3ème Année
Licence
d'Excellence

Mathématiques
Appliquées
à l'Intelligence
Machine

Statistique pour Machine Learning

Abderrahim ASLIMANI

Université Mohamed Premier
Faculté Pluridisciplinaire, Nador

Année Universitaire: 2024-2025

Table des matières

Chapitre 1

Statistique descriptive (Partie I)

1.1 Introduction

La statistique est le domaine des mathématiques qui étudie les outils de recueil, de traitement et d'interprŕtation des données. La statistique mathématique s'appuie fortement sur la théorie des probabilités et développe des outils théoriques, tandis que la statistique appliquée s'attache à proposer des méthodologies dans divers domaines scientifiques (Économie et finance, Marketing et recherche de marché, Éducation, Technologie de l'information et Intelligence Artificielle, Recherche scientifique, Gouvernement, Industrie, Médecine et santé, Environnement et sciences de la terre, Sociologie,...). La statistique désigne donc la science du recueil, du traitement et de l'interprétation des données. Notons que l'utilisation du nom au pluriel (statistiques) correspond à des données obtenues par certains type de calcul, par exemple : revenu moyen, revenu médian, taux de chômage.

La statistique descriptive est l'ensemble des méthodes et techniques permettant de présenter, de décrire, de résumer des données nombreuses et variées.
Il faut d'abord préciser l'ensemble étudié, appelé **population statistique**, dont les éléments sont des **individus**, ou unités statistiques. Il est fréquent qu'on ne puisse observer toute la population statistique, pour des raisons techniques ou budgétaires. On effectue alors une observation partielle de cette population à travers un échantillon qui est, par définition, un sous-ensemble de la population statistique. Il existe différentes procédures pour choisir un échantillon. On parle de procédure d'**échantillonnage**. Les plus courantes sont l'échantillonnage aléatoire simple et l'échantillonnage aléatoire stratifié. Pour le premier, tous les échantillons de même taille ont les mêmes chances d'être sélectionnés. Pour le second, la population statistique est divisée en strates (disjointes et relativement homogènes), et dans chacune de ces strates, un échantillonnage aléatoire simple est appliqué et ceci indépendamment d'une strate à l'autre. La statistique inférentielle est l'ensemble des méthodes permettant, à partir d'un échantillon, d'estimer des paramètres d'une population statistique et/ou de tester des hypothèses sur cette population. A l'inverse de la statistique descriptive, la statistique inférentielle fait appel à la théorie des probabilités à travers les notions de précision statistique et de risque d'erreur décisionnel. Notons qu'un individu statistique n'est pas forcément un individu biologique ni même un objet matériel. Ainsi, on peut s'intéresser à l'ensemble des accidents de la route survenus dans une région au cours d'une période donnée. L'individu statistique est alors l'accident, qui est une occurrence donc immatériel. Voici quelques exemples de population statistique :

1. Ensemble des collèges d'une académie. Pour chaque collège, on peut s'intéresser au taux de passage en seconde, au nombre d'élèves, à la présence ou pas d'une cuisine scolaire, à la commune d'implantation, au numéro de département.
2. Ensemble des parents d'élève d'un lycée. On s'intéresse à leur opinion sur un projet éducatif

selon leur profession, leur revenu, leur statut marital, le nombre d'enfants scolarisés, la distance domicile-lycée, le moyen de locomotion.

3. Ensemble des incidents de violence remontés à un rectorat au cours de l'année scolaire 2023-2024. Pour chaque incident, l'établissement concerné indique : le statut du principal acteur (élève, personnel de sécurité, personnel enseignant, personnel administratif ou technique), le type violence (physique et/ou verbale), le nombre de protagonistes, lieu (intérieur, extérieur de l'enceinte de l'établissement), le nombre de blessés.
4. Ensemble des étudiants de la licence d'excellence MAIM. L'ARS (Agence Régionale de Santé) désire étudier le comportement alimentaire chez certains jeunes et ses conséquences sur l'obésité et autres risques sanitaires. Les enquêteurs notent le poids, la hauteur, l'âge, tour de taille, tour de hanche, le sexe, la commune de résidence, le nombre de sports pratiqués, la fréquence de prise de petit-déjeuner, la taille de fratrie, régularité de consommation de divers produits.

Chaque individu statistique est donc décrit par un ou plusieurs traits distinctifs ou grandeurs physiques le caractérisant. On les appelle **variables statistiques**. Une variable statistique (ou caractère statistique) est donc ce qui est observé ou mesuré sur un individu statistique.
Quand on observe une variable statistique sur un nombre n d'individus statistiques, on obtient une suite $x_1; x_2; \ldots; x_n$ où x_i est la **modalité** ou valeur observée sur le ième individu. Cette suite est appelée **série statistique**. On parle de série statistique simple (ou univariée). Le nombre n est la taille (ou longueur) de la série. Si on observe sur chaque individu deux variables, on a alors une suite $(x_1; y_1); (x_2; y_2); \ldots; (x_n; y_n)$ appelée série statistique **double** (ou **bivariée**). D'une façon générale, si sur chaque individu statistique, il est observé un nombre de variables k (supérieur à 2), on dit que la série statistique est **multivariée**. La statistique descriptive concernant une seule variable statistique est appelée **statistique descriptive univariée** (ou unidimensionnelle). La statistique descriptive concernant plusieurs variables statistiques est dite **statistique descriptive multivariée** (ou multidimensionnelle). Cette dernière permet la description des caractères observés sur des individus et des liens éventuels entre ces caractères. Une variable peut être :

- **Quantitative :** elle concerne une grandeur mesurable. Ses valeurs sont des nombres exprimant une quantité, et sur lesquelles les opérations arithmétiques (addition, multiplication, etc,...) ont un sens. La variable peut alors être **discrète** ou **continue** selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre. Une variable quantitative discrète ne peut prendre que des valeurs isolées. Ces valeurs sont en nombre fini ou dénombrable. Le cas le plus répandu est celui où les valeurs possibles sont des nombres entiers naturels :nombre d'insectes sur une plante ; nombre de descendants dans une portée ; nombre de fruits dans un arbre ; taille de fratrie, effectif d'un établissement. Une variable quantitative continue peut prendre une infinité de valeurs sous forme d'intervalle. La taille, le poids, la surface cultivée, la température moyenne sont des variables quantitatives continues. On obtient des valeurs à la précision de l'instrument de mesure près. Je ne mesure pas exactement $1m80$ mais m'étant limité à mesurer ma taille au centimètre près, je sais seulement qu'elle est située entre $1m79$ et $1m80$.
- **Qualitative :** ses valeurs sont des modalités, ou catégories, exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens. On distingue des variables qualitatives ordinales ou nominales, selon que les modalités peuvent être naturellement ordonnées ou pas. Une variable est dichotomique si elle n'a que deux modalités.

