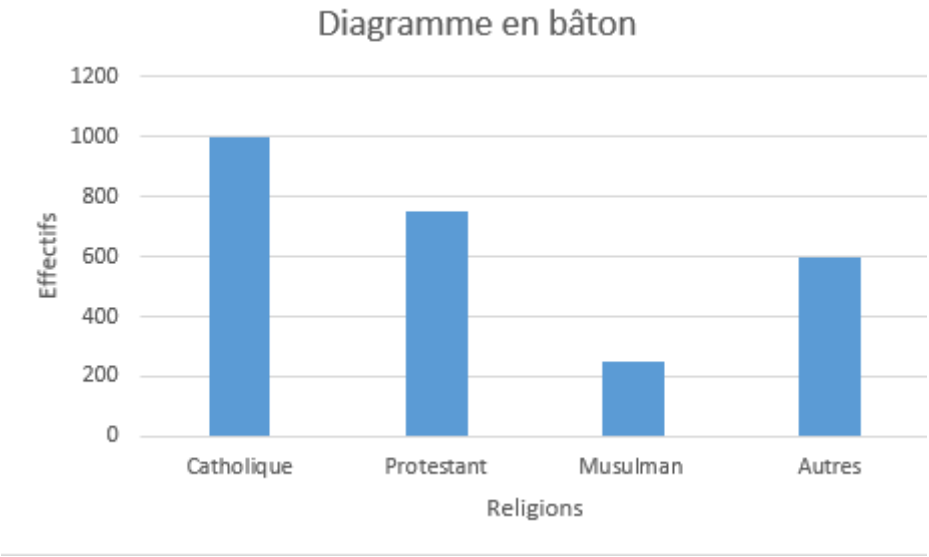


FIGURE 2.1 –

**c) Diagrammes sectoriels ou camemberts.** La representation en secteurs angulaires consiste à partager le cercle en secteurs angulaires où ce dernier représente une modalité et son angle ou la surface est propotionnelle à la fréquence de la modalité. La règle pour déterminer l'angle est $\alpha_i = \frac{360 \times n_i}{N}$ où $n_i$ est l'effectif de la modalité i. $N$ la taille de la population.

**Exemple 27** *Reprenons l'exemple portant sur les religions. Un exemple de diagramme en secteurs circulaires correspondant est donné par la figure 2.2 ci-dessous.*

#### 2.2.3.2 Caractères quantitatifs discrets

**a) Diagramme différentiel en bâtons** Lorsque la variable statistique est quantitative, on utilise des représentations graphiques similaires à celles employées pour les variables qualitatives, en s'appuyant sur les fréquences absolues ou relatives. Toutefois, une distinction essentielle réside dans la nature des modalités : les variables quantitatives possèdent un ordre naturel, puisqu'elles prennent des valeurs numériques réelles, contrairement aux variables qualitatives pour lesquelles aucun ordre intrinsèque n'est défini. Cette différence justifie l'usage privilégié des diagrammes en bâtons pour les variables quantitatives, tandis que les diagrammes circulaires (ou sectoriels) sont rarement utilisés, car ils ne permettent pas de restituer l'ordre des valeurs.

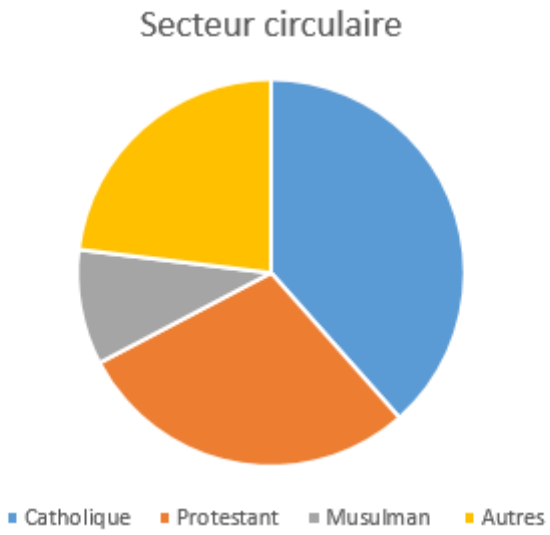


FIGURE 2.2 –

**Exemple 28** *Nous donnons les notes de* 79 *étudiants à un examen universitaire.*

68 ;84 ;75 ;82 ;68 ;90 ;62 ;88 ;76 ;93 ;73 ;79 ;88 ;73 ;60 ;63 ;93 ;71 ;59 ;85 ;75 ;
61 ;65 ;75 ;87 ;74 ;62 ;95 ;78 ;63 ;72 ;76 ;66 ;78 ;82 ;75 ;94 ;77 ;69 ;74 ;68 ;60 ;96 ;
78 ;89 ;61 ;75 ;75 ;60 ;79 ;83 ;71 ;79 ;62 ;67 ;96 ;78 ;85 ;76 ;65 ;71 ;75 ;85 ;65 ;70 ;
73 ;57 ;88 ;78 ;62 ;76 ;53 ;74 ;86 ;67 ;73 ;81 ;72 ;77.

*Un exemple de diagramme différentiel correspondant est donné par la figure 2.3 ci-dessous.*

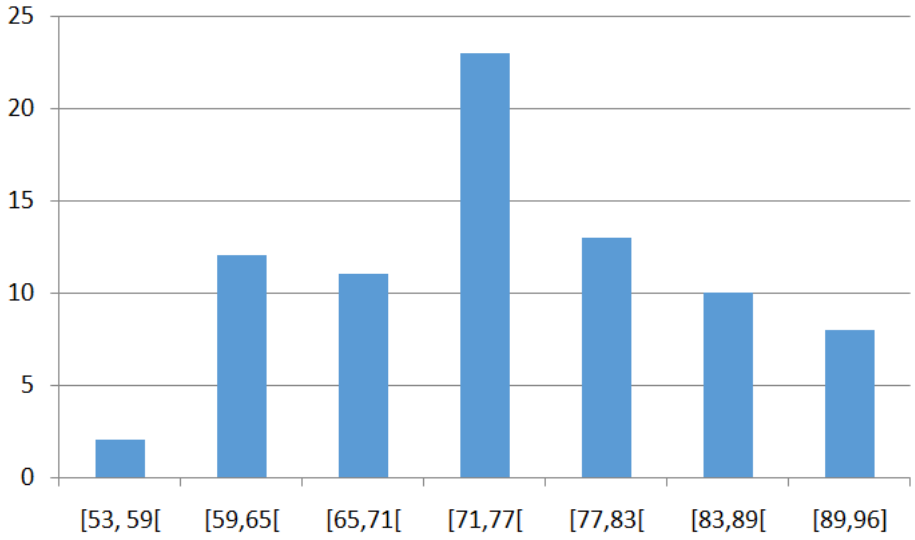


FIGURE 2.3 –

**b) Diagramme intégral**

**Definition 29** *On appelle fonction cumulative ou fonction de répartition de la distribution $X$, la fonction $F(x)$ qui représente la proportion des individus de la population dont le caractère est inférieur à $x$. Le diagramme intégral d'une distribution statistique est la représentation graphique de cette fonction définie pour toute valeur $x$ réelle comme suit :*

*Pour $x \leq x_1, F(x) = F_1 = 0, x_1$ est la première modalité de la distribution.*

*Pour $x_{i-1} < x \leq x_i, F(x) = F_i, i = 2, 3, ..., k$*

*Pour $x > x_k$, $F(x) = F_{k+1} = N$, où $N$ est l'effectif total.*

On voit bien que $F$ est une fonction en escalier. Lorsqu'on joint les points de coordonnées $(x_i, F_i)$, on obtient le polygone des effectifs ou des frequences cumulées croissants. lorsqu'on lisse ce polygone, on obtient la courbe des effectifs ou des fréquences cumulées croissants ou la courbe cumulative.

