Cours : Statistique descriptive & introduction aux Probabilités

Pr : OUAISSA HAMID

Licence Education
Spécialité : Enseignement secondaire Mathématiques

Année universitaire :
2023-2024

1

Partie 1: Statistiques descriptive

1.1 Introduction

La statistique est une discipline dont l'objet est de collecter, de traiter, d'analyser, d'interpréter et de présenter des données issues de l'observation de phénomènes aléatoires afin de rendre les données plus accessible et compréhensibles pour un public non spécialiste. C'est à la fois une branche des mathématiques appliquées, une méthode et un ensemble de techniques. La statistique fait partie de ce que l'on appelle aujourd'hui la science des données (en anglais : Data Science), où l'analyse des données est pleinement 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 différentes disciplinaires et justifie bien pourquoi elle est enseignée dans toutes les filières scientifiques universitaires, de l'économie à la biologie en passant par la physique, et bien sûr les sciences de l'ingénieur.

Nous ne nous intéresserons pas à la collecte des données, qui est une tâche importante et difficile, mais qui ne relève pas des mathématiques. Si on omet la collecte des données, les méthodes statistiques se répartissent en deux classes :

- La statistique descriptive vise à résumer l'information contenue dans les données de façon synthétique et efficace. Elle s'appuie pour cela sur des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques (par exemple des moyennes). Elle permet de dégager les caractéristiques essentielles du phénomène étudié et de suggérer des hypothèses pour une étude ultérieure plus sophistiquée. Les probabilités n'ont ici qu'un rôle mineur.

- La statistique inférentielle va au delà de la simple description des données. Elle vise de faire des prévisions et de prendre des décisions au vu des observations. En général, il faut pour cela proposer des modèles probabilistes du

phénomène aléatoire étudié et savoir gérer les risques d'erreurs. Les probabilités jouent ici un rôle fondamental.

Dans la suite de ce cours, nous nous intéresserons au statistique descriptive. On peut donc définir le mot statistique descriptive comme un ensemble de données et l'activité qui consiste à les recueillir, les traiter et les interpréter. Faire de la statistique c'est étudier un ensemble d'objets de même nature appelés individus ou unités statistique, sur lesquels on observe des caractéristiques appelées variables.

On peut donc énumérer les étapes d'une étude statistique comme suit :

1. La collecte des données.
2. Présenter et résumer ces données. En général sous forme de graphes.
3. Étude des paramètres et des indicateurs.
4. Tirer des conclusions sur le comportement de la population étudiée et prise de décision.
5. En se basant sur l'analyse des données, effectuer des prévisions.

L'analyse des données est donc définie comme l'ensemble des méthodes permettant une étude approfondie d'informations et de données de nature qualitative ou quantitative (qu'on définira dans la suite). Dans l'analyse des données, on distingue :

- L'analyse univariée, qui porte sur l'étude d'une seule variable statistique (statistique à une seule variable).
- L'analyse bivariée, qui a pour objectif définir les relations entre deux variables d'une même population (statistique à deux variables).
- L'analyse multivariée, qui a pour but l'étude de plusieurs variables d'une population précise.

Pour élaborer des analyses statistiques, il est impératif de distinguer leurs types d'échelles (échelles nominale, ordinale,. . .), car les techniques et les méthodes utilisées dépendent continûment de la nature des variables auxquelles sont appliquées. Ainsi par exemple pour la représentation graphique, on n'utilisera pas le même graphe pour présenter deux variables : nominale et ordinale ou deux variables quantitatives. Commençons par définir un peu de vocabulaire.

1.1.1 Terminologie

L'étude statistique d'un phénomène quelconque nécessite d'abord un vocabulaire assez particulier à connaître. On présentera ainsi c-joint quelques définitions de termes les plus utilisés dans la statistique.
La population : l'ensemble des éléments à étudier ayant des propriétés communes (noté Ω.)
Un individu ou unité statistique: est un élément de la population étudiée.
La taille ou l'effectif total: le nombre d'individus de la population noté N.
Un échantillon : est la partie étudiée de la population.
Variable statistique ou Caractère: est une propriété ou caractéristique commune aux individus de la population, que l'on souhaite étudier, noté par X.
Les modalités d'une variable statistique : sont les différentes valeurs que peut prendre celle-ci noté par x_i, où $i = 1, ..., N$.
Série statistique : C'est l'ensemble des données de la/les variable(s) étudié(s). Noté par $\{(x_i)/ \quad i = 1, ..., N\}$.
Enquête : étude d'une ou de plusieurs variables sur une population donnée.

Exemple 1.1.1. *On veut faire une étude sur la situation familiale de 20 employés d'une entreprise. On définit alors les différents termes de la série statistique comme suit :*

- *Population : Les 20 employés.*
- *Individu ou Unité statistique : Employé.*
- *Effectif total : $N = 20$.*
- *Variable statistique ou Caractère : La situation familiale.*
- *Les modalités : Marié (M), Célibataire (C) ou Divorcé (D).*
- *Série statistique : M,M,C,C,C,D,....*

Il faut bien maitriser les types de caractère car chacun a une étude différente de l'autre. On classe les variables selon leur nature.

Définition 1.1.1. Nous distinguons deux types de variables statistiques à savoir:

- *Caractère qualitative : lorsque les valeurs prises par la variable ne sont pas mesurables. Autrement dit les modalités ne sont pas des valeurs numériques, mais plutôt un groupe de catégories, par exemple (le sexe, la nationalité, la profession...) Ainsi, On distingue :*
  - *les variables qualitatives nominales : il n'y a pas de hiérarchie entre les différentes modalités, c'est à dire les modalités ne peuvent pas être ordonnées; exemple : sexe, couleur des yeux, couleur de pétales.*

  - *les variables qualitatives ordinales : les différentes modalités peuvent être ordonnées selon selon une certaine hiérarchie; exemple : la mention au baccalauréat, la fréquence d'une activité (jamais, rarement, parfois, souvent, très souvent).*

Remarque 1.1.1. Certaines variables qualitatives peuvent être désignées par un code numérique, qui n'a pas de valeur de quantité. Exemple : le code postal, le sexe (1=garçon, 2=fille).

- *Caractère quantitative : lorsque les valeurs prises par la variable correspondent à des quantités mesurables, càd si elle prend des valeurs numériques. On distingue ainsi :*

  - *les variables quantitatives discrètes : elles prennent leurs valeurs dans un ensemble discret (valeurs isolés), le plus souvent fini ; exemple : le nombre d'enfants, la pointure du pied, le nombre d'espèces recensées sur une parcelle.*
  - *les variables quantitatives continues : elles peuvent prendre des valeurs réelles (en général sous forme d'intervalles), exemple : la taille des individus, le poids d'un individu, le périmètre d'une coquille de moule.*

Remarque 1.1.2. Quand l'étude statistique porte sur un seul caractère, on parle d'une série statistique simple (ou univariée). Si l'étude porte sur deux ou plusieurs caractères, la série est dite respectivement double (ou bivariée). Si l'étude porte sur plusieurs caractères alors la série est dite multiple (ou multivariée).

Exemple: Étudier la longueur des pétales sur une population d'iris donne une série statistique simple ; étudier la longueur et la largeur des pétales donne une série statistique double.

Exemple 1.1.2. *Un sondage réalisé sur un groupe composé de 10 personnes pris au hasard. On leur pose la question sur la marque automobile préférée. Alors l'enquête est définie comme suit :*

- *Effectif total,* $N = 10$.
- *Population : groupe de 10 personnes.*
- *Individu : personne du groupe.*
- *Caractère : marque automobile préférée.*
- *Modalités (Valeurs) : Renault (R), Tesla (T), Audi (A), Ford (F), V (volkswagen).*

- *Type du caractère : variable qualitative nominale.*
  Données individuelles : A,A,A,A,F,F,T,V,V,V.
  ⋆ *L'effectif associé à A est 4 car quatre personnes ont répondu Audi.*
  ⋆ *L'effectif associé à F est 2 car deux personnes ont répondu Ford.*
  ⋆ *L'effectif associé à R est 0 car aucune personne n'a répondu Renault.*
  ⋆ *L'effectif associé à T est 1 car une personne ont répondu Tesla.*
  ⋆ *L'effectif associé à V est 3 car trois personnes ont répondu Volkswagen.*

On regroupe ces résultats dans un tableau appelé le tableau statistique. On rappelle que les tableaux statistiques consistent à résumer et présenter les données observées sous forme numérique :

Résponse(marque)	A	F	T	R	V
Effectif	4	2	1	0	3

Table 1.1: Tableau de distribution des effectifs selon la marque choisie sous forme horizontale

Souvent en statistique descriptive, il vaut mieux de considérer des pourcentages ou proportions pour avoir une idée significatif de la représentation de chaque modalité plutôt que de considérer le nombre d'individus discret. Cette proportion n'est rien que la **Fréquence.**

Définition 1.1.2. (Fréquence) Pour chaque modalité (ou valeur) x_i, d'effectif n_i, on définit la Fréquence comme étant le rapport :

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

c'est la proportion des individus de la population ayant cette modalité.

