MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE BATNA 2
FACULTE DE TECHNOLOGIE
DEPARTEMENT DE SCIENCE TECHNOLOGIQUE
2éme Année Socle Commun ST

Résumé de cours de Statistiques

Année universitaire 2021/2022

CHAPITRE

1

NOTION DE BASE ET VOCABULAIRE STATISTIQUE

La **statistique** est l'ensemble des techniques permettant de traiter des données issues de l'observation de phénomènes. Lobjet de la statistique est de **rassembler, organiser, analyser, interpréter,** des observations que l'on peut mesurer ou classer. Elle utilise pour cela des représentations de données sous forme de graphiques, de tableaux et d'indicateurs numériques.

1.1 Concepts de base de la statistique

Une étude **statistique descriptive** s'effectue sur un ensemble qu'il convient de définir d'une manière précise. Cet ensemble appelé **population** (des personnes, des villes, des objets...) dont les éléments sont des **individus**. Généralement, l'étude est réalisée sur une partie restreinte de la **population**, appelée **échantillon**. On s'intéresse à un ou plusieurs aspects sur chaque individu, nommé **variables** ou **caractèe** (situation familiale, salaire...). Un caractère peut prendre de différentes valeurs (pour la situation familiale : célibataire, marié et divorcé).
Il existe deux types de caractères :

1. **quantitatif** : c'est un caractère auquel on peut associer un nombre (poids, taille,). On distingue alors deux types de caractères quantitatifs :
   — discret : c'est un caractère quantitatif qui ne prend qu'un nombre fini de valeurs (le nombre d'enfants d'une famille).
   — continu : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en classes (la taille d'un individu).
2. **qualitatif** : c'est un caractère qu'on ne peut pas mesuré (la profession, la couleur..).

Exemple 1

1. On étudie le nombre d'enfants par famille en sud du pays.
   - Population : Les familles du sud
   - Individu : Chaque famille
   - Caractère : Nombre d'enfants
   - Valeurs possibles : $0;1;2;\ldots$
2. On étudie la marque du téléphone portable utilisé par les étudiants.
   - Population : Les étudiants
   - Individu : Chaque étudiant
   - Caractère : La marque du téléphone portable
   - Valeurs possibles : Condor, Samsung, Oppo, StreamSysteme ...

1.2 Les tableaux statistiques

1.2.1 Caractere quantitatif

Variable statistique discrète

Soient X la variable statistique et $x_1, x_2, \ldots, x_p$ les p valeurs possibles distinctes prises par la variable statistique X (en général si cela est possible, les valeurs x_i sont rangées par ordre croissant).

Définition 1

On appelle **effectif** de la valeur x_i, le nombre de fois que cette valeur se répète dans l'échantillon, noté par n_j. vérifie que

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

Définition 2

On appelle **effectif cumulé** en x_i la somme de tous les effectifs n_j, avec $j \leqslant i$. Ce nombre est noté $\tilde{n}_i$

$$\tilde{n}_i = \sum_{j=1}^{i} n_j = n_1 + n_2 + \ldots + n_i$$

Définition 3

On appelle **fréquence** de la valeur x_i, la proportion de la présence de cette valeur dans l'échantillon, donné par

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

Proposition 1

Soit f_i une fréquence, alors

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

Définition 4

On appelle **fréquence cumulée** en x_i, le nombre noté $\tilde{f}_i$:

$$\tilde{f}_i = \frac{\tilde{n}}{n} = f_1 + f_2 + \ldots + f_i = \sum_{j=1}^{i} f_j$$

Remarque

- La valeur de la fréquence est comprise entre 0 et 1.
- Le pourcentage est une fréquence exprimée en pour cent. Il est égal à $100 \times f_i$.

Exemple 2

Le tableau suivant représente la répartition des élèves d'une classe, de l'école primaire, selon le nombre de frères. Nous avons

- Population : l'ensemble des élèves.
- $n = 20$ la taille de la population.
- Individu : un élève.
- Caractère : le nombre de frères
- $f_3 = \frac{4}{20} = 0.2$ représente 20% d'élèves ont 3 frères.
- $\tilde{f}_3 = \frac{9}{20} = 0.45$ représente 45% d'élèves dont le nombre de frères est inférieure ou égale à 3.