**Exemple 30** *Reprenons l'exemple introductif portant sur les notes de 79 étudiants. Un exemple de diagramme différentiel correspondant est donné par la figure 2.4 ci-dessous.*

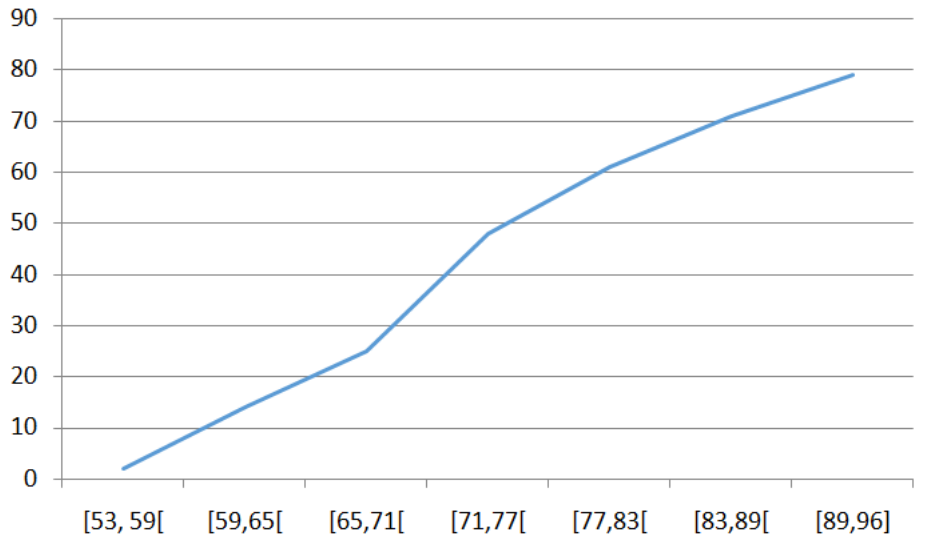


FIGURE 2.4 –

#### 2.2.3.3 Caractères quantitatifs continus

Une variable continue prend ses valeurs dans un ensemble non dénombrable, tel que $\mathbb{R}$ ou un intervalle fermé $[a, b]$. Dans ce cas, les représentations graphiques de type diagramme en bâtons perdent leur pertinence, car les observations sont généralement toutes distinctes, ce qui implique que les fréquences absolues associées à chaque valeur sont souvent égales à 1. Une telle représentation n'apporte donc aucune information synthétique utile sur la distribution des données.

Dans le cas de variables continues, nous allons étudier les représentation à l'aide d'**histogramme** et de **diagramme intégral**.

Avant de construire certaines représentations graphiques adaptées aux variables continues, il est nécessaire d'**ordonner les données**. En effet, la construction d'un histogramme ou d'une fonction de répartition empirique repose sur cette étape préalable.

Soit un échantillon de $n$ valeurs noté $(x_1, x_2, \ldots, x_n)$. On appelle *échantillon ordonné* la suite $(x_1^*, x_2^*, \ldots, x_n^*)$ obtenue après classement croissant des observations :

$$x_1^* \leq x_2^* \leq \ldots \leq x_n^*.$$

**Exemple : durée de vie d'ampoules** Considérons un échantillon de 10 ampoules dont les durées de vie (en heures) ont été mesurées comme suit :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

L'échantillon ordonné correspondant est :

5.4 9.5 24.3 35.7 57.1 67.3 91.6 118.4 170.9 251.3

On peut ainsi illustrer quelques notations utiles :

— $x_1 = 91.6$ correspond à la durée de vie de la première ampoule mesurée.

— $x_1^* = \min(x_1, \ldots, x_{10}) = 5.4$ est la plus petite durée de vie observée dans l'échantillon.

L'ordonnancement des données permet notamment de :

— construire des classes pour l'histogramme,

— calculer des fréquences cumulées,

— définir la fonction de répartition empirique.

#### 2.2.3.4 Histogramme

L'histogramme est une représentation graphique utilisée pour regrouper les observations proches en classes. Pour cela, on choisit une borne inférieure $a_0$ (souvent légèrement inférieure à $x_1^*$) et une borne supérieure $a_k$ (supérieure à $x_n^*$), puis on partitionne l'intervalle $]a_0, a_k]$ en $k$ sous-intervalles adjacents $]a_{j-1}, a_j]$ appelés **classes**. La largeur de la classe $j$ est notée $h_j = a_j - a_{j-1}$.

— Si toutes les classes sont de même largeur $h = \dfrac{a_k - a_0}{k}$, on parle d'**histogramme à pas fixe**.

— Si les classes ont des largeurs différentes, on parle d'**histogramme à pas variable**.

Le **nombre d'observations** dans la classe $j$ est noté $n_j$, c'est-à-dire :

$$n_j = \sum_{i=1}^{n} \mathbf{1}_{]a_{j-1}, a_j]}(x_i).$$

La **fréquence** (ou fréquence relative) associée à cette classe est $f_j = \dfrac{n_j}{n}$.

**Definition 31** *L'histogramme est la figure formée par des rectangles dont les bases sont les classes $]a_{j-1}, a_j]$, et dont les aires sont égales aux fréquences correspondantes. Autrement dit, la hauteur du $j^e$ rectangle est :*

$$\hat{f}_j = \frac{n_j}{n h_j}.$$

On peut interpréter cette hauteur $\hat{f}_j$ comme une approximation de la **densité de probabilité** sur la classe $]a_{j-1}, a_j]$. En effet, la fréquence associée à cette classe s'écrit comme une aire :

$$f_j = \frac{n_j}{n} = \int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx,$$

où $\hat{f}$ est une fonction en escalier, constante sur chaque classe, et valant $\hat{f}_j = \dfrac{n_j}{nh_j}$ sur la classe $]a_{j-1}, a_j]$.

Ainsi, l'histogramme fournit une estimation de la densité de la variable étudiée. Cette estimation permet d'approcher graphiquement la fonction de densité $f$ de la loi de la variable $X$, et d'en analyser la forme. L'allure de l'histogramme peut être comparée à celle de densités connues (loi normale, loi exponentielle, etc.), ce qui facilite la modélisation probabiliste.

**Remarque :** La construction d'un histogramme dépend de plusieurs paramètres :

— les bornes $a_0$ et $a_k$ choisies pour l'intervalle d'étude ;

— le nombre $k$ de classes retenues ;

— la largeur des classes, qui peut être fixe ou variable.

Des choix différents peuvent conduire à des histogrammes très différents pour un même ensemble de données, et donc à des interprétations potentiellement trompeuses.

**Recommandation :** Pour construire un histogramme pertinent, il est conseillé de suivre certaines règles pratiques (ex. : règle de Sturges, règle de Scott ou règle de Freedman–Diaconis), qui permettent de déterminer un nombre de classes adapté à la taille de l'échantillon.

**Choix du nombre de classes et des bornes** Il est recommandé de choisir un nombre de classes $k$ compris entre 5 et 20. La **règle de Sturges** propose une formule simple pour estimer ce nombre :

$$k \approx 1 + \log_2 n = 1 + \ln n / \ln 2.$$

Par exemple :

$$\begin{cases} k = 5 & \text{si } n \leq 22, \\ k = 6 & \text{si } 23 \leq n \leq 45, \quad \text{etc.} \end{cases}$$

Le choix des bornes $a_0$ et $a_k$ de l'intervalle d'étude doit garantir une certaine homogénéité dans la largeur des classes. Une méthode fréquemment utilisée consiste à fixer :

$$a_0 = x_1^* - 0{,}025(x_n^* - x_1^*) \quad \text{et} \quad a_k = x_n^* + 0{,}025(x_n^* - x_1^*).$$

Dans le cas d'un histogramme à pas fixe, les classes sont de même largeur $h = (a_k - a_0)/k$, et la hauteur de chaque rectangle est proportionnelle à l'effectif de la classe concernée.

**Exemple : histogramme des durées de vie d'ampoules** Considérons un échantillon de $n = 10$ ampoules. En appliquant la règle de Sturges, on choisit $k = 5$ classes. Les valeurs extrêmes de l'échantillon sont $x_1^* = 5.4$ et $x_n^* = 251.3$.