En résumé, la statistique descriptive a pour objectif de synthétiser l'information contenue dans les jeux de données au moyen de tableaux, figures ou résumés numériques. Les variables statistiques

sont analysées différemment selon leur nature (quantitative, qualitative).

1.2 Statistique

Définition

La statistique est une branche des mathématiques. C'est un ensemble de méthodes scientifiques basées sur la collecte de données, l'organisation, la présentation et l'analyse de ces données.

La place de la statistique et son rôle dans de nombreux domaines est incontournable. Elle est utilisée dans :
L'économétrie, La medecine, La biologie, La sociologie...
On parle de statistique descriptive lorsque on fait une analyse de données observées ou mesurées. Les statistiques (A ne pas confondre avec La Statistique) sont les résultats obtenus lors d'une étude statistique.

1.2.1 Vocabulaire

A la base de toute étude statistique, il y'a une population formée d'éléments de même nature qu'on appelle individus. Les individus peuvent être des être humains, vivants, des objets ....
La population est notée Ω, le nombre de ces éléments est noté $N = Card(\Omega)$.

Définition

Le nombre N est appelé effectif total de la population.

Exemple

Les 110000 étudiants de l'université Mohamed Premierer : Ω = {des étudiants de l'université Mohamed Ier} , card(Ω) = 110000.

- Un individu (ou unité statistique) est un élément de Ω. C'est un étudiant de l'université Mohamed Premier.
- Les 30000 étudiants de la Faculté Pluridisciplinaire de Nador est un échantillon de la population

Exemple

Les 25000 salariés de l'ONCF (Office National des Chemins de Fer). Ω = {des salariés de l'ONCF}, $card(\Omega) = 25000 = N$.

- Un individu est un salarié de l'ONCF.

1.2.2 Caractère ou variable statistique

Définition

On appelle caractère, toute propriété (ou spécifité) étudiée sur les individus. Un caractère s'appelle aussi variable statistique.

Exemple

Un abonné de l'ONE (Office National de l'Electricité) peut être étudié selon son âge, sa nationalité, son sexe, son salaire. Le salaire, l'âge sont des caractères.

Exemple

Un étudiant de la FPN (Faculté Pluridisciplinaire de Nador) peut être étudié selon son groupe sanguin, sa taille, son poids etc...Le groupe sanguin, la taille, le poids sont des caractères.

Exemple

On considère un ensemble de 50 vaches laitères dont on mesure la longueur du corps.
La population $\Omega = \{50 \text{ vaches laitières}\}$ caractère : la longueur du corps.

1.2.3 Modalité d'un caractère

Définition

On appelle modalité d'un caractère, une situation dans laquelle peut se trouver un individu selon un caractère étudié.

Les modalités sont donc les différents spécifités du caractère.

Exemple

Soit le caractère "état matrimonial." Les modalités sont : Marié, divorcé, célibataire, veuf.

Exemple

Soit le caractère "groupe sanguin." Les modalités sont : A, B, AB, O.

Exemple

Soit le caractère taille d'un individu. Les modalité sont : Grand, Moyen, Petit.

Exemple

Soit le caractère "Note d'examen de mathématiques des étudiants de la filière Web Marketing." Les modalités sont 0, 1, ..., 19, 20.

Remarque

Un caractère présente plusieurs modalités, par contre, un individu n'admet qu'une et une seule modalité.

On distingue deux types de caractères : Caractère qualitatif, Caractère quantitatif.

1.2.4 Variable Statistique qualitative

Définition

Un caractère est dit qualitatif lorsque ses différentes modalités ne sont pas mesurables (c'est à dire ne sont pas des nombres). Elles peuvent être nominales ou ordinales :

- **Variables quantitatives nominales** : Ces variables représentent des catégories ou des noms sans ordre spécifique entre eux. Les valeurs numériques attribuées à ces catégories n'ont pas de signification quantitative.
- **Variables quantitatives ordinales** : Ces variables représentent des catégories avec un ordre spécifique entre elles, mais l'intervalle entre les valeurs n'est pas nécessairement constant. Les valeurs numériques attribuées ont une signification de rang ou d'ordre, mais l'écart entre les valeurs n'est pas défini.

Exemples

- Couleur des yeux : Les catégories peuvent être "bleu", "marron", "vert", etc.. ;
- Genre : Les catégories sont généralement "masculin" et "féminin". Bien que des chiffres puissent être attribués (par exemple, 1 pour masculin, 2 pour féminin), ces chiffres n'impliquent pas un ordre ;
- Type de véhicule : Les catégories peuvent inclure "voiture", "camion", "moto", etc.. ;
- Mode de paiement : Les catégories peuvent être "carte de crédit", "espèces", "virement bancaire", etc.. ;
- Marque de téléphone : Les catégories peuvent inclure des marques spécifiques comme "Apple", "Samsung", "Sony", etc.. ;
- État civil : Les catégories peuvent être "célibataire", "marié", "divorcé", etc.. ;
- Type d'animal de compagnie : Les catégories peuvent inclure "chien", "chat", "oiseau", etc.. ;
- Domaine d'études : Les catégories peuvent être "science", "arts", "commerce", etc.. ;
- Matière préférée : Les catégories peuvent inclure "mathématiques", "finance", "art", etc.. ;

sont des variables statistiques qualitatives nominales.

Exemples

- Niveau de satisfaction : Les catégories peuvent être "insatisfait", "satisfait", "très satisfait" ;
- Classe sociale : Les catégories peuvent inclure "basse", "moyenne", "élevée" ;
- Niveau d'éducation : Les catégories peuvent être "primaire", "secondaire", "baccalauréat", "master", "Doctorat" ;
- Niveau de compétence : Les catégories peuvent être "débutant", "intermédiaire", "expert" ;
- Évaluation de la qualité : Les catégories peuvent être "mauvaise", "moyenne", "bonne".

Ces exemples illustrent des variables qualitatives ordinales où les catégories sont ordonnées.

Remarque

Dans les deux cas, bien que des valeurs numériques soient attribuées aux catégories, il est important de noter que ces chiffres ne représentent pas nécessairement des quantités mesurables de manière équidistante. Les variables quantitatives nominales n'ont pas d'ordre spécifique entre les catégories, tandis que les variables quantitatives ordinales présentent un ordre, mais l'écart entre les valeurs n'est pas défini ou significatif du point de vue quantitatif.