Remarque 1.1.3. Une fréquence est soit exprimée en pourcentage (par exemple 30%) soit par un nombre compris compris entre 0 et 1 (par exemple 0.25).
-La fréquence est toujours comprise entre 0 et 1.
-on a toujours $\sum_i f_i = 1$.

Par exemple si on veut déterminer la fréquence des personnes qui ont répondu par la marque automobile Ford, elle est donnée par :

$$f_2 = \frac{2}{10} = 0.2 \text{ ou } 20\%$$

On calcule de la même façon les proportions des modalités A,T,R, et V . On regroupe alors tous les résultats dans un tableau, appelé tableau des proportions ou fréquences

Modalités (marque)	A	F	T	R	V
Fréquence	0.4	0.2	0.1	0	0.333

Table 1.2: Tableau des fréquences selon la marque choisie

Remarque 1.1.4. Le traitement statistique d'une variable qualitative est très limité. En effet, il existe peu de grandeurs ou indicateurs dont on peut s'inspirer pour étudier une distribution statistique. Le moyen le plus crédible est de représenter graphiquement les données. Par contre, aucune étude paramétrique sera envisagé après.

La représentation graphique d'une variable statistique dépend de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les diagrammes ou graphes. Dans la suite nous étudierons la représentation graphique d'une série statistique à caractère qualitatif.

Représentation graphique

Dans la littérature, de nombreuses représentations plus ou moins informatives peuvent être utilisées. Parmi ces représentations, on trouve les deux diagrammes assez classiques permettent de représenter une variable qualitative : le diagramme en bandes (dit tuyaux d'orgue) et le diagramme à secteurs angulaires (dit camembert).

Les tuyaux d'orgue:

Pour illustrer les données sous forme de tuyaux d'orgue, il faut suivre la démarche suivante :

- les modalités associés à la variable statistique sont placées sur une droite horizontale (attention : cette droite ne doit pas orienté car les modalités ne sont pas mesurables et il n'y a donc pas d'ordre entre elles).
- les effectifs (ou les fréquences) sont placées sur un axe vertical. La hauteur du tuyau est proportionnelle à l'effectif.

Exemple 1.1.3. *On reprend l'exemple du sondage. La représentation graphique associée à cette distribution est donnée par un diagramme tuyaux d'orgue (voir Figure 1.2). De la même façon, on peut représenter la marque préférée des 10 personnes par fréquences en utilisant un diagramme tuyaux d'orgue.*

Remarque 1.1.5. Les tuyaux ont une certaine épaisseur pour qu'il n'y ait pas de confusion avec les diagrammes en bâtons réservés à la variable quantitative discrète.

Diagramme par secteur (diagramme circulaire) :

Les diagrammes circulaires, ou semi-circulaires, consistent à répartir un disque ou un demi-disque, en secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence de la modalité. Pour présenter

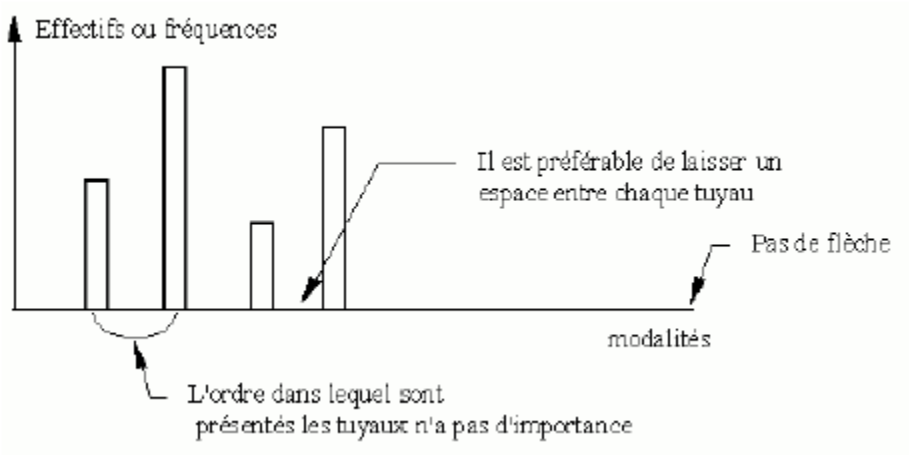

Figure 1.1: Le diagramme tuyaux d'orgue.

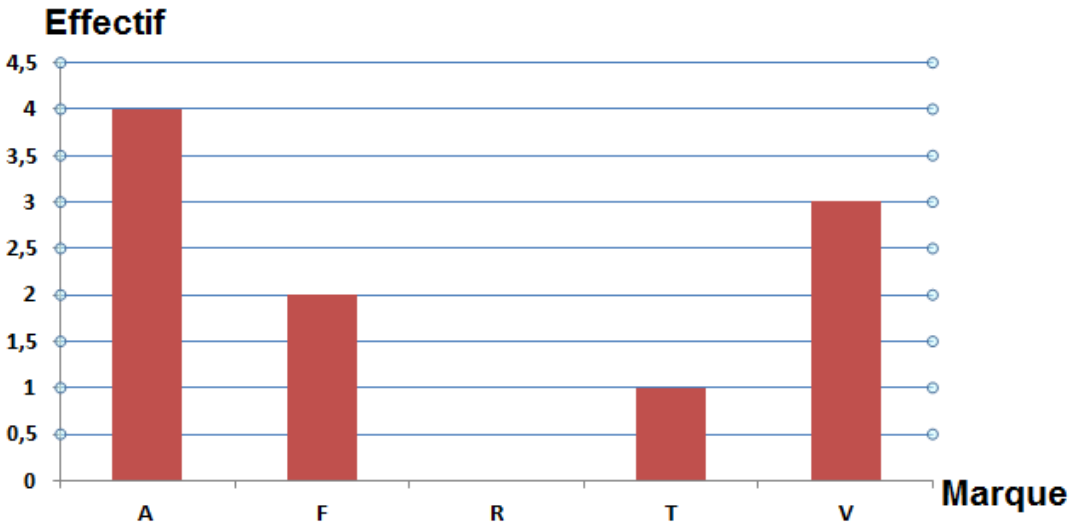

Figure 1.2: La marque préférée des 10 personnes par effectif.

les données statistiques sous forme de diagrammes à secteurs (ou camemberts), il faut suivre la démarche suivante :
- L'effectif total est représenté par un disque.
- Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement : l'angle au centre) est proportionnelle à l'effectif correspondant (voir

figure 1.3). En effet, le degré d'un secteur est déterminé à l'aide de la règle de

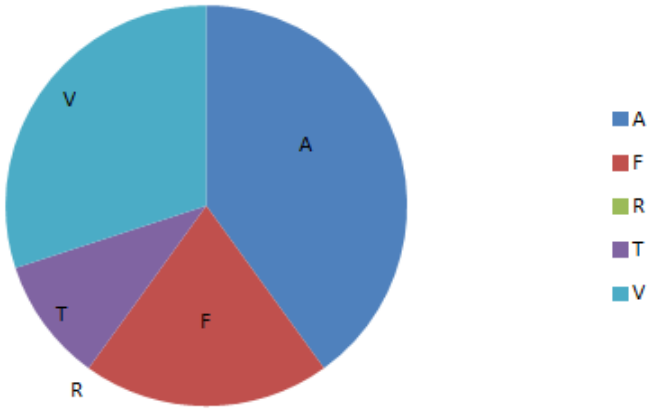

Figure 1.3: La représentation d'un diagramme circulaire exemple de sondage de marque de voitures.

trois, de la manière suivante :

$$N \to 360°$$
$$n_i \to d_i$$

où (d_i le degré de la modalité x_i d'effectif n_i).
On trouve alors l'angle représentatif de chaque modalité par la relation suivante :

$$d_i = \frac{n_i}{N}$$

1.2 Variables quantitatives discrètes

Le caractère statistique peut prendre un nombre fini dénombrable de valeurs (note, nombre d'enfants, nombre de couple, nombre de pièces, ...). Dans ce cas, le caractère statistique étudié est alors appelé un caractère discret. On traite ce type de variable comme une variable qualitative si elle prend un nombre assez petit de valeurs et comme une variable quantitative continue si elle prend un nombre considérable de valeurs.

Exemple 1.2.1. *Une province marocaine décide de recenser le nombre d'enfants des familles de son territoire. Les données ainsi obtenues ont été présentés dans le tableau de distribution des effectifs suivant :*

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

Table 1.3: Tableau de distribution des effectifs

- *Population : Les familles du village.*
- *Individu : une famille.*
- *Effectif total : 500+250+175+75=1000 familles.*
- *Variable : nombre d'enfants d'une famille.*
- *Type de variable : caractère quantitatif discret.*

On peut aussi considérer le tableau de distribution des fréquences (proportions) : Par exemple

$$f_1 = \frac{500}{1000} = 0.5 \text{ } ou \text{ } f_1 = 50\%$$

Nombre d'enfants	1	2	3	4
Fréquence	0.5	0.25	0.175	0.75

Table 1.4: Tableau de distribution des fréquences

Représentation graphique

On s'intéresse aux diagrammes différentiels qui mettent en évidence les différences d'effectifs (ou de fréquences) entre les différentes modalités. En effet, Pour représenter une variable quantitative discrète, on peut utiliser le diagramme différentiel appelé diagramme en bâtons. Les valeurs discrètes x_i prises par les variables sont placées sur l'axe des abscisses, et les effectifs (ou les fréquences) sur l'axe des ordonnées. La hauteur du bâton est proportionnelle à l'effectif.