En appliquant la méthode des bornes élargies :

$$a_0 = 5.4 - 0{,}025(251.3 - 5.4) \approx -0.747 \quad \text{et} \quad a_5 = 251.3 + 0{,}025(251.3 - 5.4) \approx 257.4.$$

Par commodité, on arrondit à $a_0 = 0$ et $a_5 = 260$, ce qui donne une largeur de classe :

$$h = \frac{260 - 0}{5} = 52.$$

On obtient alors la répartition suivante :

| **Classes** $]a_{j-1}, a_j]$ | $]0, 52]$ | $]52, 104]$ | $]104, 156]$ | $]156, 208]$ | $]208, 260]$ |
|---|---|---|---|---|---|
| Effectifs $n_j$ | 4 | 3 | 1 | 1 | 1 |
| Fréquences $n_j/n$ | 40% | 30% | 10% | 10% | 10% |
| Hauteurs $n_j/(nh)$ | 0.0077 | 0.0058 | 0.0019 | 0.0019 | 0.0019 |

**Tableau 2.4** — Répartition des durées de vie d'ampoules en classes de largeur égale.

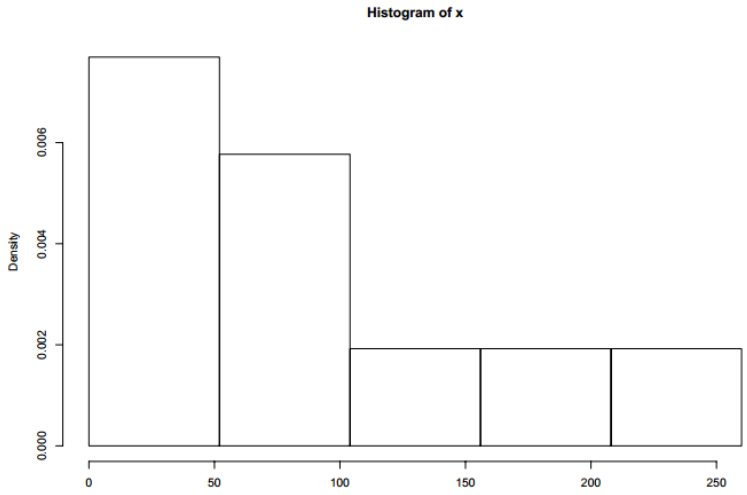


FIGURE 2.5 –

**Interprétation de l'histogramme** Le **mode de l'histogramme** est estimé comme le centre de la classe correspondant à la plus grande hauteur. Dans l'exemple précédent, cette classe est $]0, 52]$, dont le milieu est 26. Cela donne une approximation du point où la densité est maximale, que l'on appelle aussi *mode de la densité*.

L'histogramme fournit une bonne visualisation de la répartition des données. Dans le cas des durées de vie des ampoules, on remarque une forte concentration d'observations pour les petites valeurs, et une raréfaction des observations à mesure que la durée augmente. Autrement dit, la densité de la variable aléatoire représentant la durée de vie est ici **décroissante**.

**Limites des histogrammes à pas fixe** Un inconvénient majeur de l'histogramme à pas fixe est qu'il ne garantit pas une répartition homogène des observations. Certaines classes peuvent être très chargées tandis que d'autres sont quasi vides. Par exemple, dans l'exemple des ampoules :

— la première classe ($]0, 52]$) contient 4 observations,

— alors que les trois dernières classes ne regroupent qu'une observation chacune.

Cela peut nuire à la lisibilité et à l'interprétation de l'histogramme. Pour pallier ce problème, on peut :

— **scinder les classes trop chargées** en sous-classes plus fines,

— **regrouper les classes peu remplies** pour gagner en lisibilité.

Une solution consiste à adopter un histogramme à **effectifs égaux**, dans lequel chaque classe contient approximativement le même nombre d'observations. Dans ce cas, les bornes des classes ne sont plus également espacées, mais déterminées de façon à équilibrer les effectifs. Elles deviennent donc dépendantes de la distribution des données observées, ce qui rend l'analyse plus robuste dans certains cas.

**Histogramme à effectifs égaux** Une alternative à l'histogramme à pas fixe consiste à constituer des classes contenant un **nombre égal d'observations**. Ce type de découpage permet de mieux répartir les données sur l'axe horizontal, notamment lorsque la densité varie fortement.

Dans l'exemple des ampoules, on choisit de répartir les 10 observations en 5 classes contenant chacune 2 observations. On peut déterminer les bornes des classes en prenant les milieux des valeurs de l'échantillon ordonné. Cela donne l'intervalle suivant :

$$\text{Classes :} \quad ]0, 17] \quad ]17, 46] \quad ]46, 79] \quad ]79, 145] \quad ]145, 260]$$

On obtient alors le tableau suivant :

| **Classes** $]a_{j-1}, a_j]$ | $]0, 17]$ | $]17, 46]$ | $]46, 79]$ | $]79, 145]$ | $]145, 260]$ |
|---|---|---|---|---|---|
| Largeurs $h_j$ | 17 | 29 | 33 | 66 | 115 |
| Effectifs $n_j$ | 2 | 2 | 2 | 2 | 2 |
| Fréquences $n_j/n$ | 20% | 20% | 20% | 20% | 20% |
| Hauteurs $n_j/(nh_j)$ | 0.0118 | 0.0069 | 0.0061 | 0.0030 | 0.0017 |

**Tableau 2.5** — Répartition des durées de vie d'ampoules en classes de même effectif.

Ce type de représentation, bien que les classes ne soient plus de même largeur, permet de visualiser la densité relative des observations sur différents intervalles. Plus la classe est étroite, plus la concentration locale d'observations est forte.

**Remarque :** dans l'histogramme à effectifs égaux, les bornes sont ajustées en fonction des données. L'uniformité des effectifs rend la comparaison des hauteurs de rectangles directement liée à la concentration locale d'observations.

**Bilan sur les types d'histogrammes** On observe que l'histogramme à classes de même effectif permet une description plus fine de la distribution des données, notamment lorsque la densité n'est pas uniforme. Toutefois, ce type d'histogramme est moins fréquemment utilisé que celui à classes de même largeur, car il est généralement plus difficile à tracer manuellement.

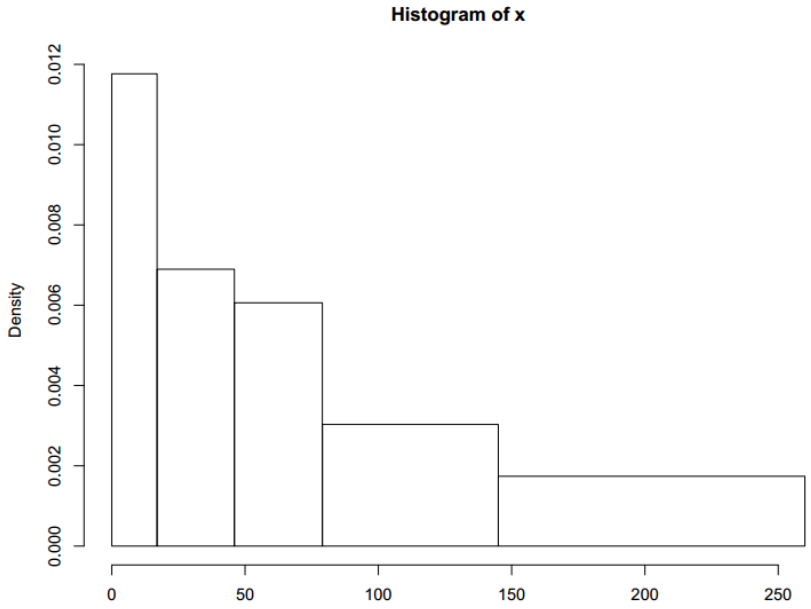


FIGURE 2.6 –

Il est important de noter que deux histogrammes construits à partir des mêmes données, mais avec des choix de classes différents, peuvent donner des représentations visuelles sensiblement différentes. Cela implique qu'il faut rester prudent lorsqu'on souhaite utiliser un histogramme pour estimer la densité d'une variable : l'histogramme fournit avant tout une **allure générale** de la distribution, et non une estimation rigoureuse.