1.2.5 Variable Statistique quantitative

Définition

— Un caractère est dit quntitatif lorsque ses différentes modalités sont des valeurs numériques (c'est à dire sont mesurables). Dans ce cas, le caractère est appelé variable quantitative.

— Les variables statistiques quantitatives peuvent être divisées en deux catégories principales en fonction de la nature de leurs valeurs : discrètes et continues.

1. **Variable statistique quantitative discrète** : Les variables quantitatives discrètes prennent des valeurs distinctes et comptables. Les valeurs ne peuvent pas être subdivisées en unités plus petites.
2. **Variable statistique quantitative continue :** Les variables quantitatives continues peuvent prendre une infinité de valeurs (ou toutes) dans un intervalle spécifié. Les valeurs peuvent être subdivisées en unités plus petites sans limite.

Exemples

— Nombre de frères et sœurs : Cette variable mesure le nombre exact de frères et sœurs d'une personne, prenant des valeurs entières ;

— Nombre de pièces dans une maison : Cette variable représente le nombre total de pièces dans une maison, telles que les chambres, la cuisine, le salon, etc.. ;

— Nombre d'étudiants dans une classe : La taille d'une classe est une variable discrète, car elle est déterminée par un nombre entier d'élèves ;

— Nombre d'employés dans une entreprise : Le nombre d'employés travaillant pour une entreprise est une variable discrète ;

— Nombre de livres dans une bibliothèque : La taille d'une bibliothèque peut être mesurée en comptant le nombre de livres, ce qui représente une variable discrète ;

— Nombre de téléphones dans un foyer : Le nombre de téléphones dans un foyer est une variable discrète, car il est dénombrable en nombres entiers ;

— Nombre de cours suivis par un étudiant : Cette variable mesure le nombre exact de cours qu'un étudiant suit au cours d'un semestre ou d'une année académique ;

— Nombre de réponses correctes dans un questionnaire à choix multiples : Lors d'un test à choix multiples, le nombre de réponses correctes est une variable discrète, car il est dénombrable en nombres entiers ;

sont des exemples de variables statistiques quantitatives discrètess

Exemples

- Âge : L'âge est une variable quantitative continue, mesurée en années ;
- Revenu annuel : Le revenu annuel est une variable quantitative continue qui représente la somme d'argent gagnée au cours d'une année ;
- Taille : La taille d'une personne est une variable quantitative continue, mesurée en centimètres ou en mètres ;
- Poids : Le poids d'un objet ou d'une personne est une variable quantitative continue, mesurée en kilogrammes ou en livres ;
- Température : La température est une variable quantitative continue, mesurée en degrés Celsius ou Fahrenheit ;
- Nombre d'enfants dans une famille : Le nombre d'enfants dans une famille est une variable quantitative discrète, car il ne peut prendre que des valeurs entières ;
- Score à un test : Le score obtenu à un test est une variable quantitative, mesurée sur une échelle numérique, représentant la performance d'un individu ;
- Durée d'un événement : La durée d'un événement, comme la durée d'une réunion ou la durée d'un processus, est une variable quantitative continue mesurée en unités de temps ;
- Vitesse : La vitesse d'un objet en mouvement est une variable quantitative continue, mesurée en unités de distance par unité de temps (par exemple, kilomètres par heure) ;

Remarque

Si la variable est discrète mais prend beaucoup de valeurs, on la traite comme une variable continue.

1.2.6 Série et distribution statistique

Définition

On appelle série statistique, une liste de N observations d'un caractère sur la population Ω.

Une série statistique quantitative est donc une liste de valeurs de la variable.

Exemples

1. Le nombre d'appels téléphonique réalisé au moyen d'un GSM au cours d'une journée pour un échantillon de 15 personnes est :

   $$0,\ 1,\ 0,\ 0,\ 1,\ 2,\ 1,\ 3,\ 1,\ 0,\ 2,\ 2,\ 3,\ 2,\ 1.$$

   Cette suite de valeurs constitue une série statistique quantitative discrète. (Remarquer qu'elle n'est pas ordonnée !).

2. Le groupe sanguin de 15 étudiants est

   $$O, A, A, AB, AB, O, O, O, AB, B, B, B, A, AB, B.$$

   Cette suite de valeurs constitue une série statistique qualitative nominale.

1.3 Distributions d'effectifs et fréquences

Soit Ω une population de taille N et soit X une variable statistique quantitative discrète (ou qualitative) définie sur Ω dont les valeurs possibles sont rangées dans l'ordre croissant, sont $x_1, x_2 \ldots, x_p$. L'ensemble des valeurs prises par la variable X est noté $X(\Omega) = \{x_1, x_2, \ldots, x_p\}$.

1.3.1 Effectifs et fréquences

Définition

L'effectif (ou fréquence absolue) d'une valeur x_i (ou modalité x_i) est le nombre n_i d'individus présentant cette modalité ou prenant cette valeur.

Exemple

La série statistique quantitative indiquant le nombre d'appels téléphonique réalisés au moyen d'un GSM au cours d'une journée pour un échantillon de 100 personnes est donné dans le tableau suivant :

Nombres d'appels	0	2	3	5	7	8	9
Effectifs	15	20	25	15	10	8	7

L'effectif de la valeur 0 est 15, l'effectif de la valeur 7 est 10, etc...

Exemple

La série statistique qualitative indiquant le groupe sanguin de 15 étudiant se la filière Web marketing est :

Groupe Sanguin	A	B	AB	O	...
Effectifs	2	3	6	4	...

L'effectif de la modalité A est 2, l'effectif de la modalité O est 4 etc...

On constate au passage la remarque suivante :

Remarque

La somme de tous les effectifs d'une série statistique est l'effectif total

$$\sum_{i=1}^{p} n_i = n_1 + n_2 + \ldots n_p = N.$$

Définition

La fréquence d'une valeur x_i notée f_i est le rapport

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

Le pourcentage de x_i est la quantité

$$100 \times f_i.$$

Exemple

La série statistique qualitative suivante :

Groupe Sanguin	A	B	AB	O	Somme
Effectifs	$150 = n_1$	$70 = n_2$	$180 = n_3$	$100 = n_4$	500
Fréquences	$0.3 = f_1$	$0.14 = f_2$	$0.36 = f_3$	$0.2 = f_4$	100%

— Population : les 500 étudiants de la faculté des sciences.

— Caractère étudié : Groupe sanguin.

— Modalité : A, B, AB, O.

— Effectifs :

(a) Le nombre d'étudiants ayant le groupe sanguin A ; $n_1 = 150$.

(b) Le nombre d'étudiants ayant le groupe sanguin B ; $n_2 = 70$.

(c) Le nombre d'étudiants ayant le groupe sanguin AB ; $n_3 = 180$.