Exemple 1.2.2. *On retourne à l'exemple, de tout à l'heure, où un recensement des nombre d'enfants dans des familles d'un village est étudié. La représentation graphique par un diagramme en bâtons de cette distribution des effectifs est illustré dans la Figure 1.4:*

Remarque 1.2.1. Avant toute tentative de représentation, il faut d'abord distinguer entre variable discrète et variable continue (regroupements en classes). Lorsque les modalités d'une variable discrète sont trop nombreuses, il est préférable de regrouper des modalités pour obtenir une variable classée afin que les graphiques synthétisent l'information et garantir une lisibilité. On passe alors du cas discret

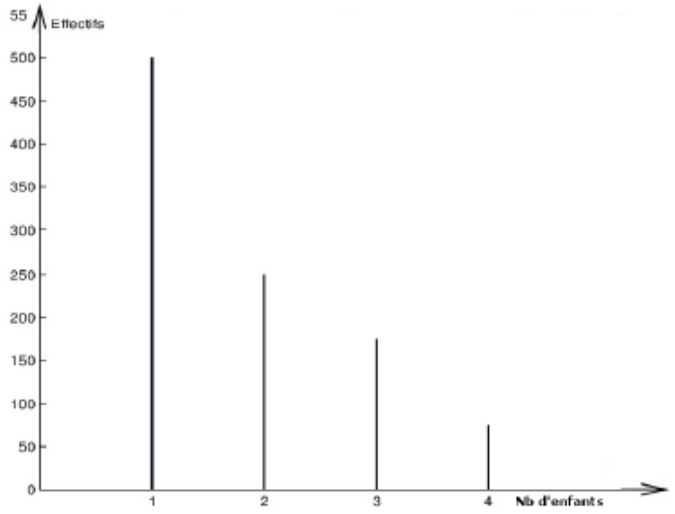

Figure 1.4: Le diagramme en bâtons des effectifs de l'exemple 1.2.1.

au cas continu (les classes ou intervalles). Une variable continue est systématiquement synthétisée dans un tableau sous forme de variables classées. Deux types de graphiques sont intéressants à représenter :

1. les diagrammes différentiels qui mettent en évidence les différences d'effectifs (ou de fréquences) entre les différentes modalités ou classes.
2. les diagrammes cumulatifs qui permettent de répondre aux questions du style "combien d'individus ont pris une valeur inférieure (ou supérieure) ?".

1.3 Variables quantitatives continues

Nous rappelons qu'une variable statistique quantitative continue concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité. On peut donc parler des valeurs décimales. Nous allons dans cette section se focaliser sur la variable statistique quantitative continue. En général, une variable continue est utilisée lorsque le nombre des modalités est assez grand, ce qui justifie le fait d'utiliser les classes. chaque classe $[x_i; x_{i+1}[$ est caractérisé par :

- son amplitude donné par la quantité $a_i = x_{i+1} - x_i$.
- son centre donné par la quantité $c_i = \frac{x_i + x_{i+1}}{2}$.
- sa borne inférieure x_i.

- sa borne supérieure x_{i+1}.

Exemple 1.3.1. - *L'amplitude de la classe* $[15;30[$ *est égale à* $30-15=15$.
- *Le centre de la classe* $[15;30[$ *est donné par* $\frac{30+15}{2}=22,5$

La répartition en classes pose deux problèmes :

- Comment alors on détermine les classes, leur nombre et leurs amplitudes?
- Comment on arrange les différentes valeurs dans ces classes ?

Voici les étapes qu'il faut suivre :

- On détermine d'abord la plus petite et la plus grande valeur prise par la variable. Puis on choisit un nombre d'intervalles appelés classes de la forme $[x_i;x_{i+1}[$ (un intervalle semi-ouvert) couvrant l'ensemble des valeurs prises par la variable.
- Ensuite, pour chaque intervalle $[x_i;x_{i+1}[$ on compte le nombres d'individus pour lesquels la variable prend une valeur supérieure ou égal à x_i et strictement inférieure à x_{i+1}. On appelle ce nombre, noté $n_{[x_i;x_{i+1}[}$, l'effectif de la classe $[x_i;x_{i+1}[$.
- Finalement, on regroupe dans un tableau à deux lignes ou à deux colonnes les différentes classes et leurs effectifs correspondants

Exemple 1.3.2. *Un sondage est réalisé pour savoir la durée en minutes du trajet domicile-ENS de dix étudiants de l'ENS pris au hasard. Nous exprimons tout d'abord l'étude statistique comme suit:*

- *Population : 10 étudiants de l'ENS*
- *Variable statistique : durée (exprimée en minutes) du trajet domicile-ENS.*
- *Type de variable : quantitative continue.*
- *Données individuelles (Modalités): 5 ;6 ;7;10;12;13;20;25;29;39*
  - La plus petite valeur $x_{\min}=5$
  - La plus grande valeur $x_{\max}=39$

- *On propose alors le choix des classes* $[5;15[$, $[15;30[$ *et* $[30;40[$.
  • *L'effectif de la classe* $[5;15[$ *est* $n_{[5;15[}=6$ *car six étudiants mettent moins de 15 minutes pour rejoindre l'ENS.*
  • *L'effectif de la deuxième classe* $[15;30[$ *est* $n_{[15;30[}=3$ *car trois étudiants mettent plus de 15 minutes dans le trajet mais strictement inférieur à 30 minutes pour rejoindre la l'ENS.*

• *L'effectif de la dernière classe* $[30;40[$ *est* $n_{[30;40[} = 1$ *car un étudiant a besoin de plus de 30 minutes pour rejoindre l'ENS.*
On présente alors le tableau statistique en classes représentant la durée du trajet pour arriver à l'ENS pour les 10 étudiants comme suit :

Durée	[5;15[	[15;30[	[30;40[
Effectif	6	3	1

Table 1.5: Tableau de distribution des effectifs

Remarque 1.3.1. -La somme des effectifs des différentes classes doit être égal à l'effectif total.
- Le tableau de distribution des effectifs contient moins d'informations que les données individuelles. En effet, connaître l'effectif d'une classe ne donne pas d'information sur par exemple la répartition des données individuelles à l'intérieur de la classe.

Attention: Le choix des classes est très important ! En effet plus l'amplitude des classe est grande moins on a des informations concrètes sur la répartition des données.
Le choix du nombre de classes peut se faire selon :
-la formule de Sturge

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

-ou bien la formule de Yule

$$k = 2.5 * N^{1/4}$$

Nous mentionnons que les deux formules sont presque pareils si $N << 200$.
où N désigne l'effectif total. Mais ce choix n'est pas toujours efficace surtout si les données ne sont pas bien réparties sur les classes.
On supposera dans toutes les études qui suivent que la distribution à l'intérieur des classes est uniforme (voir Figure 1.5). Cette hypothèse permet de justifier le fait qu'on choisisse le centre des classes comme une moyenne des extrémités.
Représentation graphique:
Nous pouvons représenter le tableau statistique d'une variable statistique par un Histogramme. On porte en abscisse les extrémités de classes. Chaque intervalle ainsi délimité devient la base d'un rectangle dont l'aire est proportionnelle à l'effectif n_i (ou la fréquence f_i) associée. En d'autre termes, c'est un ensemble de rectangles contigus, chaque rectangle associé à chaque classe ayant une surface proportionnelle à l'effectif (ou la fréquence) de cette classe (voir la Figure 1.6) Quand chaque classe a même amplitude l'histogramme est dit régulier, mais un problème se pose dans le cas d'amplitudes inégales.

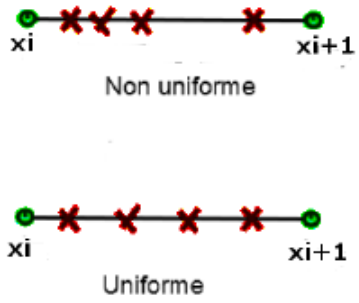

Figure 1.5: Une représentation de la distribution des valeurs à l'intérieur d'une classe uniforme et non-uniforme.