Dans notre exemple, la forme des deux histogrammes obtenus évoque une densité décroissante, similaire à celle d'une loi exponentielle de densité $f(x) = \lambda e^{-\lambda x}$. En revanche, ces histogrammes sont très éloignés de l'allure d'une densité normale (en forme de cloche). Cela suggère que la durée de vie des ampoules ne suit probablement pas une loi normale, mais pourrait raisonnablement être modélisée par une loi exponentielle. Ce constat reste toutefois qualitatif, et nécessitera d'être confirmé par des outils statistiques plus rigoureux.

**Remark 32** *Si, au lieu des effectifs $n_j$, on considère les* ***effectifs cumulés***

$$m_j = \sum_{i=1}^{j} n_i,$$

*on construit un* ***histogramme cumulé****, qui permet d'obtenir une estimation graphique de la fonction de répartition de la variable étudiée.*

#### 2.2.3.5 Fonction de répartition empirique

**Definition 33** *La* ***fonction de répartition empirique*** *(FdRE), notée $F_n$, est associée à un échantillon $x_1, x_2, \ldots, x_n$ de valeurs réelles. Elle est définie pour tout $x \in \mathbb{R}$ comme le pourcentage d'observations inférieures ou égales à $x$ :*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}}.$$

*Si l'on considère les données classées par ordre croissant* $x_1^* \leq x_2^* \leq \ldots \leq x_n^*$*, la fonction* $F_n$ *s'écrit également :*

$$F_n(x) = \begin{cases} 0 & si\ x < x_1^* \\ \frac{i}{n} & si\ x_i^* \leq x < x_{i+1}^*, \quad 1 \leq i \leq n-1 \\ 1 & si\ x \geq x_n^* \end{cases}$$

La fonction de répartition empirique $F_n(x)$ est donc une **fonction en escalier croissante**, qui effectue des sauts de hauteur $1/n$ à chaque donnée de l'échantillon. Elle constitue une estimation naturelle de la **fonction de répartition théorique** $F(x) = P(X \leq x)$ de la variable aléatoire $X$.

Cette estimation est particulièrement robuste et converge rapidement vers $F(x)$ lorsque la taille de l'échantillon augmente. Elle est largement utilisée pour :

— visualiser la distribution d'un échantillon,

— comparer empiriquement deux échantillons,

— effectuer des tests d'ajustement (comme le test de Kolmogorov-Smirnov).

**Exemple :** Dans l'exemple des durées de vie d'ampoules, la figure 2.7 représente graphiquement la fonction $F_n$ associée à l'échantillon. Cette fonction présente 10 sauts successifs de hauteur 1/10, chacun correspondant à une des valeurs ordonnées.

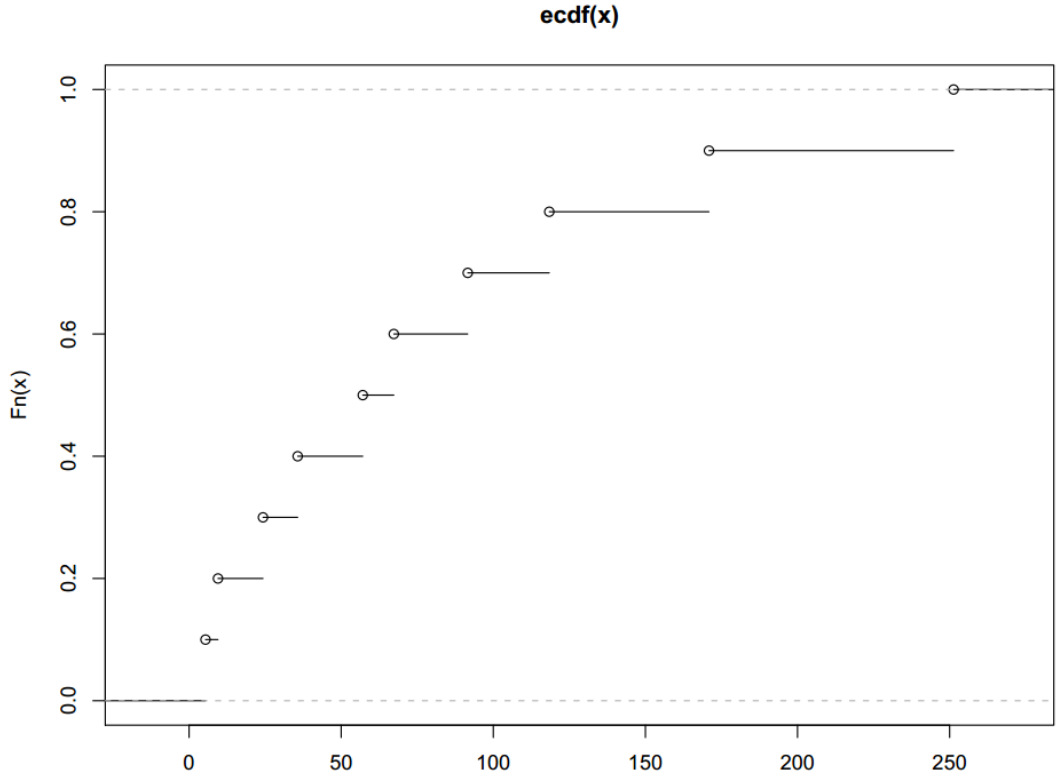


FIGURE 2.7 –

#### 2.2.3.6 Les graphes de probabilités

La fonction de répartition empirique est un outil très utile en statistique. Dans cette section, nous nous intéressons à son application dans le but de déterminer si un modèle probabiliste donné est compatible avec les observations issues d'un échantillon.

Une première idée consiste à tracer le graphe de la fonction de répartition empirique $F_n$ et à le comparer visuellement à celui de la fonction de répartition théorique $F$ d'une loi connue (par exemple la loi normale, exponentielle, etc.). Toutefois, cette méthode est peu efficace, car les fonctions de répartition de nombreuses lois probabilistes sont visuellement très proches. À l'œil nu, il est souvent difficile de distinguer la fonction de répartition d'une loi normale de celle d'une loi exponentielle, alors même que leurs densités sont très différentes.

Une alternative plus robuste consiste à transformer la fonction de répartition empirique de façon à faire apparaître les caractéristiques spécifiques d'une loi. C'est le principe du **graphe de probabilités** (ou *probability plot*, aussi appelé *Q-Q plot*). Ce graphe consiste en un nuage de points construit à partir de la fonction $F_n$ de manière que :

> *si les données suivent bien une certaine loi de probabilité, alors les points du nuage devraient s'aligner approximativement sur une droite.*

Soit $F$ la fonction de répartition théorique d'une loi dépendant d'un paramètre $\theta$, et $F_n$ la fonction empirique obtenue à partir des données. L'idée est de chercher une transformation affine de la forme :

$$h[F(x)] = \alpha(\theta)g(x) + \beta(\theta),$$

où $h$ et $g$ sont des fonctions connues et $\alpha(\theta)$, $\beta(\theta)$ des paramètres dépendant de $\theta$.

Ainsi, si la véritable fonction de répartition des données est $F$, alors pour tout $x$, on devrait avoir :

$$h[F_n(x)] \approx \alpha(\theta)g(x) + \beta(\theta).$$

Pour les données ordonnées $x_1^*, \ldots, x_n^*$, la fonction de répartition empirique prend les valeurs $F_n(x_i^*) = \dfrac{i}{n}$. En appliquant la fonction $h$ à ces valeurs, on obtient :

$$h\left(F_n(x_i^*)\right) = h\left(\frac{i}{n}\right).$$

En reportant les couples de points $\left(g(x_i^*), h\left(\frac{i}{n}\right)\right)$ sur un graphique, on obtient le **graphe de probabilité**. Si les données suivent bien la loi associée à $F$, alors ces points s'alignent approximativement sur une droite. La pente et l'ordonnée à l'origine de cette droite permettent d'estimer les paramètres $\alpha(\theta)$ et $\beta(\theta)$, et donc indirectement le paramètre $\theta$ de la loi supposée.