(d) Le nombre d'étudiants ayant le groupe sanguin O ; $n_4 = 100$.

— L'effectif total est $N = n_1 + n_2 + n_3 + n_4 = 500$.

— La fréquence de la modalité A ; $f_1 = \dfrac{150}{500} = 0.3$.

— La fréquence de la modalité B; $f_2 = \dfrac{70}{500} = 0.14$.

— La fréquence de la modalité AB; $f_3 = \dfrac{180}{500} = 0.36$.

— La fréquence de la modalité O; $f_4 = \dfrac{100}{500} = 0.2$.

On vérifie sur cette exemple que $f_1 + f_2 + f_3 + f_4 = 1$

Remarque

Pour tout $1 \leq i \leq p$ on a : $0 \leq f_i \leq 1$ et nous avons

$$\sum_{i=1}^{p} f_i = f_1 + f_2 + \ldots + f_p = 1.$$

En effet on a :

$$\sum_{i=1}^{p} f_i = f_1 + f_2 + \ldots + f_p = \frac{n_1}{N} + \frac{n_2}{N} + \ldots + \frac{n_p}{N} = \frac{n_1 + n_2 + \ldots + n_p}{N} = \frac{N}{N} = 1.$$

Définition

— L'effectif cumulé croissant (ECC en abrégé) d'une valeur x_i , noté N_i est la somme des effectifs de cette valeur et des valeurs inférieures. C'est à dire

$$N_i = n_1 + n_2 + n_3 + \ldots + n_{i-1} + n_i$$

— L'effectif cumulé décroissant (ECD en abrégé) d'une valeur x_i , est la somme des effectifs de cette valeur et des valeurs supérieures. C'est à dire

$$ECD_i = n_i + n_{i+1} + \ldots + n_{p-1} + n_p$$

Définition

— La fréquence cumulée croissante (FCC en abrégé) d'une valeur x_i , notée F_i est la somme des fréquences de cette valeur et des valeurs inférieures. C'est à dire

$$F_i = f_1 + f_2 + f_3 + \ldots + f_{i-1} + f_i$$

— La fréquence cumulée décroissante (FCD en abrégé) d'une valeur x_i , est la somme des fréquences de cette valeur et des valeurs supérieures. C'est à dire

$$FCD_i = f_i + f_{i+1} + \ldots + f_{p-1} + f_p$$

Il est commode de présenter une série statistique sous forme de tableau, contenant les valeurs possibles de cette variable, rangées dans un ordre l'ordre croissant, et pour chacune de ces valeurs l'effectif (où fréquence) correspondant (correspondante).

Valeur de la variable	Effectif n_i	Fréquences f_i	Effectifs cumulés croissants N_i	Fréquences cumulées croissantes F_i
x_1	n_1	f_1	n_1	$F_1 = f_1$
x_2	n_2	f_2	$n_1 + n_2$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	$n_1 + n_2 + \ldots + n_i$	$F_i = f_1 + f_2 + \ldots + f_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_p	f_p	$n_1 + n_2 + \ldots + n_p$	$F_p = f_1 + \ldots + f_p = 1$

Exemple

Le nombre d'enfants dans 1000 ménages ayant au moins un enfants est indiqué dans le tableau suivant :

Nombres d'enfants	Effectif n_i	ECC N_i	Fréquence f_i	FCC F_i
1	282	282	0.282	0.282
2	273	555	0.273	0.555
3	248	803	0.248	0.803
4	120	923	0.120	0.923
5	35	958	0.035	0.958
6	42	1000	0.042	1

1.3.2 Groupement de données en classes, Variable statistique continue

Pour une série statistique qui présente un grand nombre de valeurs distincts, on a intérêt à grouper les données. Cela signifie qu'au lieu d'énumérer les valeurs de la variable, on partitionne le domaine de celle-ci en intervalle appelés classes.

Exemple

On désire étudier la taille des étudiants de la faculté Pluridisciplinaire de nador. Pour cela, on range les tailles en classes : [155, 160[, [160, 165[, [165, 170[, [170, 175[, [175, 180[, [180, 185[, [185, 190[

Définition

Pour une classe $c_i = [a_{i-1}, a_i[$:

1. a_{i-1} et a_i s'appellent les bornes ou la les limites de la classe c_i.
2. $m_i = \dfrac{a_{i-1} + a_i}{2}$ s'appelle le centre (ou le milieu) de la classe c_i.
3. L'écart $a_i - a_{i-1}$ est l'amplitude (ou la longueur) de la classe c_i.
4. L'effectif n_i de la classe c_i , est le nombre d'individus pour lesquels la variable statistique prend une valeur dans l'intervalle c_i .
5. La fréquence f_i de la classe c_i , est le rapport $f_i = \dfrac{n_i}{N}$
6. La fréquence cumulée croissante F_i de la classe c_i est la somme des fréquence de cette classe et des classes précédentes. C'est à dire : $F_i = f_1 + f_2 + \ldots + f_i$.
7. La densité (ou fréquence unitaire) de la classe c_i est le rapport $d_i = \dfrac{f_i}{a_i - a_{i-1}}$.
8. La distribution des fréquences d'une variable est un tableau contenant les classes de cette variable et pour chaque classe, la fréquence correspondante.

Remarque

Pour avoir une "étude homogène" de la série, le nombre de classe ne doit pas être ni trop grand ni trop petit et souvent les classes sont de même amplitude.

Exemple

Le revenu mensuel d'un groupe de 200 ingénieurs se répartissent comme suit en l'an 2022.

Revenu en Dh	effectifs n_i	Fréquences f_i	Fréquence cumulées F_i	milieu m_i	densité d_i
[8000, 10000[	30	15%	0.15	9000	0.15/2000
[10000, 12000[	37	18.5%	0.335	11000	0.185/2000
[12000, 14000[	40	20%	0.535	13000	0.2/2000
[14000, 16000[	16	8%	0.615	15000	0.08/2000
[16000, 18000[	12	6%	0.675	17000	0.06/2000
[18000, 20000[	10	5%	0.725	19000	0.05/2000
[20000, 22000[	20	1%	0.825	21000	0.1/2000
[22000, 24000[	18	9%	0.915	23000	0.09/2000
[24000, 26000[	17	8.5%	1	25000	0.085/2000
La somme	200	100%			

1.4 Représentation graphique

Il est souvent utile de représenter graphiquement une série statistique. Un graphique permet de visualiser le comportement de la varaible statistique étudiée.