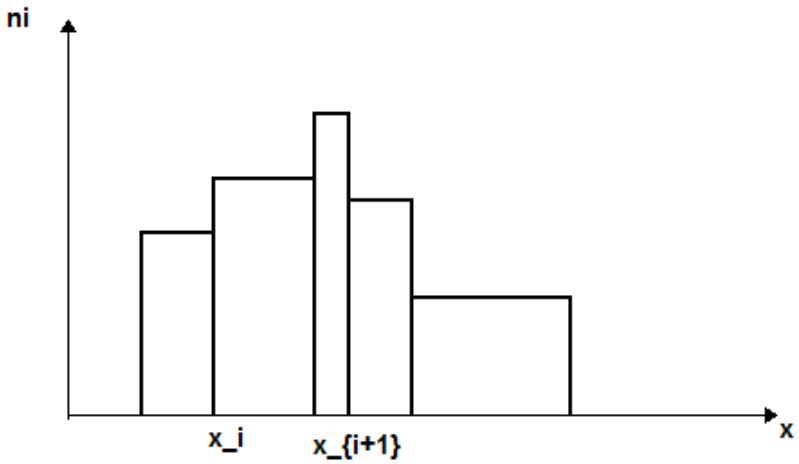

Figure 1.6: Exemple d'un Histogramme

Remarque 1.3.2. Attention : Avant d'établir tout histogramme, il faut faire attention et regarder si les classes sont d'amplitudes égales ou inégales. Si les classes sont de même amplitude, on représente directement les classes avec hauteur proportionnelle à l'effectif (ou fréquence). Sinon il faut procéder par une correction de l'effectif ou bien une détermination préalable des densités de fréquence des différentes classes. D'abord, nous utilisons les densités de fréquence des différentes classes pour tracer le Histogramme lorsque les classes ne sont pas de même ampli-

tude.

Définition 1.3.1. (densité de fréquence)
Étant donnée une classe $[x_i; x_{i+1}[$ de fréquence $f_{[x_i;x_{i+1}[} = \frac{n_{[x_i;x_{i+1}[}}{N}$.
La densité de fréquence, notée d_i, de la classe $[x_i; x_{i+1}[$ est alors donnée par

$$d_i = \frac{f_{[x_i;x_{i+1}[}}{x_{i+1} - x_i}$$

En utilisant la même notation, on peut aussi définir la densité d'effectif comme suit :

Définition 1.3.2. (densité d'effectif)
Étant donnée une classe $[x_i; x_{i+1}[$ d'effectif $n_{[x_i;x_{i+1}[}$
La densité d'effectif, notée d_i, de la classe $[x_i; x_{i+1}[$ est alors donnée par

$$d_i = \frac{n_{[x_i;x_{i+1}[}}{x_{i+1} - x_i}$$

Pour mieux comprendre comment il faut tracer l'histogramme pour les classes qui n'ont pas la même amplitude, on considère l'exemple précédant:

Exemple 1.3.3. *Durée du trajet domiciile-ENS de 10 étudiants:*

Classe	[5;15[	[15;30[	[30;40[
Fréquence	0.6	0.3	0.1
Amplitude	10	15	10

Table 1.6: Tableau de distribution des fréquences

On commence par calculer les densités pour chaque classe :
La densité de fréquence de la classe $[5; 15[$ *est donnée par* $d_1 = \frac{0.6}{10} = 0.06$.
La densité de fréquence de la classe $[15; 30[$ *est donnée par* $d_2 = \frac{0.3}{15} = 0.02$.
La densité de fréquence de la classe $[30; 40[$ *est donnée par* $d_3 = \frac{0.1}{10} = 0.01$.
On résume alors les densités calculées dans le tableau suivant:

Classe	[5;15[	[15;30[	[30;40[
Densité de fréquence	0.06	0.02	0.01

On passe maintenant la représentation graphique.

Représentation graphique par Histogramme:
Pour représenter une variable quantitative continue, on utilise le histogramme. On positionne sur l'axe des abscisses les différentes classes $[x_i; x_{i+1}[$. Ensuite, on trace des rectangle dont la longueur est proportionnelle à la densité de fréquence correspondante à ces classes. Ci-dessous le histogramme associé à l'exemple en haut.

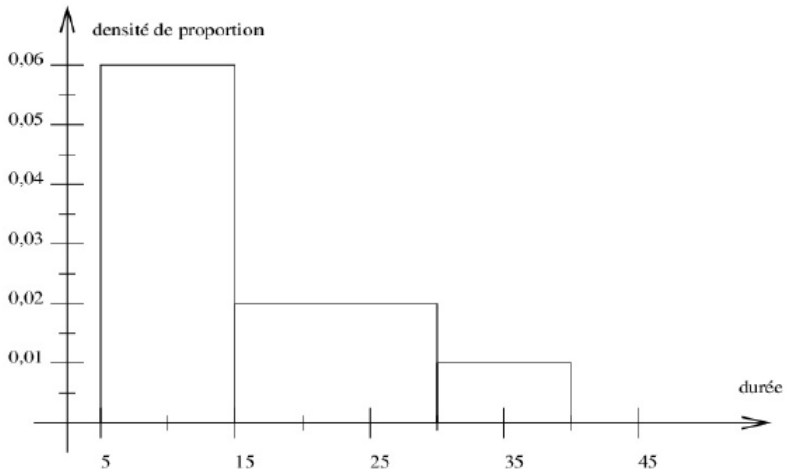

Figure 1.7: Le histogramme correspondant aux densités de fréquence

Exemple 1.3.4. *(Durée du trajet domicile-ENS) Ainsi, on peut bien constater que la surface de chaque rectangle est exactement la fréquence de la classe correspondante. On peut même le démontrer facilement, soit S la surface du rectangle dans l'histogramme, on a alors :*

$$S = hauteur \times largeur = \frac{f_{[xi;xi+1[}}{x_{i+1} - x_i} \times (x_{i+1} - x_i) = f_{[xi;xi+1[}.$$

La surface totale de l'histogramme est égale à 1 puisqu'elle est la somme des surfaces des rectangles dont leur surface ne sont que les fréquences correspondantes. Parfois on ne dispose pas du tableau statistique mais à la place on a un histogramme. Dans ce cas il faut savoir estimer la fréquence à partir de l'histogramme.
Exercice *Estimation d'une fréquence à partir de l'histogramme.*
Déterminer la fréquence des données comprises entre 10 min et 20 min à partir de l'histogramme ?

1.4 Répartitions des effectifs et des fréquences

Nous nous intéressons par la suite à définir les effectifs et les fréquences cumulés qui nous aident à trouver une réponse aux questions de genre combien d'individus ou bien la proportion de la population admettant au moins (ou bien au plus) une certaine valeur de modalité.

1.4.1 Cas discret

Soit X une variable quantitative discrète définie dans une population Ω, prenant les valeurs $x_1; ...; x_k$ classées par ordre croissant avec les effectifs partiels $n_1; ...; n_k$ respectivement.

Effectifs cumulés

Définition 1.4.1. On appelle effectif cumulé croissant de la modalité x_k la somme des effectifs partiels des valeurs $x_1, .., x_k$. c'est à dire

$$N_k = n_1 + ... + n_k = \sum_{i=1}^{k} n_i$$

Ainsi, l'effectif cumulé croissant associé à la modalité x_i du caractère X est le nombre d'individus de la population Ω dont les modalités sont inférieures ou égales à x_i.

Définition 1.4.2. On appelle effectif cumulé décroissant de la valeur x_k le nombre

$$\tilde{N} = N - \sum_{i=1}^{k-1} n_i$$

On considère l'exemple suivant pour mieux comprendre l'effectif cumulé croissant et décroissant.

Exemple 1.4.1. *Dans le tableau suivant, on présente le nombre d'enfants de 1000 familles dans un village.*

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

L'effectif cumulé croissant de la première modalité est $N_1 = 500$.
L'effectif cumulé croissant de la deuxième modalité est $N_2 = 500 + 250 = 750$.
L'effectif cumulé croissant de la troisième modalité est $N_3 = 500+250+175 = 925$.
L'effectif cumulé croissant de la quatrième modalité est $N_4 = 500+250+175+75 = 1000$.
De même on détermine les effectifs cumulés décroissants de la façon suivante:
L'effectif cumulé décroissant de la modalité x_1 *est* ($\tilde{N}_1 = 1000 - 0 = 1000$)
L'effectif cumulé décroissant de la modalité x_2 *est* ($\tilde{N}_2 = 1000 - 500 = 500$).
L'effectif cumulé décroissant de la troisième modalité x_3 *est* ($\tilde{N}_3 = 1000 - 500 - 250 = 250$).
L'effectif cumulé décroissant de la quatrième modalité x_4 *est* ($\tilde{N}_4 = 1000 - 500 -$

$250 - 175 = 75$*).*
On résume alors les effectifs cumulés dans le tableau des effectifs croissants et décroissants suivant :

Nombre d'enfants	1	2	3	4
Effectifs cumulés croissant	500	750	925	1000
Effectifs cumulés décroissant	1000	500	250	75

Fréquences cumulées

Définition 1.4.3. On appelle fréquence cumulée croissante associée à la modalité x_k la somme des fréquences des modalités $x_1; ...; x_k$. c'est à dire

$$F_k = f_1 + ... + f_k = \sum_{i=1}^{k} f_i$$

Ainsi, la fréquence cumulée croissante associée à la modalité x_k du caractère X est la fréquence d'individus de la population Ω pour lesquels la modalité correspondante est inférieure ou égale à x_k.