**Conclusion :** le graphe de probabilités est une méthode puissante pour tester l'adéquation d'un modèle probabiliste à un échantillon, notamment en amont de tests statistiques formels.

**Definition 34** *Soit $F$ la fonction de répartition d'une loi de probabilité, dépendant d'un paramètre inconnu $\theta$. S'il existe des fonctions $h$, $g$, ainsi que des fonctions $\alpha$ et $\beta$ dépendant de $\theta$, telles que :*

$$\forall x \in \mathbb{R}, \quad h[F(x)] = \alpha(\theta)g(x) + \beta(\theta),$$

*alors le nuage de points*

$$\left(g(x_i^*),\ h\left(\frac{i}{n}\right)\right), \quad i \in \{1, \ldots, n\}$$

*est appelé* ***graphe de probabilités*** *(ou Q-Q plot) pour la loi de fonction de répartition $F$.*

*Si les points de ce nuage sont approximativement alignés, on considérera que la fonction $F$ constitue un modèle de répartition plausible pour les observations.*

**Exercice :** Tracer le graphe de probabilités pour la loi exponentielle pour l'exemple des ampoules.

### 2.2.4 Représentations graphiques des séries bidimensionnelles

Dans cette partie, on s'interressera aux données quantitatives.

#### 2.2.4.1 Données non groupées

On représente dans un répère orthonormé les points de coordonnées $(x_i, y_i)$. L'ensemble de ces points forme le nuage de points. Le nombre de ces points est égale au nombre d'individus s'il n'y a pas superposition. Pour l'exemple 139 a) ci-dessus, nous obtenons le nuage de points de la figure 5.1.

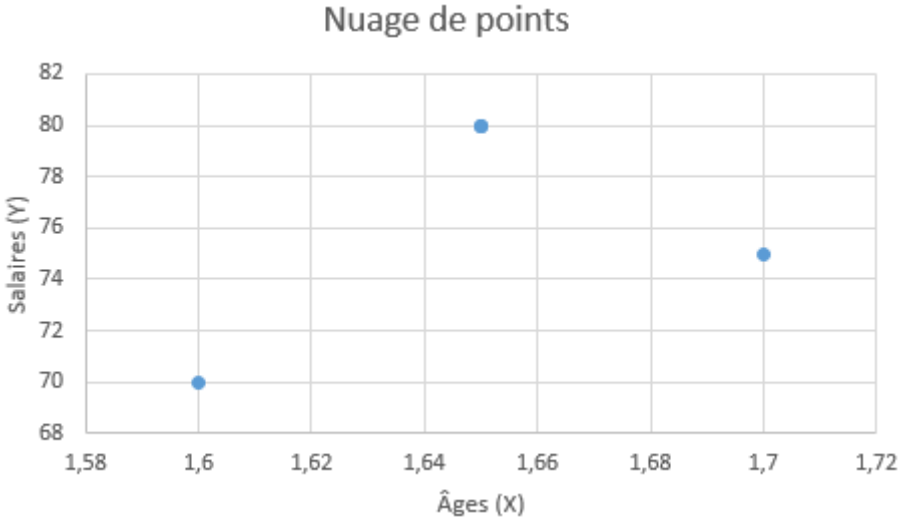


FIGURE 2.8 –

#### 2.2.4.2 Données groupées

On considère ici le tableau de contingence. On représente dans un répère orthonormé les disques aux points de coordonnées $(X_i, Y_j)$ dont la surface ou le rayon est proportionnelle aux effectifs. Le nombre de disque est égale à $I \times J$. La représentation graphique correspondant à l'exemple 140 a) ci-dessus est donnée par la figure 5.2.

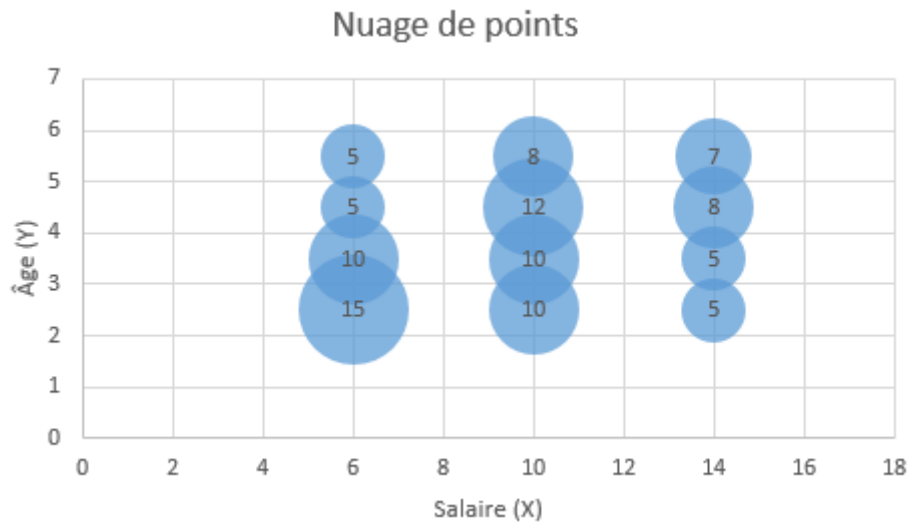


FIGURE 2.9 –

### 2.2.5 Mesures de tendance centrale

#### 2.2.5.1 Le mode ou la dominante

**Definition 35** *Le mode est la valeur ou la modalité qui présente la plus grande fréquence dans une distribution. Il est généralement noté $M_0$. Lorsqu'une distribution présente un seul mode (resp. deux, trois ou plusieurs), on parle de distribution unimodale (resp. bimodale, trimodale ou multimodale).*

Dans le cas d'un caractère quantitatif continu avec regroupement en classes, on parle plutôt de *classe modale* : celle qui présente la plus grande densité (ou hauteur). Le mode est alors estimé par le centre de cette classe.

#### 2.2.5.2 La médiane

**Definition 36** *La médiane d'une variable statistique est la valeur qui partage une population, classée par ordre croissant, en deux groupes d'effectifs égaux. Elle est notée $M_e$ et correspond à la valeur pour laquelle la fréquence cumulée atteint 0,5 : $F(M_e) = 0{,}5$.*

**Détermination pour des données non groupées :**

— Si $n$ est impair : $M_e = x_{\frac{n+1}{2}}$.

— Si $n$ est pair : $M_e = \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$.

**Détermination pour des données groupées :**

**(a) Caractère discret**

— Si $\dfrac{N}{2}$ correspond exactement à une valeur dans les effectifs cumulés, alors l'intervalle médian est $[x_G, x_D]$ et $M_e = \dfrac{x_G + x_D}{2}$.

— Sinon, la médiane est la modalité dont l'effectif cumulé encadre $\frac{N}{2}$.

**(b) Caractère continu**

— Si $\frac{N}{2}$ tombe exactement sur un total cumulé entre deux classes $[a, b]$ et $[c, d]$, alors $M_e = \frac{b+c}{2}$.

— Sinon, on utilise l'interpolation linéaire dans la classe $[e_{i-1}, e_i[$ qui contient la médiane :

$$M_e = e_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{n_i}(e_i - e_{i-1}) = e_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{n_i}a_i$$

**Remark 37** *Lorsque les fréquences sont relatives, on remplace* $\frac{N}{2}$ *par* 0,5.

#### 2.2.5.3 Autres quantiles : quartiles, déciles, centiles

**Quartiles**

**Definition 38** *Les quartiles divisent la population en quatre parties égales. On note* $Q_1$, $Q_2$ *(la médiane), et* $Q_3$. *Le* $k$*-ième quartile correspond à la valeur pour laquelle* $F(Q_k) = \frac{kN}{4}$.