1.4.1 Diagramme en secteurs

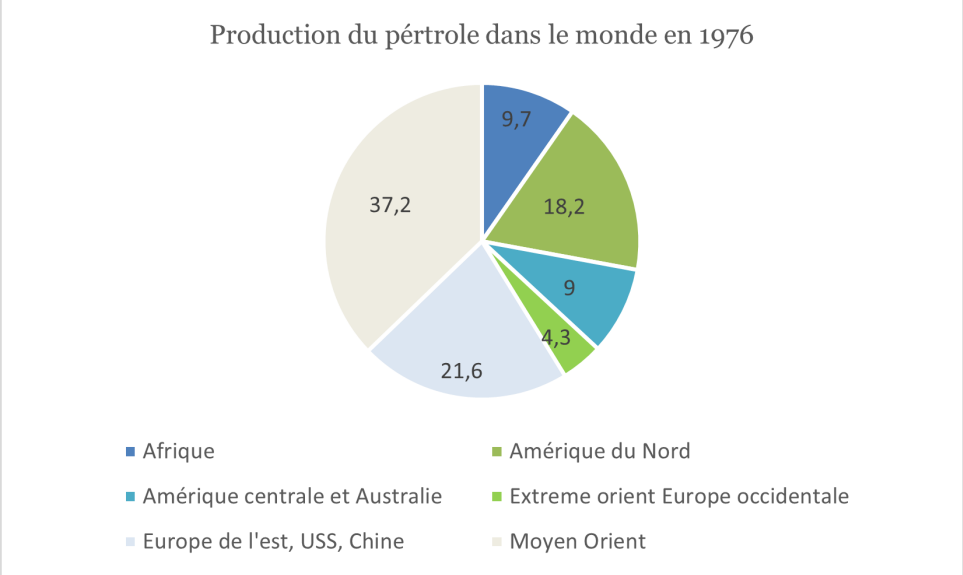
Sur un cercle, on présente une modalité par un secteur angulaire dont la valeur de l'angle θ_i est proportionnelle à l'effectif (ou à la fréquence) de la modalité c'est à dire :

$$\theta_i = 360 \times f_i.$$

Exemple

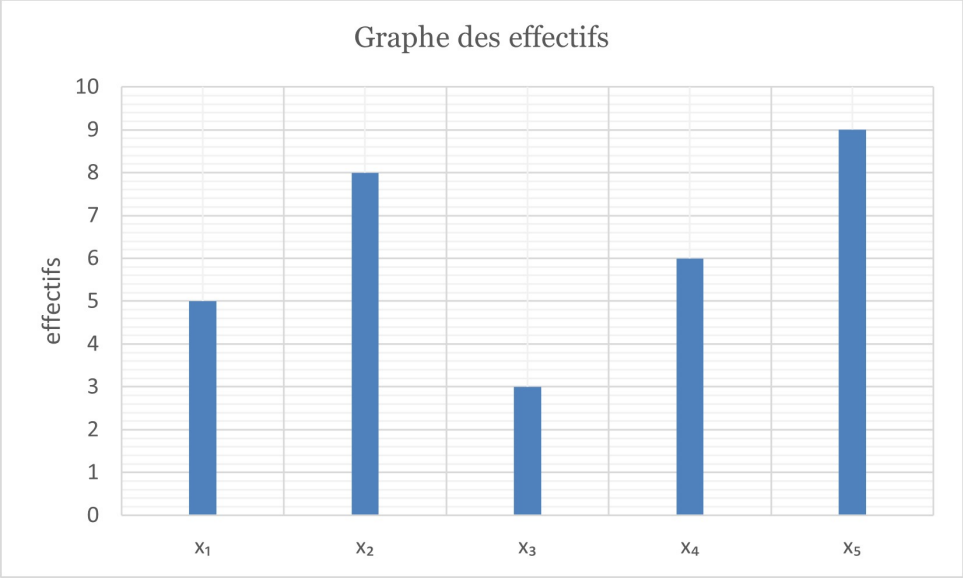
Production du pétrole dans le monde en 1976.

Afrique	Amérique du Nord	Amérique Centrale et Australie	Extrême orient Europe Occiden-tale	Europe de l'est URSS Chine	Moyen Orient
9.7%	18.2%	9%	4.3%	21.6%	37.2%



1.4.2 Digramme en bâtons

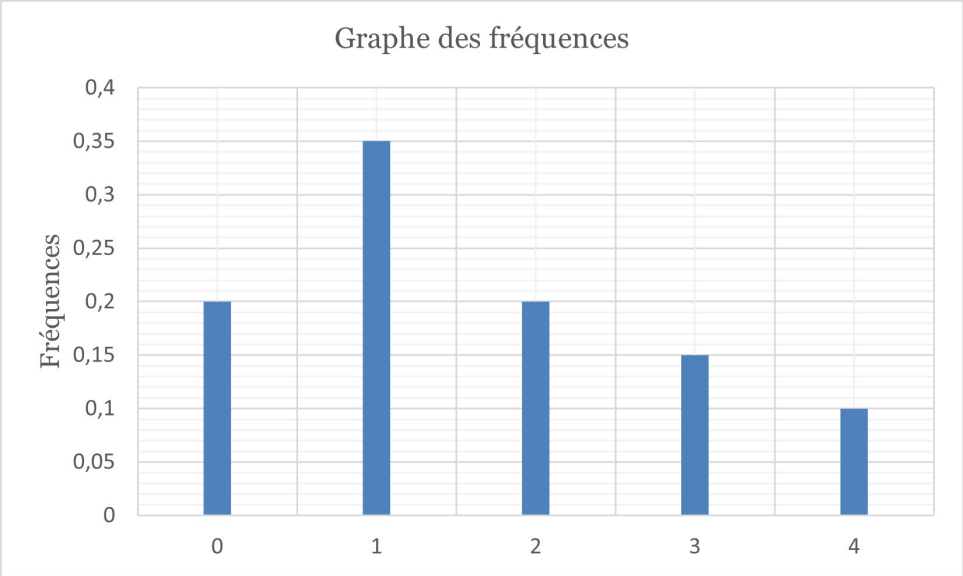
Ce diagramme consiste à porter les valeurs observées x et à tracer en regard de chacune d'elles un segment vertical de longueur égale à son effectif ou sa fréquence. Ce diagramme est utilisé pour les variables qualitatives nominales, ordinales et quantitatives discrètes.



Exemple

Soit X la variable statistique associant à chaque étudiant le nombre de livres empreintés à la bibliothèque.