Définition 1.4.4. On appelle fréquence cumulée décroissante de la valeur x_k le nombre

$$\tilde{F}_k = 1 - (f_1 + ... + f_{k-1}) = 1 - \sum_{i=1}^{k-1} f_i$$

Exemple 1.4.2. *On considère l'exemple du nombre d'enfants des familles d'un village. on présente ci-dessous le tableau statistique des fréquences cumulés croissants et décroissants correspondant à cette exemple.*

Nombre d'enfants	1	2	3	4
Fréquences	0.5	0.25	0.175	0.075
Fréquences cumulés croissant	0.5	0.75	0.925	1
Fréquences cumulés décroissant	1	0.5	0.25	0.075

1.4.2 Cas continue

Le calcul des effectifs et fréquences cumulés se fait de la même façon que pour le cas discret. En effet, la fréquence (ou effectif) cumulée d'une modalité x_k est la fréquence (ou effectif) des observations dont les modalités sont inférieures ou égales cette valeur x_k. Cependant, dans le cas d'une variable continue dont les valeurs ont été regroupées sous forme de classes, c'est équivalent à chercher la fréquence d'observations qui sont strictement inférieures à la classe actuelle. On considère l'exemple suivant pour avoir une idée du calcul des fréquences cumulées.

Exemple 1.4.3. *Durée du trajet domicile-ENS*

Classe	[5;15[	[15;30[	[30;40[
Fréquence	0.6	0.3	0.1
Fréquence cumulé croissant	0.6	0.9	1
Fréquence cumulé décroissant	1	0.4	0.1

1.5 Paramètres de position et de dispersion

Les indicateurs statistiques de tendance centrale (dits aussi de paramètres de position) considérés fréquemment sont la moyenne, la médiane et le mode. Tous les trois sont faciles à calculer, mais parfois on a tendance à les confondre. Tandis que , les indicateurs statistiques de dispersion usuels sont l'étendue, l'intervalle interquartiles et l'écart type. Ces paramètres de dispersion indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur de référence. Un paramètre de dispersion s'exprime toujours dans l'unité de mesure de la variable considérée. Ainsi, si par exemple, on étudie la densité de population des régions mondiales, l'unité de mesure de la dispersion de ce caractère sera exprimée en habitants par km^2.

1.5.1 Le mode et la classe modale

Définition 1.5.1. Le Mode est un indicateur complémentaire à la moyenne et à la médiane. Il permet de donner une indication statistique de tendance centrale à un ensemble de données. Le mode d'un ensemble d'observations est la valeur la plus fréquemment rencontrée. On appelle mode d'une variable (qualitative ou quantitative discrète), la modalité ayant le plus grand effectif ou la plus grande fréquence. On note le mode par M_o.

Remarque 1.5.1. le Mode correspond aussi à (ou aux valeur(s)) ayant la plus grande hauteur dans la représentation graphique de la variable.

Exemple 1.5.1. *-Le Mode de la variable "nombre d'enfants" est $M_o = 1$ car la fréquence de x_1 (50%) est la plus élevée, ou bien car l'effectif de x_1 ($n_1 = 500$) est le plus élevé.*
- Le Mode de la variable "Matière préférée" est $M_e =$ "Chimie" car la fréquence de la matière de chimie est ($f_1 = 40\%$) est la plus élevée, ou bien car l'effectif de la matière de chimie ($n_1 = 4$) est le plus élevé.

On passe maintenant au cas continue où on parle de classe modale.

Définition 1.5.2. Si les classes sont de même amplitude, on appelle classe modale d'une variable quantitative continue la (ou les classe(s)) ayant le plus grand effectif

où la plus grande fréquence. C'est la classe correspondant au rectangle le plus haut dans l'histogramme des effectifs ou fréquences.

Exemple 1.5.2. *Durée du trajet domicile-ENS : La classe modale de la variable*

Classe	[5;15[	[15;30[	[30;40[
Densité de fréquence	0.06	0.02	0.01

"Durée du trajet domicile-école" est "[5; 15[".

Remarque 1.5.2. -Il peut y avoir une ou plusieurs classes modales.
-Attention : Si les classes ne sont pas de même amplitude, il faut corriger les effectifs ou bien les fréquences. Donc comment corriger les effectifs.

Correction des effectifs:

Si les classes ne sont pas de même amplitude, on procède par correction de l'effectif en suivants ces étapes :

- Choisir une amplitude de référence, en général celle qui se répète le plus. On la note a_r.
- Corriger les effectifs, noté n'_i en utilisant la relation suivante

$$n'_i = \frac{a_r \times n_i}{a_i} = a_r \times d_i$$

où d_i est la densité d'effectif de la classe qu'on veut corriger son effectif.

Remarque 1.5.3. si l'on raisonne en fréquences relatives (ou en pourcentage), on parle de correction de fréquences et on procède alors de la même façon, en utilisant la formule suivante :

$$f'_{[x_i,x_{i+1}[} = \frac{a_r \times f_{[x_i,x_{i+1}[}}{a_i} = a_r \times d_i$$

où $f_{[x_i,x_{i+1}[}$ est la fréquence et d_i la densité de fréquence de la classe $[x_i, x_{i+1}[$ qu'on veut corriger son effectif.

Pou bien comprendre ce processus, on considère l'exemple suivant, où une statistique sur la durée du trajet entre domicile et l'ENS de 10 étudiants :

Classe	[5;15[	[15;30[	[30;40[
Effectif (n_i)	6	3	1
Amplitude	10	15	10

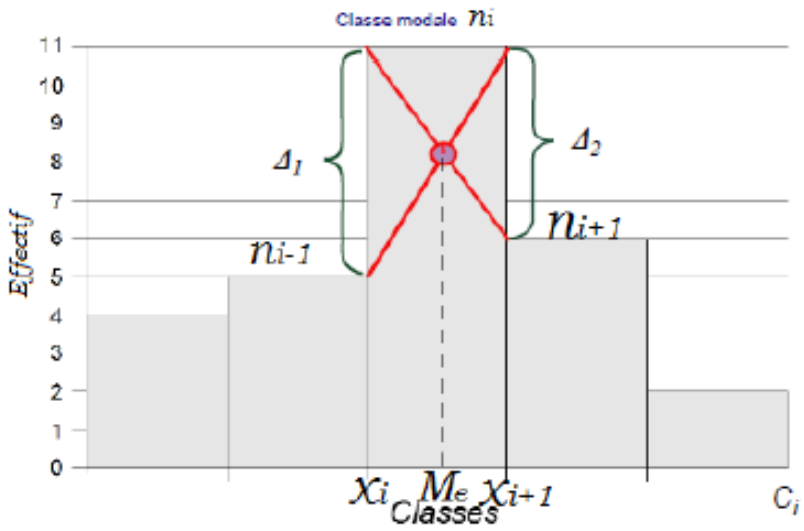

Figure 1.8: Détermination graphique du mode dans le cas continu

Exemple 1.5.3. *La classe modale est [5; 15[et non* $[15;30[$. *On choisit comme classe de référence* $[5;15[$ *puisque son amplitude se répète deux fois. Ensuite on corrige l'effectif de la deuxième classe [15; 30[comme suit*

$$n_2' = \frac{10 \times 3}{15} = 2$$

Donc le tableau avec l'effectif corrigé est donné comme suit : La classe modale est

Classe	[5;15[	[15;30[	[30;40[
Effectif corrigé (n_i')	6	2	1

donc $[5;15[$ *puisqu'elle a le plus grand effectif.*

La question qui se pose maintenant est comment déterminer le Mode M_o. En effet nous utilisons une interpolation linéaire qu'on peut la voir graphiquement.
Détermination du mode dans une classe : Pour déterminer le mode à partir de la classe modale, on utilise une interpolation linéaire, inspiré de la Figure 1.8. Le mode est donc déterminé par l'expression suivante :

$$M_o = \frac{\Delta_1}{\Delta_1 + \Delta_2}(x_{i+1} - x_i) + x_i$$

avec

$$\Delta_1 = n_i - n_{i-1},$$

et

$$\Delta_2 = n_i - n_{i+1},$$

Si on revient maintenant à l'exemple précédent, on a $\Delta_1 = 6 - 0 = 6$ et $\Delta_2 = 6 - 2 = 4$, alors on a

$$M_o = \frac{6}{6+4}(15-5) + 5 = 11$$

1.5.2 Le médiane

La Médiane désigne la valeur centrale d'une série statistique dont les valeurs observées ont été rangées dans l'ordre croissant, est la valeur qui partage la population étudiée en deux sous-ensembles de même effectif. C'est la valeur de la variable statistique qui divise la population en deux sous-populations de même effectif. on la note par m. Comment alors calculer la médiane ?