**Déciles et centiles**

**Definition 39** — *Les déciles divisent la population en 10 groupes égaux :* $d_1, \ldots, d_9$.

— *Les centiles divisent la population en 100 groupes égaux :* $c_1, \ldots, c_{99}$.

*La valeur* $d_k$ *(resp.* $c_k$*) est définie par* $F(d_k) = \frac{kN}{10}$ *(resp.* $F(c_k) = \frac{kN}{100}$*).*

La méthode de calcul est similaire à celle de la médiane.

#### 2.2.5.4 La moyenne arithmétique

**Definition 40** *La moyenne arithmétique, notée* $\overline{x}$*, représente le centre de gravité d'une distribution.*

— Pour des données non groupées : $\overline{x} = \frac{1}{N}\sum n_i x_i = \sum f_i x_i$

— Pour des données groupées : remplacer $x_i$ par les centres de classe $c_i$.

**Astuce de calcul par changement de variable :** Si $x'_i = \frac{x_i - b}{a}$, alors :

$$\overline{x'} = \frac{1}{N}\sum n_i x'_i \quad \text{et} \quad \overline{x} = a\overline{x'} + b$$

#### 2.2.5.5 La $\varphi$-moyenne

**Definition 41** *La $\varphi$-moyenne est une généralisation de la moyenne arithmétique. Elle est définie par :*

$$\overline{x}_{\varphi} = \frac{1}{N}\sum n_i \varphi(x_i)$$

Cas particuliers :

— $\varphi(x) = x \Rightarrow$ moyenne arithmétique

— $\varphi(x) = x^2 \Rightarrow$ moyenne quadratique : $\overline{x}_Q$

— $\varphi(x) = \dfrac{1}{x} \Rightarrow$ moyenne harmonique : $\overline{x}_H$

— $\varphi(x) = \log x \Rightarrow$ moyenne géométrique : $\overline{x}_G = (\prod x_i^{n_i})^{1/N}$

**Remark 42** — $\overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq \overline{x}_Q$

— *La moyenne s'exprime dans les mêmes unités que les données.*

— *La moyenne arithmétique est égale à l'espérance mathématique.*

### 2.2.6 Mesures de dispersion

#### 2.2.6.1 L'étendue

**Definition 43** *L'étendue, notée $E$, est la différence entre la plus grande et la plus petite valeur d'une distribution. Elle mesure l'amplitude totale des données.*

— Pour une variable discrète non groupée : $E = x_{\max} - x_{\min}$.

— Pour une variable groupée : $E = e_k - e_1$, où $[e_i, e_{i+1}[$ sont les bornes des classes.

#### 2.2.6.2 L'écart interquartile et le semi-interquartile

**Definition 44** *L'écart interquartile, noté $E_Q$, mesure l'étendue des 50% centrales des observations. Il est défini par :*

$$E_Q = Q_3 - Q_1$$

*où $Q_1$ et $Q_3$ sont respectivement le premier et le troisième quartile.*

**Definition 45** *Le semi-interquartile est la moitié de l'écart interquartile. Il donne une mesure robuste de la dispersion autour de la médiane :*

$$SIQ = \frac{E_Q}{2}$$

#### 2.2.6.3 L'écart moyen et l'écart médiant

**Definition 46** *L'écart moyen, noté $\overline{E}$, est la moyenne des écarts absolus entre les observations et la moyenne :*

$$\overline{E} = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - \overline{x}|$$

*Pour les données groupées, on remplace $x_i$ par les centres des classes.*

**Definition 47** *L'écart médiant, noté $E_M$, est la moyenne des écarts absolus entre les observations et la médiane :*

$$E_M = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - M_e|$$

#### 2.2.6.4 La variance et l'écart-type

**Definition 48** *La variance, notée $V$, mesure la dispersion moyenne des données autour de la moyenne :*

$$V = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$$

Formule alternative (formule de König) :

$$V = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2$$

**Definition 49** *L'écart-type, noté $\sigma$, est la racine carrée de la variance :*

$$\sigma = \sqrt{V}$$

L'écart-type est la mesure de dispersion la plus utilisée. En cas de changement de variable $x_i' = \frac{x_i - b}{a}$, on a :

$$\sigma = |a| \cdot \sigma'$$

**Remark 50** *Pour les données groupées, les $x_i$ sont remplacés par les centres des classes $c_i$.*

#### 2.2.6.5 Le score et le coefficient de variation

**Definition 51** *Le score (ou note standardisée) d'un individu, noté $s_i$, mesure le nombre d'écarts-types qui le séparent de la moyenne :*

$$s_i = \frac{x_i - \overline{x}}{\sigma}$$

**Definition 52** *Le coefficient de variation, noté $C_v$, évalue la dispersion relative par rapport à la moyenne :*

$$C_v = \frac{\sigma}{\overline{x}}$$

Il permet de comparer la variabilité de deux distributions même si elles n'ont pas la même unité ou échelle.

#### 2.2.6.6 Les moments

**Definition 53** *Le moment non centré d'ordre $r$ est la moyenne des puissances $r$-ième des valeurs :*

$$m_r = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^r$$

**Definition 54** *Le moment centré d'ordre $r$ par rapport à une valeur $x_0$ est :*

$$\mathcal{M}_{r,x_0} = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - x_0)^r$$

**Remark 55** *Cas particuliers :*

- $r = 0$ *:* $m_0 = 1$, $\mathcal{M}_{0,x_0} = 1$
- $r = 1$ *:* $m_1 = \overline{x}$, $\mathcal{M}_{1,\overline{x}} = 0$
- $r = 2$ *:* $\mathcal{M}_{2,\overline{x}} = V$

## 2.2.7 Mesures de forme

### 2.2.7.1 Le coefficient d'asymétrie de Fisher

**Definition 56** *Le coefficient d'asymétrie de Fisher permet d'apprécier le sens et l'intensité de l'asymétrie d'une distribution par rapport à sa moyenne. Il est noté $\gamma_1$ et défini par :*

$$\gamma_1 = \frac{\mathcal{M}_{3,\overline{x}}}{\sigma^3}$$

*où $\mathcal{M}_{3,\overline{x}}$ est le moment centré d'ordre 3 et $\sigma$ l'écart-type.*

- Si $\gamma_1 = 0$, la distribution est parfaitement symétrique.
- Si $\gamma_1 > 0$, la distribution est asymétrique à droite (queue à droite).
- Si $\gamma_1 < 0$, la distribution est asymétrique à gauche (queue à gauche).

Une alternative graphique consiste à considérer la formule :

$$d = \frac{Q_1 + Q_3 - 2M_e}{2M_e}$$

où $Q_1$, $Q_3$ sont les quartiles et $M_e$ la médiane. Ce coefficient fournit une approximation simple de l'asymétrie.

### 2.2.7.2 Le coefficient d'aplatissement (kurtosis)

**Definition 57** *Le coefficient d'aplatissement de Fisher permet d'évaluer le niveau de concentration des valeurs autour de la moyenne. Il est noté $\gamma_2$ et défini par :*

$$\gamma_2 = \frac{\mathcal{M}_{4,\overline{x}}}{\sigma^4} - 3$$

*où $\mathcal{M}_{4,\overline{x}}$ est le moment centré d'ordre 4.*

- Si $\gamma_2 = 0$, la distribution a le même aplatissement que la loi normale (courbe de Gauss).
- Si $\gamma_2 > 0$, la distribution est plus effilée (leptokurtique).
- Si $\gamma_2 < 0$, la distribution est plus aplatie (platykurtique).

Ces deux coefficients ($\gamma_1$ et $\gamma_2$) permettent une caractérisation fine de la forme de la distribution, complémentaire aux mesures de tendance centrale et de dispersion.

## 2.2.8 Mesures de concentration

### 2.2.8.1 La médiale

Soit $X$ une variable statistique à valeurs positives, avec des modalités $x_i$ et des effectifs associés $n_i$.

**Definition 58** *On appelle **masse** associée à la modalité $x_i$ le produit $m_i = n_i x_i$.*

La **masse totale** est donnée par $M = \sum_{i=1}^{k} m_i$, et la **masse cumulée croissante** est $M_i = \sum_{j=1}^{i} m_j$.