Nombres d'ouvrages	0	1	2	3	4	Somme
Effectifs	20	35	20	15	10	$N = 100$
Fréquences	0.2	0.35	0.2	0.15	0.1	1

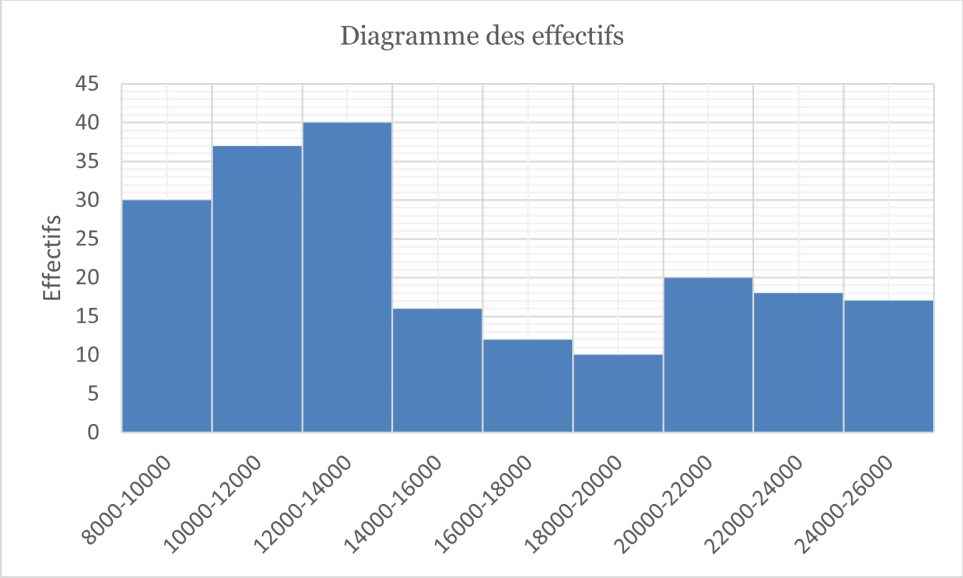


1.4.3 Histogramme

Il s'agit d'un diagramme composé de recctangles dont les bases sont les classes de la variable statistique et dont les surfaces sont proportionnelles aux fréquences de ces classes. (Les hauteurs des rectangles sont les densités des classes). Ce diagramme est utilisé pour les variables quantitatives continues.

Exemple

Reprenons l'exemple des salires de 200 ingénieurs



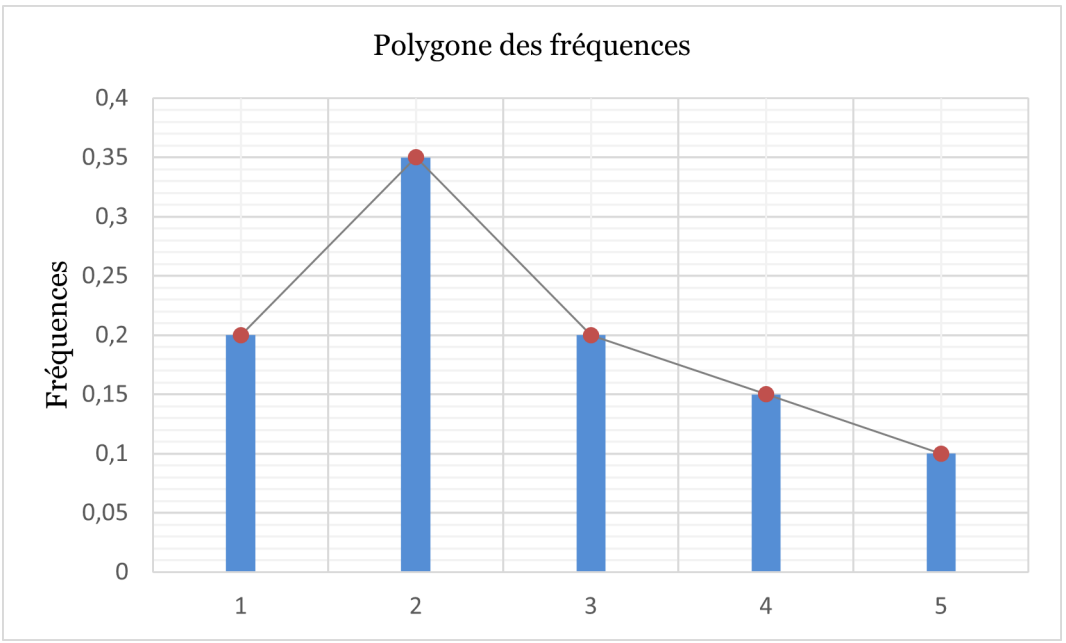
1.4.4 Polygone des fréquences ou des effectifs

Cas d'une varibale statistique quantitative discrète

Ce polygone s'obtient en joignant par un segment de droite les extrémités des segments voisins des diagrammes en bâtons pour les fréquences. C'est une ligne polygonale joignant les points $(x_1, f_1), (x_2, f_2), \ldots, (x_p, f_p)$.
Si on remplace les fréquences f_i par les effectifs n_i, la ligne polygonale obtenue en joignant les points $(x_1, n_1), (x_2, n_2), \ldots, (x_p, n_p)$ s'appelle polygone des effectifs.

Exemple

Reprenons l'exemple du nombre d'ouvrages associés aux étudiants :

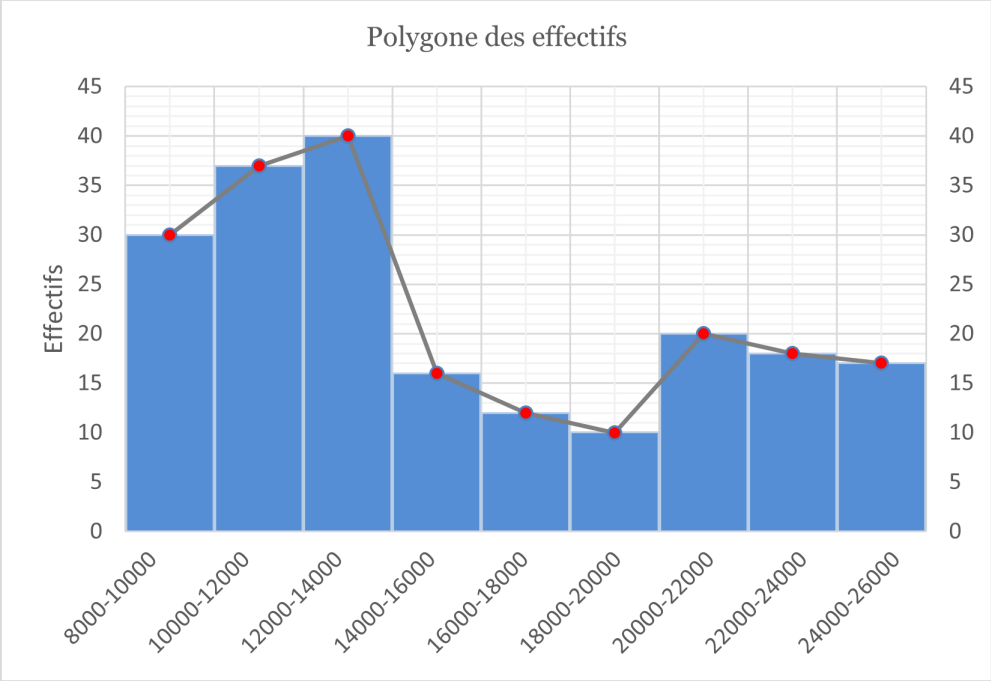


Cas d'une varibale statistique quantitative discrète

Ce polygone s'obtient en joignant par des segments de droite les milieux des bases supérieures des rectangles successifs.