Cas d'une variable discrête

— Si l'effectif total N est impair, la médiane est la modalité du rang $\frac{N+1}{2}$, c'est :

$$m = x_{\frac{N+1}{2}}$$

— Si l'effectif total N est pair, la médiane est la moyenne entre les deux variables de rang $\frac{N}{2}$ et $\frac{N}{2}+1$, c-à-d :

$$m = \frac{x_{\frac{N}{2}+1} + x_{\frac{N}{2}}}{2}$$

On considère l'exemple suivant pour illustrer le calcul de la médiane :

Exemple 1.5.4. *Soit un échantillon de 10 personnes dont l'âge est donné comme suit :*

$$32, 18, 45, 38, 60, 64, 19, 20, 26, 24$$

Il faut d'abord arranger les valeurs de cette série par un ordre croissant, tel que :

$$18, 19, 20, 24, 26, 32, 38, 45, 60, 64.$$

Puisque l'effectif total $N = 10$ est pair, alors on applique directement la règle tel que :

$$m = \frac{x_{\frac{N}{2}+1} + x_{\frac{N}{2}}}{2} = \frac{26+32}{2} = 29$$

1.5.3 Cas d'une variable continue

Pour les données groupées en classe, il faut dans un premier temps définir la classe médiane (classe qui contient la médiane) à partie des effectif cumulés N_i. Ensuite, il faut envisager une interpolation linéaire autour de la médiane. On sait que $m \sim \frac{N}{2}$, donc il faut chercher la modalité de rang $\frac{N}{2}$, dans le tableau statistique. On considère l'exemple suivant pour mieux comprendre comment déterminer la médiane dans le cas continu.

Exemple 1.5.5. *On considère l'exemple qu'on a utilisé avant, où une statistique sur la durée du trajet entre domicile et l'ENS de 10 étudiants : On a* $N = 10$,

Durée	[5;15[	[15;30[	[30;40[
Effectif	3	4	3
Effectif cumulé	3	7	10

Table 1.7: Tableau de distribution des effectifs selon la durée du trajet

puisque $m \sim \frac{N}{2} = 5$. *Il faut alors chercher la modalité du rang 5 dans le tableau des effectifs cumulés. Ceci correspond à la deuxième classe, donc* $m \in [15; 30[$. *Il reste à définir la médiane par interpolation. On a alors :*

$$\begin{aligned} &(1) \quad m \to 5 \\ &(2) \quad 15 \to 3 \\ &(3) \quad 30 \to 7 \end{aligned}$$

On commence par l'opération suivante $(3) - (2)$, *nous donne*

$$(4) \quad 15 \to 4$$

Après, on applique l'opération suivante $(1) - (2)$, *ce qui donne :*

$$(5) \quad m - 15 \to 2$$

On reprend maintenant les relations (4) *et* (5) *et on utilise une règle à trois pour déterminer* m, *ce qui donne :*

$$m - 15 = \frac{15 \times 2}{4}$$

D'où :

$$m = 22.5$$

1.5.4 La moyenne

La moyenne, notée x est la moyenne arithmétique des modalités x_i pondérées par n_i. C'est un paramètre de position qui correspond au centre de gravité de la distribution.

Définition 1.5.3. Étant donné une série statistique quantitative discrète X prenant les valeurs $x_1; ...; x_p$ avec les effectifs partiels $n_1; ...; n_p$. Le nombre

$$\bar{x} = \frac{n_1 x_1 + ... + n_p x_p}{N} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i$$

où $N = \sum_{i=1}^{p} n_i$ désigne l'effectif total de la série, la moyenne arithmétique de la série X.

Remarque 1.5.4. -La formule donnant $\bar{x}$ peut se réécrire

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

où $f_i = n_i/N$ est la fréquence de la valeur x_i.
-La moyenne de n nombres $x_1; ...; x_n$ est le quotient par n de la somme de ces nombres.

Exemple 1.5.6. *On considère l'exemple du nombre d'enfants de* 1000 *familles dans un village.*

Nombre d'enfants	1	2	3	4
Pourcentage	50%	25%	17.5%	7.5%

Table 1.8: Tableau de distribution des effectifs

$$\bar{x} = 0.5 \times 1 + 0.25 \times 2 + 0.175 \times 3 + 0.075 \times 4 = 1.825$$

Définition 1.5.4. Pour une variable quantitative continue la moyenne arithmétique est aussi définie par l'une des formules précédentes en retenant pour x_i les centres des classes c_i.

Exemple 1.5.7. *Durée su trajet domicile-ENS*

$$\bar{x} = \frac{6 \times 10 + 3 \times 22.5 + 1 \times 35}{10} = 16.25$$

La durée du trajet domicile-ENS moyenne est de 16.25 minutes.

Classe	[5;15[	[15;30[	[30;40[
Effectif (n_i)	6	3	1
Centre (c_i)	10	22.5	35

Remarque 1.5.5. La moyenne calculée à partir des données individuelles est :

$$\bar{x} = \frac{5+6+7+10+12+13+20+25+29+39}{10} = 16.60$$

La moyenne calculée sur les données regroupées n'est pas toujours égale à celle calculée sur les données individuelles. Problème d'arrondissement.
— La "vraie" valeur de la moyenne est celle calculée sur les données individuelles.
— La moyenne calculée sur les données regroupées est une valeur approchée de la bonne valeur de la moyenne; il y a certainement une perte d'informations lors du regroupement des données en classes.

1.5.5 Autres moyennes

Définition 1.5.5. Soit r un entier rationnel non nul ($r \in \mathbb{Q}*$). On appelle la moyenne d'ordre r de la série statistique $(x_i; n_i)_{1\leq i\leq p}$ la quantité

$$[\frac{1}{N}\sum_{i=1}^{p} n_i x_i^r]^{\frac{1}{r}}$$

Selon la valeur de r, on définit plusieurs moyennes

- Si $r = -1$, on parle de la moyenne harmonique définit par

$$\bar{x}_h = \frac{N}{\sum_{i=1}^{p} \frac{n_i}{x_i}}$$

- si $r \sim 0$, on parle de moyenne géométrique

$$\bar{x}_g = \left[\prod_{i=1}^{p} x_i^{n_i}\right]^{\frac{1}{N}}$$

- si $r = 1$, on parle de moyenne arithmétique donné par

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

- si $r = 2$, on parle de moyenne quadratique donné par

$$\bar{x}_q = \left[\frac{1}{N}\sum_{i=1}^{p} n_i x_i^2\right]^{\frac{1}{2}}$$

Remarque 1.5.6. — On ne peut pas parler de la moyenne harmonique que si toutes les observations sont non nulles. Cependant, la moyenne géométrique n'est calculable que si toutes les observations sont strictement positives.
— La comparaison entre les différentes moyennes donne

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_q$$

Maintenant, on s'intéresse aux paramètres de dispersion, on commence par l'étendu.

1.5.6 L'étendu

Définition 1.5.6. La différence entre la plus grande valeur et la plus petite valeur prise par la variable

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

s'appelle l'étendue de la Variable statistique X. Le calcul de l'étendue est très simple.

Exemple 1.5.8. *L'étendu de la variable statistique décrivant le nombre d'enfants de 1000 familles d'un village vaut*

$$e = 4 - 1 = 3$$

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).

1.5.7 Variance

Définition 1.5.7. Soit X une variable quantitaive discrète définit sur une population Ω, d'effectif total N prenant les valeurs $x_1; ...; x_p$ avec les effectifs partiels $n_1; ...; n_p$. On appelle variance, notée σ^2 ou $V(X)$, de X le nombre

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

Nombre d'enfants	1	2	3	4
Effectif	500	250	175	75

Table 1.9: Tableau de distribution des effectifs

Exemple 1.5.9. *on a déja calculé la moyenne* $\bar{x} = 1.825$. *La variance est donnée*

Nombre d'enfants	1	2	3	4
$x_i - \bar{x}$	-0.825	0.175	1.175	2.175
$(x_i - \bar{x})^2$	0.681	0.031	1.381	4.731
Effectif (n_i)	500	250	175	75

par

$$\begin{aligned}\sigma^2 &= \frac{(500 * 0.681) + (250 * 0.031) + (175 * 1.381) + (75 * 4.731)}{1000} \\ &= 0.945\end{aligned}$$

Remarque 1.5.7. On montre que l'on peut encore écrire la variance sous la forme

$$\sigma^2 = \frac{\sum_{i=1}^{p} n_i x_i^2}{N} - \bar{x}^2 \tag{1.1}$$

Avec cette relation, le calcul devient beaucoup plus rapide et plus pratique.
Exercice : Démontrer la formule 1.1?