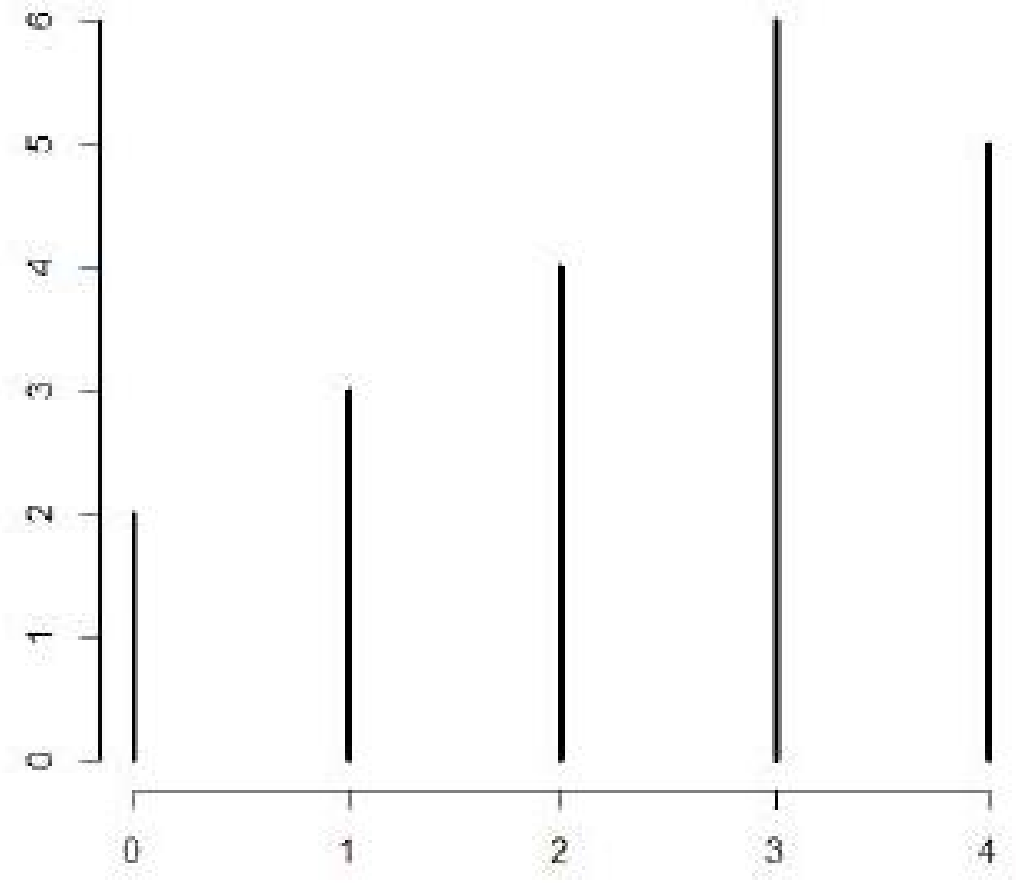
x_i	0	1	2	3	4
n_i	2	3	4	6	5
f_i	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{4}{20}$	$\frac{6}{20}$	$\frac{5}{20}$
$\tilde{n}_i$	2	5	9	15	20
$\tilde{f}_i$	$\frac{2}{20}$	$\frac{5}{20}$	$\frac{9}{20}$	$\frac{15}{20}$	$\frac{20}{20} = 1$

Représentation graphique

La représentation graphique du caractère quantitatif discret est **le diagramme des effectifs, en bâtons**. On obtient le diagramme en bâton des fréqueces par le changement d'échelle sur l'axe des ordonnées.

Exemple 3

Le diagramme des effectifs, en bâton de l'exemple précédent.



Variable statistique continue

Les caractères continus sont ceux qui ont une infinité de modalités.
L'intervalles $[a,b]$ se devise en k sous intervalles distincts $[a_0,a_1[,[a_1,a_2[,\ldots,[a_{k-1},a_k]$, tel que

$$[a,b]=[a_0,a_1[\cup[a_1,a_2[\cup\ldots[a_{k-1},a_k]$$

avec $a=a_0$ et $b=a_k$. Chaque intervalle est appellé classe.

Définition 5

On appelle **effectif** de la classe $[a_{i-1}, a_i[$, le nombre de valeur appartenant à cette classe, noté par n_i.vérifie que

$$\sum_{i=1}^{k} n_i = n$$

Définition 6

On appelle **effectif cumulé** en $[a_{i-1}, a_i[$ la somme de tous les effectifs n_j, avec $j \leqslant i$. Ce nombre est noté $\tilde{n}_i$

$$\tilde{n}_i = \sum_{j=1}^{i} n_j = n_1 + n_2 + \ldots + n_i$$

Définition 7

On appelle **fréquence** de la classe $[a_{i-1}, a_i[$, le raport :

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

avec

$$\sum_{i=1}^{k} f_i = 1$$

Définition 8

On appelle **fréquence cumulée** en $[a_{i-1}, a_i[$, le nombre noté $\tilde{f}_i$:

$$\tilde{f}_i = \frac{\tilde{n}}{n} = \sum_{j=1}^{k} f_j$$

Exemple 4

Les notes du contrôle continue au module statistique, observées un échantillon de 100 étudiants sont données dans le tableau suivant