Exemple

Reprenons l'exemple des salaires de 200 ingénieurs :



1.5 Fonction de répartition des variables statistiques quantitatives

1.5.1 Fonction de répartition d'une variable statistique discrète

Définition

— La fonction de répartion d'une varibale statistique discrète X notée F_X est définie de la manière suivante pour tout $x \in \mathbb{R}$, (on suppose que la variable statistique X prend les valeurs $x_1, x_2, \ldots, x_p$, avec $x_1 < x_2 < \ldots < x_p$

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_1 = f_1 & \text{si } x_1 \leq x < x_2 \\ F_2 = f_1 + f_2 & \text{si } x_2 \leq x < x_3 \\ \vdots & \vdots \\ F_i = f_1 + f_2 + \ldots + f_i & \text{si } x_{i-1} \leq x < x_i, i = 1, 2, \ldots, p \\ \vdots & \vdots \\ F_p = f_1 + f_2 + \ldots + f_p = 1 & \text{si } x \geq x_p \end{cases}$$

— Le graphique de la fonction de répartition F_X est appelé le polygone des fréquences cumulées.

Représentation graphique de la fonction de répartition

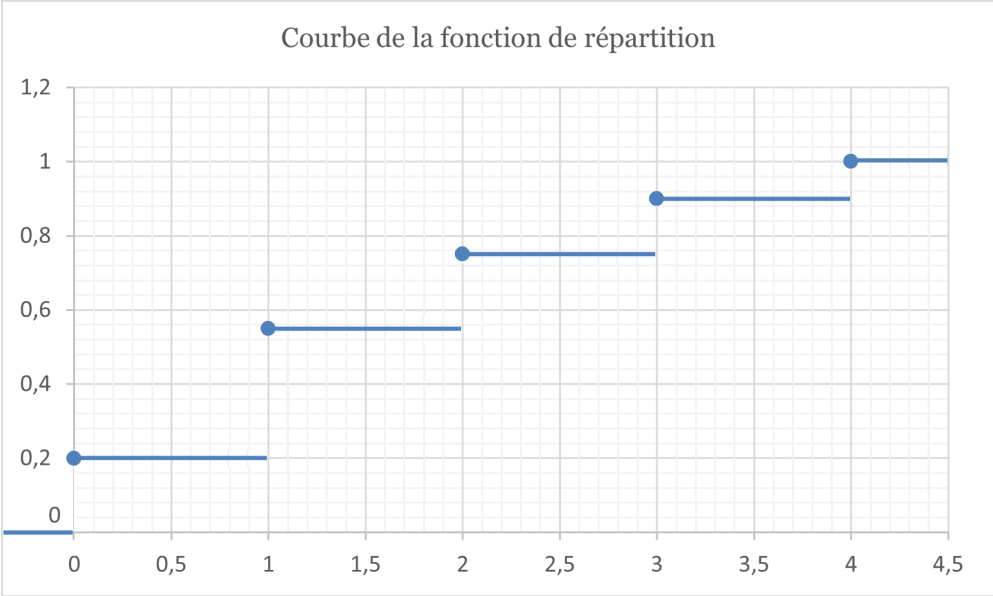
Exemple

Reprenons l'exemple du nombre d'ouvrages associés aux étudiants

Nombre d'ouvrage	Effectifs	Fréquences	Fréquences cumulées
0	20	0.2	0.2
1	35	0.35	0.55
2	20	0.2	0.75
3	15	0.15	0.9
4	10	0.1	1
La somme	100	1	

La fonction de répartition de cette variable est donnée par :

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ F_1 = 0.2 & \text{si } 0 \leq x < 1, \\ F_2 = 0.55 & \text{si } 1 \leq x < 2, \\ F_3 = 0.75 & \text{si } 2 \leq x < 3, \\ F_4 = 0.9 & \text{si } 3 \leq x < 4, \\ F_5 = 1 & \text{si } x \geq 4 \end{cases}$$



Il en résulte de la définition que la fonction de réparttitrion possède les proprétés suivantes :

Proposition

1. La fonction F est définie sur $\mathbb{R}$ et à valeurs dans [0,1].
2. F est en escalier.
3. F est croissante.
4. F est continue à droite et admet une limite à gauche en tout point $x \in \mathbb{R}$.
5. $\lim\limits_{x \to +\infty} F(x) = 1$, $\lim\limits_{x \to -\infty} F(x) = 0$.

1.5.2 Fonction de répartition (ou courbe cumulative) d'une variable aléatoire continue

Définition

On appelle fonction de répartition d'une série statistique, la fonction F, définie pour tout $x \in \mathbb{R}$ par

$$\begin{aligned} F(x) &= \textit{fréquence des observations } \leq x \\ &= \textit{proportion des observations } \leq x. \end{aligned}$$