Exemple 1.5.10. *On considère l'exemple précédant pour le calcul de la moyenne*

Nombre d'enfants	1	2	3	4
x_i^2	1	4	9	16
Effectif (n_i)	500	250	175	75

$$\begin{aligned}\sigma^2 &= \frac{(500 \times 1) + (250 \times 4) + (175 \times 9) + (75 \times 16)}{1000} - (1.825)^2 \\ &= 0.945\end{aligned}$$

Remarque 1.5.8. La variance peut s'écrire en utlisant les fréquences sous forme suivante :

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

où $f_i = \dfrac{n_i}{N}$ est la fréquence de la valeur x_i.
De même on montre que

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

Exemple 1.5.11. *on a déja calculé la moyenne* $\bar{x} = 1.825$. *La variance se calcule rapidement en utilisant la formule* 1.1.

$$\begin{aligned}\sigma^2 &= (0.5 \times 1) + (0.25 \times 4) + (0.175 \times 9) + (0.075 \times 16) - (1.825)^2 \\ &= 0.945\end{aligned}$$

Nombre d'enfants	1	2	3	4
Fréquence	0.5	0.25	0.175	0.075

Table 1.10: Tableau de distribution de fréquences

Nombre d'enfants	1	2	3	4
x_i^2	1	4	9	16
Fréquence	0.5	0.25	0.175	0.075

Définition 1.5.8. Pour une variable quantitative continue la variance est aussi définie par les mêmes formules précédentes en remplaçant les x_i par les centres des classes c_i.

De même que pour la moyenne arithmétique, dans le cas de variable quantitative continue, la variance peut être calculer à partir des données individuelles où à partir des données regroupées.

Exemple 1.5.12. *Durée du trajet domicile-ENS Données regroupées en classes : -Le calcul de la variance à partir*

Classe	[5;15[	[15;30[	[30;40[
Effectif	6	3	1
Fréquence	0.6	0.3	0.1
Centre	10	22.5	35

$$\sigma^2 = \frac{5^2 + 6^2 + 7^2 + 10^2 + 12^2 + 13^2 + 20^2 + 25^2 + 29^2 + 39^2}{10} - (16.60)^2 = 115.44$$

-Calcul de la variance à partir des effectifs :

$$\sigma^2 = \frac{(6 \times 10^2) + (3 \times 22.5^2) + (1 \times 35^2)}{10} - (16.25)^2 = 70.31$$

-Calculer la variance à partir des fréquences ?

Remarque 1.5.9. — La variance calculée à partir des données regroupées n'est pas toujours égale à celle calculée sur les données individuelles. Mais en général la différence n'est pas énorme c'est dû aux erreurs d'arrondissements.
— La valeur de la variance la plus précise est celle calculée sur les données individuelles.
— La variance calculée à partir des données regroupées est une valeur approchée de la valeur précise de la variance ; il y a eu certes une perte d'information lors du regroupement des données en classes.

On passe maintenant à un autre paramètre de dispersion qui n'est en fait que la racine de la variance.

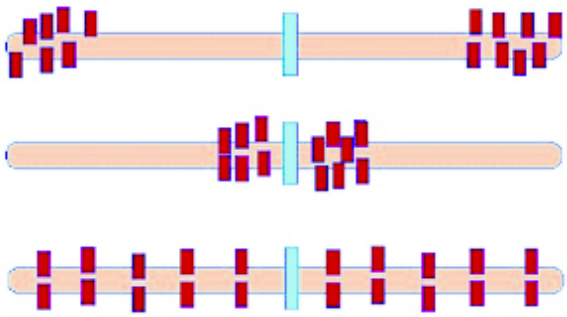

Figure 1.9: La dispersion d'une série statistique autour de sa moyenne.

1.5.8 L'écart type

L'écart-type est une mesure de la dispersion d'une série statistique autour de sa moyenne. C'est le paramètre de dispersion le plus utilisé dans la littérature. On définit l'écart-type, noté σ, comme étant la racine carrée de la variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(X)}$$

Remarque 1.5.10. Le paramètre σ mesure la distance moyenne entre $\bar{x}$ et les valeurs de X (voir Figure 1.5.8). 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).

1.5.9 Coefficient de variation

Définition 1.5.9. Etant donné une variable quantitative X, de moyenne $\bar{x}$ et d'écart-type σ. On appelle coefficient de variation de la variable X le rapport

$$C.V = \frac{\sigma}{\bar{x}}$$

Le coefficient de variation mesure la dispersion relative des données individuelles par rapport à la moyenne. C'est un coefficient "sans dimension", invariant si on change l'unité de mesure.
Ce coefficient permet de comparer les dispersions de distributions qui ne sont pas exprimées dans la même unité (comme les salaires dans deux pays différents), ou

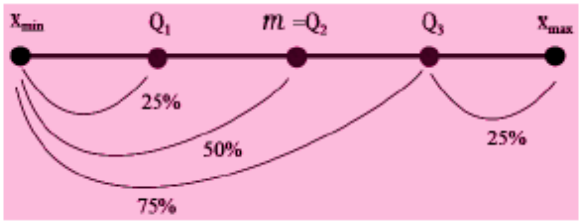

Figure 1.10: La répartition des quartiles.

deux distributions exprimées dans la même unité mais de moyennes différentes (comme les salaires dans deux entreprises du même pays). Les avantage de ce paramètre par rapport à l'écart-type :
— L'écart-type seul ne permet le plus souvent pas de juger de la dispersion des valeurs autour de la moyenne. Si par exemple une distribution a une moyenne de 10 et un écart-type de 1 (CV de 10 %), elle sera beaucoup plus dispersée qu'une distribution de moyenne 1000 et d'écart-type 10 (CV de 1 %).
— Ce nombre est sans unité, c'est une des raisons pour lesquelles il est parfois préféré à l'écart type qui lui ne l'est pas. En effet, pour comparer deux séries de données d'unités différentes, l'utilisation du coefficient de variation est plus judicieuse.

1.5.10 Le quartiles

En statistique descriptive, un quartile est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population. Le quartile fait partie des quantiles et il généralise la médiane. Le quartile est calculé en tant que 4-quartiles (voir Figure 1.10).
— Le 1er quartile est la donnée de la série qui sépare les 25% inférieurs des données (notation Q1);
— Le 2e quartile est la donnée de la série qui sépare les 50(notation Q_2) ; il est également appelé médiane (noté aussi m) ; Le 3e quartile est la donnée de la série qui sépare les 75% inférieurs des données (notation Q_3) ; Par extension : le 0 ème quartile est la donnée de la série qui sépare les 0% inférieurs des données (notation Q_0, c'est le minimum) et le 4ème quartile est la donnée de la série qui sépare les 0% supérieurs des inférieurs des données (notation Q_4, c'est le maximum)

Détermination des quartils

La détermination des quartiles dans le cas continu se fait de la même façon que la médiane. C'est à partir de l'effectif cumulé du tableau statistique tel que :

$$Q_1 \to \frac{N}{4}$$

$$Q_2 \to \frac{N}{2}$$

$$Q_3 \to \frac{3N}{4}$$

Ensuite une interpolation linéaire est utilisée comme pour le cas de la médiane.
Dans **le cas discret**, on range les données par ordre croissant : si l'effectif total est N valeurs :
— Le quartile zéro (minimum) est celui qui a le rang 1.
— Le premier quartile est la modalité de rang $(N+3)/4$.
— La deuxième quartile (médiane) est celui qui a le rang $(2N+2)/4$ que l'on simplifie en $(N+1)/2$.
— Le troisième quartile est celui qui a le rang $(3N+1)/4$.
— le quatrième quartile est celui qui a le rang N.

Exemple 1.5.13. *L'âge de 13 personnes pris au hasard d'un village au Maroc est donné par ordre croissant comme suit :*

$$4; 6; 13; 22; 28; 34; 40; 44; 50; 60; 62; 64; 70.$$

Calcul de Q_1 ***:*** *On applique la règle qui dit que* $Q_1 \to (N+3)/4 = \frac{16}{4} = 4$. *Donc le premier quartile est la modalité du 4éme ordre, d'où :*

$$Q_1 = 22$$

Calcul de Q_3 ***:*** *On applique la règle qui dit que* $Q_3 \to (3N+1)/4 = \frac{3*13+1}{4} = 10$. *Donc la troisième quartile est le modalité du 10éme ordre, d'où*

$$Q_3 = 60$$

Calcul de Q_2 ***:*** *On applique la règle classique pour déterminer la médiane* $Q_2 \to (N+1)/2 = \frac{13+1}{2} = 7$ *(car N est impaire). Donc la médiane est la modalité du 7éme ordre, d'où :*

$$Q_2 = 40$$

1.5.11 Boite à moustaches

On s'intéresse maintenant au diagramme illustrant cette notion de quartiles:

Définition 1.5.10. La Boîte à moustaches, connue aussi sous le nom de Diagramme de Tukey ou box-plot, est un graphique représentatif (voir Figure 1.11), où on présente à la fois la médiane et les quartiles. Ceci permet de donner une idée sur l'asymétrie et la dispersion des valeurs ou modalités de la distribution statistique.