**Definition 59** *La **médiale**, notée $M_d$, est la valeur de la variable qui partage la masse totale $M$ en deux parts égales. Autrement dit, c'est la valeur pour laquelle $M_i = \frac{M}{2}$.*

Sa détermination suit une méthode analogue à celle de la médiane, en utilisant les masses cumulées plutôt que les effectifs cumulés.

#### 2.2.8.2 Courbe de concentration et indice de Gini

**Definition 60** *La **courbe de concentration** (ou courbe de Lorenz) permet de visualiser la répartition d'une variable positive parmi une population. Elle est définie par l'ensemble des points :*

$$\left(P_i = \frac{F_i}{N}, \quad Q_i = \frac{M_i}{M}\right),$$

*où $F_i$ est l'effectif cumulé jusqu'à $x_i$, $M_i$ la masse cumulée, $N$ l'effectif total, et $M$ la masse totale.*

Plus la courbe de Lorenz est proche de la diagonale, plus la répartition est égalitaire. Plus elle est éloignée, plus la concentration est forte (inégalités accrues).

**Definition 61** *L'**indice de Gini**, noté $g$, mesure l'inégalité de la répartition. Il est défini par :*

$$g = 1 - \sum_{i=1}^{k-1} (P_{i+1} - P_i)(Q_{i+1} + Q_i)$$

*Il correspond au double de l'aire entre la courbe de Lorenz et la diagonale.*

- $g = 0$ : distribution parfaitement égalitaire.
- $g = 1$ : concentration totale (un seul individu détient toute la masse).

## 2.3 Régression et corrélation statistique

### 2.3.1 Statistiques conditionnelles et distributions croisées

On considère un tableau de contingence croisant deux variables statistiques $X$ et $Y$, où $n_{ij}$ représente l'effectif des individus présentant simultanément la modalité $X_i$ et la modalité $Y_j$.

**Distribution de $X$ conditionnellement à $Y = Y_j$ :** La distribution de $X$ sachant $Y_j$ est la série $(X_i, n_{ij})$ pour $1 \leq i \leq I$. Les statistiques conditionnelles associées sont :

$$\overline{x}_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{I} n_{ij} x_i, \qquad \sigma_j^2 = \frac{1}{n_{\cdot j}} \sum_{i=1}^{I} n_{ij} (x_i - \overline{x}_j)^2$$

**Distribution de $Y$ conditionnellement à $X = X_i$ :** La distribution de $Y$ sachant $X_i$ est la série $(Y_j, n_{ij})$ pour $1 \leq j \leq J$. Les statistiques associées sont :

$$\overline{y}_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{J} n_{ij} y_j, \qquad \sigma_i^2 = \frac{1}{n_{i\cdot}} \sum_{j=1}^{J} n_{ij} (y_j - \overline{y}_i)^2$$

**Lien entre les statistiques conditionnelles et globales :** La moyenne marginale s'exprime comme une moyenne pondérée des moyennes conditionnelles :

$$\overline{x} = \sum_{j=1}^{J} p_j \overline{x}_j, \qquad \overline{y} = \sum_{i=1}^{I} p_i \overline{y}_i$$

La variance marginale se décompose en :

$$\sigma_X^2 = \sum_{j=1}^{J} p_j \sigma_j^2 + \sum_{j=1}^{J} p_j (\overline{x}_j - \overline{x})^2$$

Le premier terme est la variance intra-groupe (résiduelle), le second est la variance inter-groupe (expliquée).

### 2.3.2 Covariance et corrélation linéaire

**Covariance :**

$$\mathrm{Cov}(X,Y) = \frac{1}{n_{..}} \sum_{i,j} n_{ij} x_i y_j - \overline{x} \cdot \overline{y}$$

**Coefficient de corrélation linéaire :**

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \quad \text{où } -1 \leq \rho \leq 1$$

— $\rho = 1$ : corrélation linéaire parfaite positive

— $\rho = -1$ : corrélation linéaire parfaite négative

— $\rho = 0$ : absence de corrélation linéaire

### 2.3.3 Droites de régression et estimation linéaire

**Droite de régression de $Y$ sur $X$ :**

$$Y - \overline{Y} = a_X (X - \overline{X}), \quad \text{où } a_X = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} = \rho \cdot \frac{\sigma_Y}{\sigma_X}$$

**Droite de régression de $X$ sur $Y$ :**

$$X - \overline{X} = a_Y (Y - \overline{Y}), \quad \text{où } a_Y = \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}$$

Les deux droites se coupent au point moyen $(\overline{x}, \overline{y})$. Plus $|\rho|$ est proche de 1, plus l'alignement est marqué.

### 2.3.4 Rapport de corrélation

**Définition :** Le rapport de corrélation mesure la proportion de la variance expliquée par la variable indépendante :

$$\eta^2_{Y/X} = \frac{\sum n_{i\cdot} (\overline{y}_i - \overline{y})^2}{n_{..} \cdot \sigma_Y^2} = 1 - \frac{\sum n_{i\cdot} \sigma_i^2}{n_{..} \cdot \sigma_Y^2}$$

— $\eta^2 \in [0,1]$, plus $\eta^2$ est proche de 1, plus la liaison est forte.

— Le rapport est invariant par changement d'origine et d'unité.

### 2.3.5 Indépendance

Les variables $X$ et $Y$ sont dites indépendantes si toutes les moyennes conditionnelles $\overline{x}_j = \overline{x}$ et $\overline{y}_i = \overline{y}$. Dans ce cas :

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}}$$

Les courbes de régression sont alors parallèles aux axes.

# Chapitre 3

# Fondamentaux de la Probabilité

## 3.1 Éléments de la théorie des ensembles et dénombrement

### 3.1.1 Rappel sur les ensembles

La théorie des probabilité est efficacement modélisée à l'aide des concepts de la théorie des ensemble. Un *ensemble* est une "collection" d'objets bien définis appelées *éléments* de cet ensemble. Ces éléments peuvent être énumérés entre accolades ou définis à l'aide d'une propriété qui les caractérise tous. Un ensemble est généralement noté par une lettre majuscule de l'alphabet latin et ses éléments en lettres miniscules. Si $x$ est un élément de $E$, on note $x \in E$. Sinon, on note $x \notin E$.

Dénombrer un ensemble $E$ c'est compter ses éléments et en déterminer éventuellement le nombre. Ce nombre est appelé le *Cardinal* de $E$ et est noté $|E|$ ou $Card(E)$.

**Exemple 62** $Card(\emptyset) = 0$

**Definition 63** *Soit $E$ un ensemble. On dit qu'un ensemble $A$ est un sous-ensemble (ou une partie) de $E$ si tout élément de $A$ est un élément de $E$. Lorsque tel est le cas, on note $A \subseteq E$. On dit aussi que $A$ est inclus dans $E$.*

Si $A$ n'est pas inclus dans $E$, on note $A \not\subseteq E$, et cela veut dire que même si certains éléments de $A$ peuvent appartenir à $E$, il existe au moins un élément de $A$ qui n'est pas dans $E$. Pour tout ensemble $E$, on a toujours $\emptyset \subseteq E$ et $E \subseteq E$.

**Definition 64** *On définit l'ensemble $\mathcal{P}(E)$ lire ("P de E") par $\mathcal{P}(E) = \{A | A \subseteq E\}$ comme étant l'ensemble des parties de $E$.*

Soit $A$ et $B$ deux ensembles, et $A_1$, $A_2$, ..., $A_n$, ... une famille d'ensembles. On a :

**Definition 65** *(Intersection d'ensembles) $A \cap B$ se lit "A inter B" et désigne l'ensemble des éléments appartenant aussi bien à $A$ qu'à $B$. $A_1 \cap A_2 \cap ... \cap A_n$ noté encore $\bigcap_{i=1}^{n} A_i$ est l'ensemble des éléments appartenant à chaque $A_i$.*

Deux ensembles $A$ et $B$ sont dits *disjoints* si leur intersection est vide, i.e. ; si $A \cap B = \emptyset$. Dans ce cas on a : $|A \cap B| = |A| + |B|$.