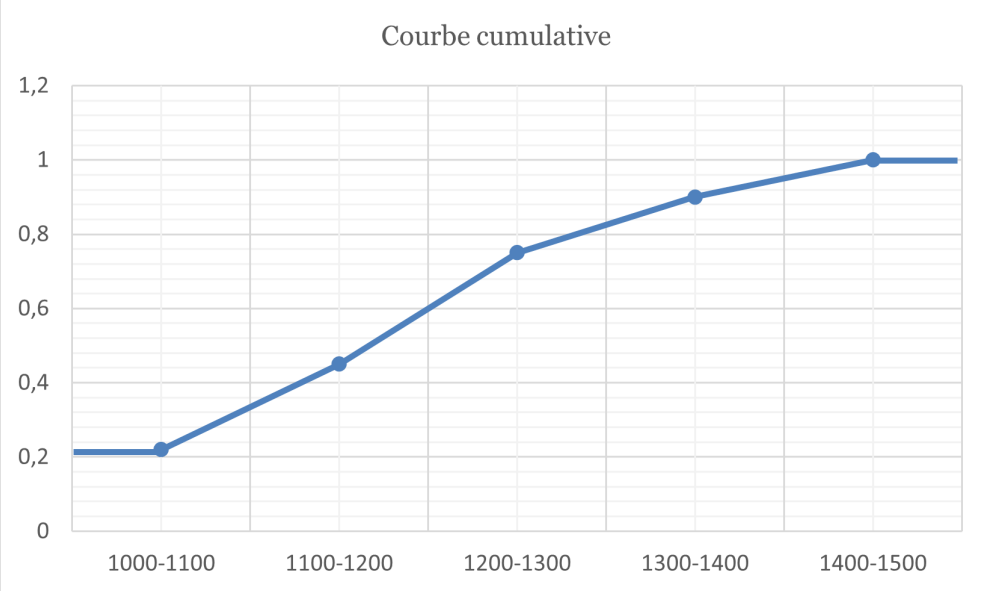
La courbe cumulative s'obtient en joignant les points d'abscisses : la borne supérieure de la classe, et en ordonnée : la fréquence cumulée correspondante. Autrement dit en joignant les points de coordonnées (a_i, F_i).

Exemple

Considérons la série statistique indiquant les salaires des 100 ouvriers d'un atelier. On a le tableau suivant :

Les salaires	Effectifs	Fréquences	Fréquences cumulées
[1000,1100[	22	0.22	0.22
[1100,1200[	23	0.23	0.45
[1200,1300[	30	0.3	0.75
[1300,1400[	15	0.15	0.9
[1400,1500[	10	0.1	1
La somme	100	1	

La courbe des fréquences cumulées permet de lire, pour toute valeure de x, le pourcentage des ouvriers dont le salaire est $\leq x$. Par exemple, 35% des ouvriers ont un salaire $\leq$ 1150Dh.



Remarque

Si $a < b$ alors $F(b) - F(a)$ = proportion des observations se trouvant dans $[a, b[$.

Proposition

$$F(x) = F_{i-1} + \frac{F_i - F_{i-1}}{a_i - a_{i-1}}(x - a_{i-1}),\ x \in [a_{i-1}, a_i[$$

où $F_i = F(a_i)$ est la fréquence cumulée de a_i,

$$F_i = F(a_i) = f_1 + f_2 + \ldots + f_i.$$

Démonstration. : Pour démontrer la proposition, on va utiliser l'interpolation linéaire. On connaut $F(a_i) = F_i$ et $F(a_{i-1}) = F_{i-1}$. Donc, la droite passant par les points $(a_{i-1}, F_{i-1}), ((a_i, F_i))$ a pour équation pour $x \in [a_{i-1}, a_i, [$

$$F(x) = F_{i-1} + \frac{F_i - F_{i-1}}{a_i - a_{i-1}}(x - a_{i-1})$$

La propsition suivante résume les propriétés de la fonction de répartition : □

Proposition

1. F est une fonction définie sur $\mathbb{R}$.
2. F est fonction continue sur $\mathbb{R}$.
3. F est une fonction croissante.
4. $\forall x \in 0 \leq F(x) \leq 1$.

Remarque

Étant donné l'importance des données, il est parfois utile de regrouper les modalités d'une variable discrète en sous-classes afin de la transformer en une variable continue

Exemple

Le responsable de stock d'un atelier a noté, au cours de 98 jours de travail, le nombre de boulon d'un certain type utilisé dans son atelier. Les données brutes sont données cidessous.

72	51	56	95	68	66	77	81	83	75
41	79	92	78	85	55	104	76	80	61
65	70	83	92	88	59	75	75	81	69
71	96	101	87	65	74	68	73	78	68
73	86	84	51	85	75	79	90	68	71
75	74	81	64	88	78	77	66	91	75
69	73	82	75	76	71	74	96	72	74
102	74	80	82	86	78	87	61	80	78
48	68	71	66	59	92	77	76	81	70
85	77	68	82	78	75	91	77		

Vue l'importance des données, on a intérêt à les grouper sous forme de classe. Ici on va prendre des classes de longueurs 10. Les résultats regroupés sont résumes dans le tableau suivant :

Classes	Effectifs	Fréquences	Effectifs cumulés croissants	Fréquences cumulées croissantes
[40, 50[	2	2.04	2	2.04
[50, 60[	6	6.12	8	8.16
[60, 70[	16	16.33	24	24.49
[70, 80[	40	40.82	64	65.31
[80, 90[	22	22.45	86	87.76
[90, 100[	9	9.18	95	96.94
[100, 110[	3	3.06	98	100

1.5.3 Quelques règles et précautions

Une représentation graphique est un outil de communication : il a pour objectif de montrer des données de façon claire et adéquate. Plusieurs règles et précautions doivent être présentes à l'esprit de ceux qui construisent, analysent et utilisent de tels outils. Mentionnonsen quelquesunes :

1. Le graphique doit contenir le maximum d'informations utiles (légendes, sources, valeurs numériques, ...) ; il doit être compris par luimême, sans que l'on soit obligé de recourir à la lecture d'un texte explicatif.
2. Les informations ne doivent pas être cachées par des lignes, dessins ou mentions inutiles. Elles ne doivent pas être déformées en raison d'un choix des unités peu judicieux ; ceci implique en particulier qu'il ne faut pas nécessairement qu'une échelle montre toutes les graduations à partir de zéro ! Par ailleurs, comme nous l'avons indiqué implicitement lors de notre étude du diagramme en bâtons, on s'efforce en général de construire des graphiques inscriptibles dans un carré ou dans un rectangle peu allongé par rapport à l'un des axes afin de ne pas influencer l'utilisateur. 3 Il n'est pas nécessaire de vouloir représenter des situations simples par des graphiques sophistiqués.
3. Il ne faut pas interrompre une échelle sans le signaler explicitement et être certain que cela n'a pas d'influence sur l'interprétation du graphique.
4. La façon de présenter graphiquement un phénomène statistique doit mettre en évidence ses caractéristiques essentielles.
5. Il ne faut comparer des graphiques que si l'on a choisi des unités communes sur les axes. Un esprit critique et une pratique minimale permettent d'éviter les écueils d'un mauvais graphique.

Type de variable	Diagramme circulaire	Diagramme en bâtons	Histogramme	Polygone des fréquences	Fonction de répartition
Qualitative nominale	✓	✓(non ordonné)	✖	✖	✖
Qualitative ordinale	✖	✓(ordonné)	✖	✖	✖
Quantitative discrète	✖	✓(ordonné)	✖	✓	✓
Quantitative continue	✖	✖	✓	✓	✓