Exemple 1.5.14. *Durée du trajet domicile-ENS.*

Classe	[5;15[	[15;30[	[30;40[
Fréquence	6	3	1

Table 1.11: Tableau de fréquences

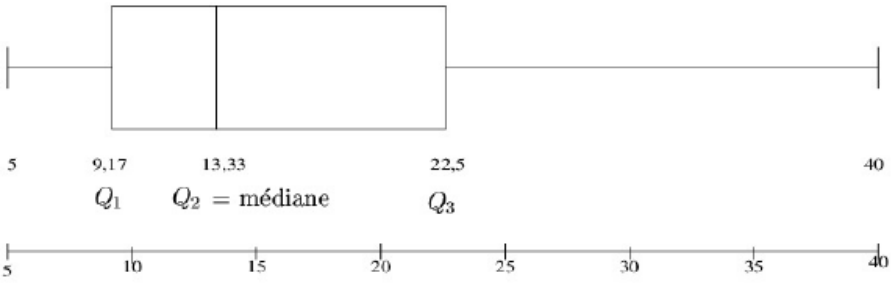

Figure 1.11: Le diagramme à moustache

Remarque 1.5.11. Pour mettre un lien entre le diagramme à moustache et la courbe statistique associée à la distribution étudiée, on considère ci-dessous une visualisation des caractéristiques d'une distribution à l'aide d'un box-plot.

1.5.12 Ecart inter-quartiles

Définition 1.5.11. Un écart inter-quartiles est la différence $EI = Q_3 - Q_1$. Il mesure la dispersion des données autour de la médiane.

Exemple 1.5.15. *Durée du trajet domicile-ENS.*
$[Q_1; Q_3] = [9; 17; 22; 5]$*: 50% des étudiants mettent entre 9,17 et 22,5 minutes pour aller à l'ENS.*

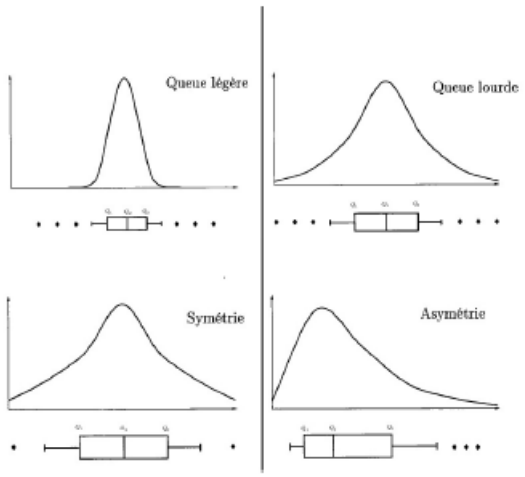

Figure 1.12: Visualisation des données statistiques en fonction du box-plot.

25% font moins de 9.17 minutes tandis que 25% font plus de 22.5.
-L'intervalle inter-quartile est l'intervalle $[Q1; Q3]$. *Il contient* 50% *des observations.*
-25% des observations sont inférieures à Q_1 *;* 25% *des observations sont supérieures à* Q_3.

Remarque 1.5.12. — Dans la boite de moustache, on peut lire $Q_3 - Q_1$ comme taille de la moustache.
— L'écart inter-quartile est généralement utilisé pour détecter l'existence éventuelle des valeurs aberrantes.
En effet, plusieurs logiciels statistiques comme le S-plus ou le SPSS, considèrent comme valeur aberrante, toute valeur x qui se trouve à l'extérieur de l'intervalle $[Q_1 - 1.5 \times (Q_3 - Q_1); Q3 + 1.5 \times (Q_3 - Q_1)]$

1.5.13 Paramètres de forme

Il existe plusieurs paramètres de forme dans la littérature, nous allons traiter ici seulement quelques-uns.

Moments centrés d'ordre r

Afin de distinguer les formes des distributions statistiques, on utilise des paramètres dites **moments centrés** de la variable statistique.

Définition 1.5.12. Soit X une variable quantitative définit sur une population Ω, d'effectif total N prenant les valeurs $x_1; ...; x_p$ avec les effectifs partiels $n_1; ...; n_p$

et m_r, le moment centré d'ordre $r \in \mathbb{Q}$ s'écrit

$$m_r = \frac{1}{N} \sum_{i=1}^{p} n_i (x_i - \bar{x})^r.$$

Lorsque $r = 1$, le moment centré d'ordre 1 est nul : $m_1 = 0$, lorsque r = 2, le moment centré d'ordre 2 n'est rien d'autre que la variance.

Coefficient d'asymétrie (Skewness)

On distingue trois types de distributions selon leur symétrie par rapport à la moyenne comme la distribution normale, dissymétrique à gauche ou à droite.
Comparaison Mode-Moyenne-Médiane
Si la distribution est symétrique (voir la figure 1.13) Si la distribution est dis-

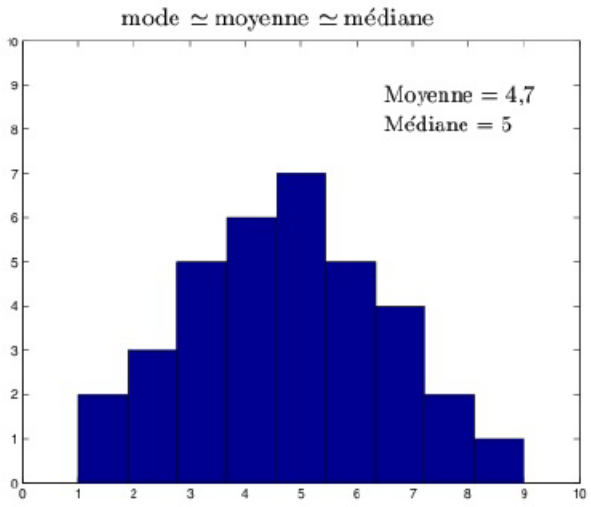

Figure 1.13: distribution symétrique

symétrique étalée à droite (voir la figure 1.14): Si la distribution est dissymétrique étalée à gauche (voir la figure 1.15): Le paramètre le plus utilisé pour caractériser l'asymétrie d'une distribution $(x_i; n_i)_i$ est le moment d'ordre 3.
On constate en effet que pour analyser la symétrie d'une distribution, il faut prévoir les cas suivant :

$$\begin{cases} \text{dissymétrique à gauche} & m_3 > 0 \\ \text{symétrique} & m_3 = 0 \\ \text{dissymétrique à droite} & m_3 < 0 \end{cases}$$

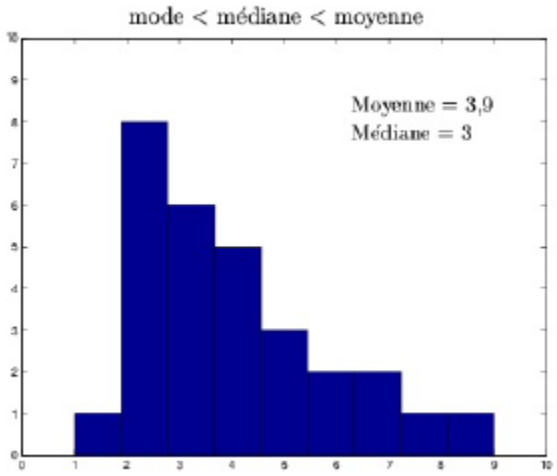

Figure 1.14: distribution dissymétrique étalée à droite

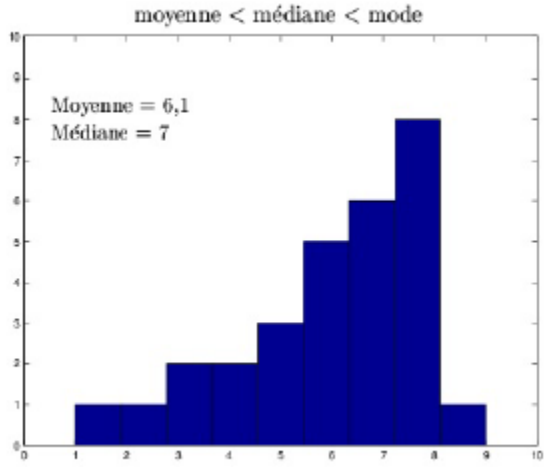

Figure 1.15: distribution dissymétrique étalée à gauche

Coefficient d'aplatissement (Kurtosis)

L'aplatissement d'une distribution est basé sur le moment centré d'ordre 4 définit comme suit :

$$m_4 = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^4$$

Définition 1.5.13. Les paramètres d'aplatissement sont :

- Le coefficient d'aplatissement de Pearson

$$\eta = \frac{m_4}{\sigma^4}$$

- Le coefficient d'aplatissement de Fisher

$$\beta = \frac{m_4}{\sigma^4} - 3$$

Plus la distribution observée est effilée, plus ces coefficients sont grands; plus la distribution des données est aplatie, plus ces coefficients sont petits.

Remarque 1.5.13. Le calcul de ces coefficients pour la distribution normale donne $\eta = 3$ est par conséquent $\beta = 0$. Donc on prend l'aplatissement de la distribution normale comme aplatissement "référentiel" ainsi pour analyser l'aplatissement d'une distribution, on a les cas suivants :

$$\begin{cases} \text{Applatie} & \text{si } \beta < 0 \\ \text{Normale} & \text{si } \beta = 0 \\ \text{Effilé} & \text{si } \beta > 0 \end{cases}$$