**Definition 66** *(Réunion d'ensembles) $A \cup B$ se lit "A inter B" et désigne l'ensemble des éléments appartenant à moins un des ensembles $A$ ou $B$. $A_1 \cup A_2 \cup ... \cup A_n$ noté encore $\bigcup_{i=1}^{n} A_i$ est l'ensemble des éléments appartenant à au moins un des $A_i$.*

**Proposition 67** *Soient A,B et C trois ensembles, on a :*

1. $|A \cup B| = |A| + |B| - |A \cap B|$
2. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
3. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

**Definition 68** *(Partition d'un ensemble) On dit qu'une famille $A_1$, $A_2$,...,$A_n$ de parties d'un ensemble $E$ est une partition de $E$ si on a :*

— *pour tout $i = 1, ..., n$, $A_i \neq \emptyset$,*

— *pour tous $i, j = 1, ..., n$ tels que $i \neq j$, on a $A_i \cap A_j \neq \emptyset$,*

— *et $\cup_{i=1}^{n} A_i = E$.*

**Exemple 69** *L'ensemble des entiers naturels pairs et l'ensemble des entiers naturels impairs constituent une partition de $\mathbb{N}$.*

**Proposition 70** *Si $A_1$, $A_2$, ..., $A_n$ forment une partition d'un ensemble fini $E$, alors on a : $|A_1| + |A_2| + ... + |A_n| = |E|$*

**Definition 71** *(Complémentaire et différence) . Si $A$ et $B$ sont deux parties de $E$, on définit la différence de $A$ et $B$, notée $A - B$ comme étant l'ensemble des éléments de $A$ qui n'appartienent pas à $B$. $E - A$ est appelé complémentaire de $A$ dans $E$ est noté $A^c$ ou $\overline{A}$.*

**Proposition 72** *Soit $A$ et $B$ deux parties de $E$, on a :*

1. *$A \cap A^c = \emptyset$, $A \cup A^c = E$, et $(A^c)^c = A$.*
2. *$(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$, et cela s'étend à une famille $A_1$, $A_2$,...,$A_n$ d'ensembles de la même manière.*

**Definition 73** *(Produit cartésien d'ensembles) On définit le produit cartésien de $A$ et $B$, noté $A \times B$ (lire "A croix B"), l'ensemble des couples (x, y) avec $x$ dans $A$ et $y$ dans $B$ respectivement. $A_1 \times A_2 \times ... \times A_n$ noté encore $\Pi_{i=1}^{n} A_i$ est l'ensemble des n-listes $(x_1, x_2, ..., x_n)$ où les éléments $x_i$ appartenant respectivement à $A_i$.*

**Proposition 74** *Si $A_1$, $A_2$, ..., $A_n$ une famille d'ensembles finis, alors on a :*

$$|A_1 \times A_2 \times ... \times A_n| = |A_1||A_2|...|A_n|$$

## 3.1.2 Rappels sur le dénombrement

Une partie importante du calcul des probabilités repose sur le dénombrement d'ensembles. Cette section présente les principes de base du dénombrement.

### 3.1.2.1 Principe fondamentale

Si $E_1$, $E_2$, ...,$E_k$ sont des ensembles finis non vides ayant chacun $m_1$, $m_2$, ...,$m_k$ objets respectivement, alors il y a $m_1 \times m_2 \times ůů \times m_k$ manières de choisir, d'abord un objet dans $E_1$, ensuite un objet dans $E_2$, puis un objet dans $E_3$, ...,et finalement un objet dans $E_k$.

**Exemple 75** *Supposons qu'un étudiant dispose de* 3 *chemises,* 4 *pantalons et de* 2 *paires de chaussures. Alors il existe* $3 \times 4 \times 2 = 24$ *tenues différentes qu'ils peut arborer.*

### 3.1.2.2 Arrangements

**Arrangements avec répétition**

**Definition 76** *Soit E un ensemble non vide à n objets, et p un entier naturel quelconque. On appelle arrangement avec répétition de p objets de E, toute liste de p objets choisis dans l'ordre avec répétition éventuellement parmi les n.*

**Proposition 77** *Le nombre d'arrangements avec répétition de p objets choisis parmi n objets est* $n^p$.

**Arrangements sans répétition**

**Definition 78** *Soit E un ensemble non vide à n objets, et p un entier naturel quelconque. On appelle arrangement sans répétition de p objets de E ou simplement arrangement, toute liste de p objets choisis dans l'ordre sans répétition parmi les n. Ici il faut que* $p \leq n$.

**Proposition 79** *Le nombre d'arrangements sans répétition de p objets choisis parmi n objets est*

$$A_n^p = n \times (n-1) \times (n-2) \times ... \times (n-p+1)$$

Pour entier naturel $n$, $A_n^n$ est le nombre de *permutations* de $n$ objets, il est noté $n!$ (lire factoriel $n$). Par convention $0! = 1$ et on vérifie facilement que $A_n^p = \frac{n!}{(n-p)!}$.

**Permutations avec répétition**

**Definition 80** *Une permutation avec répétition (anagramme) de p objets distincts est une liste de longueur n constitué des p objets dans un ordre donnée, avec éventuellement des répétions.*

**Proposition 81** *En supposant que l'objet* 1 *apparaît* $k_1$*, l'objet* 2 $k_2$ *fois,..., et l'objet p* $k_p$ *fois, le nombre de permutations avec répétition n objets est*

$$\frac{n!}{k_1!k_2!...k_p!}$$

#### 3.1.2.3 Combinaisons

**Combinaisons sans remise**

**Definition 82** *Soit $E$ un ensemble non vide à $n$ objets, et $p$ un entier naturel quelconque. On appelle combinaison sans répétition de $p$ objets de $E$ ou simplement combinaison, tout sous-ensemble de $p$ éléments choisis parmi les $n$ sans remise. On suppose alors que $p \leq n$.*

**Proposition 83** *Le nombre de combinaisons sans remise de $p$ objets choisis parmi $n$ objets est*

$$C_n^p = \frac{A_n^p}{p!}$$

**Combinaisons avec remise**

**Definition 84** *Soit $E$ un ensemble à $n$ objets. On appelle combinaison avec remise de $p$ objets de $E$, tout groupement non-ordonné de $p$ objets de $E$, chaque objet pouvant y figurer plusieurs fois.*

**Proposition 85** *Le nombre de combinaisons avec remise de $p$ objets choisis parmi $n$ objets est*

$$C_{n+p-1}^p$$

#### 3.1.2.4 Tableau récapitulatif

| Tirage de $p$ boules dans $n$ | ordoné/ dissernable | non ordonné / non dissernable |
|---|---|---|
| avec remise ou repétition | $n^p$ | $C_{n+p-1}^p$ |
| sans remise ou repétition | $A_n^p$ | $C_n^p$ |

**Remark 86** *Tirer simultanément $p$ objets d'un ensemble contenant $n$ objets est équivalent à tirer successivement sans remise $p$ objets parmi les $n$ objets de l'ensemble.*

## 3.2 Définitions et axiomes de la probabilité, espaces probabilisés

## 3.3 Probabilités conditionnelles, théorème de Bayes et ses applications en chimie analytique, indépendance des événements

## 3.4 Variables aléatoire

### 3.4.1 Variables aléatoires discrètes

#### 3.4.1.1 Définition et exemples

#### 3.4.1.2 Fonction de masse

#### 3.4.1.3 Espérance mathématique et variance

#### 3.4.1.4 Distributions discrètes courantes

### 3.4.2 Variables aléatoires continues

#### 3.4.2.1 Définition et exemples

#### 3.4.2.2 Fonction de densité

#### 3.4.2.3 Espérance mathématique et variance

#### 3.4.2.4 Distributions gaussiennes

#### 3.4.2.5 Théorème central limite

#### 3.4.2.6 Autres distributions continues importantes