X	$[0;4[$	$[4;8[$	$[8;12[$
C_i	2	6	10
n_i	30	50	20
f_i	$\frac{30}{100} = 0.3$	$\frac{30}{100} = 0.5$	$\frac{30}{100} = 0.2$
$\tilde{n}_i$	30	80	100
$\tilde{f}_i$	$\frac{30}{100} = 0.3$	$\frac{80}{100} = 0.8$	$\frac{100}{100} = 1$

Le centre d'une classe $[a_{i-1}, a_i[$, noté C_i :

$$C_i = \frac{a_{i-1} + a_{i+1}}{2}$$

Représentation graphique

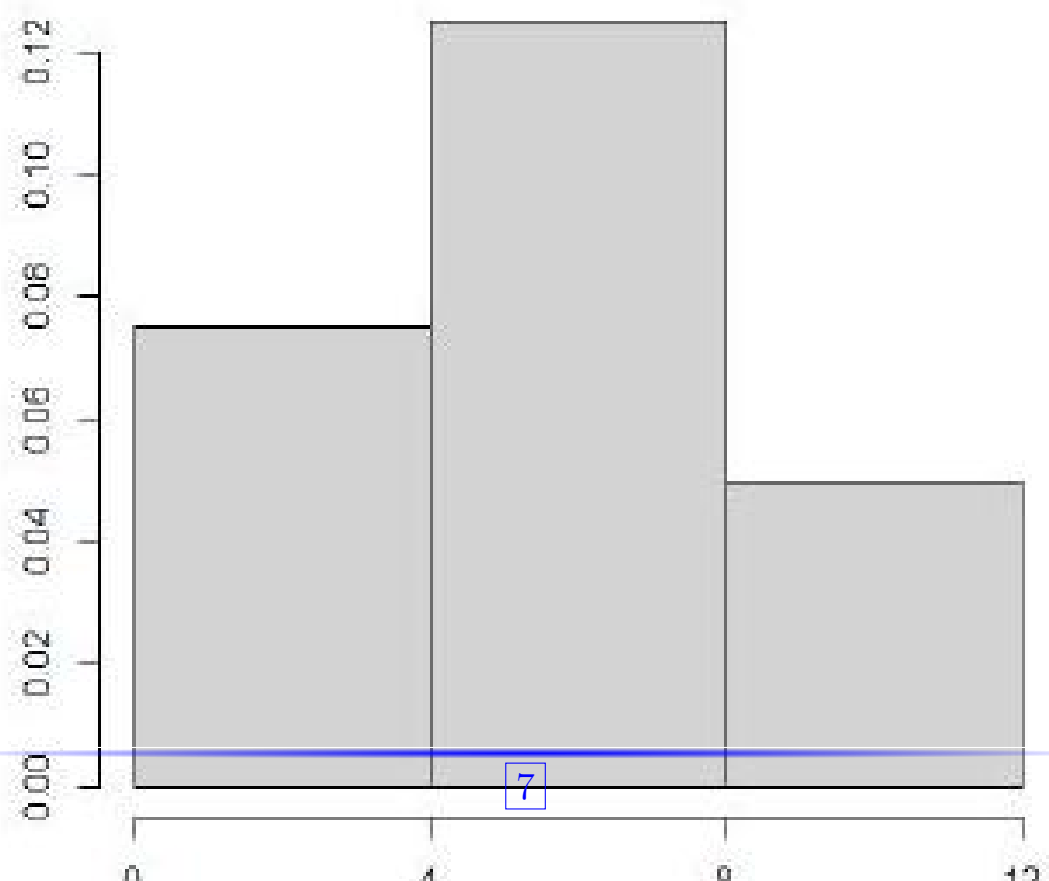
Nous pouvons représenter le tableau statistique par un histogramme (un diagramme comparé d'un ensemble de rectangles contigus(adjacents), chaque rectangle, associé à chaque classe, ayant une surface proportionnelle à l'effectif (ou à la fréquence) de cette classe.). Autrement dit, la hauteur du rectangle correspondant à la classe i est donc donnée par

$$h_i = \frac{n_i}{a_i}$$

alors, la surface du rectangle est égale à l'effectif (ou à la fréquence) de la classe

$$h_i \times a_i = \frac{n_i}{a_i} \times a_i = n_i.$$

La même procédure du calcul pour une surface proportionnelle à la fréquence.



Remarque

— **Le polygone des effectifs** est la ligne brisée joint dans le cas discret les sommets des bâtonnets, et dans le cas continu les milieux des bases.

1.2.2 Caractère qualitatif

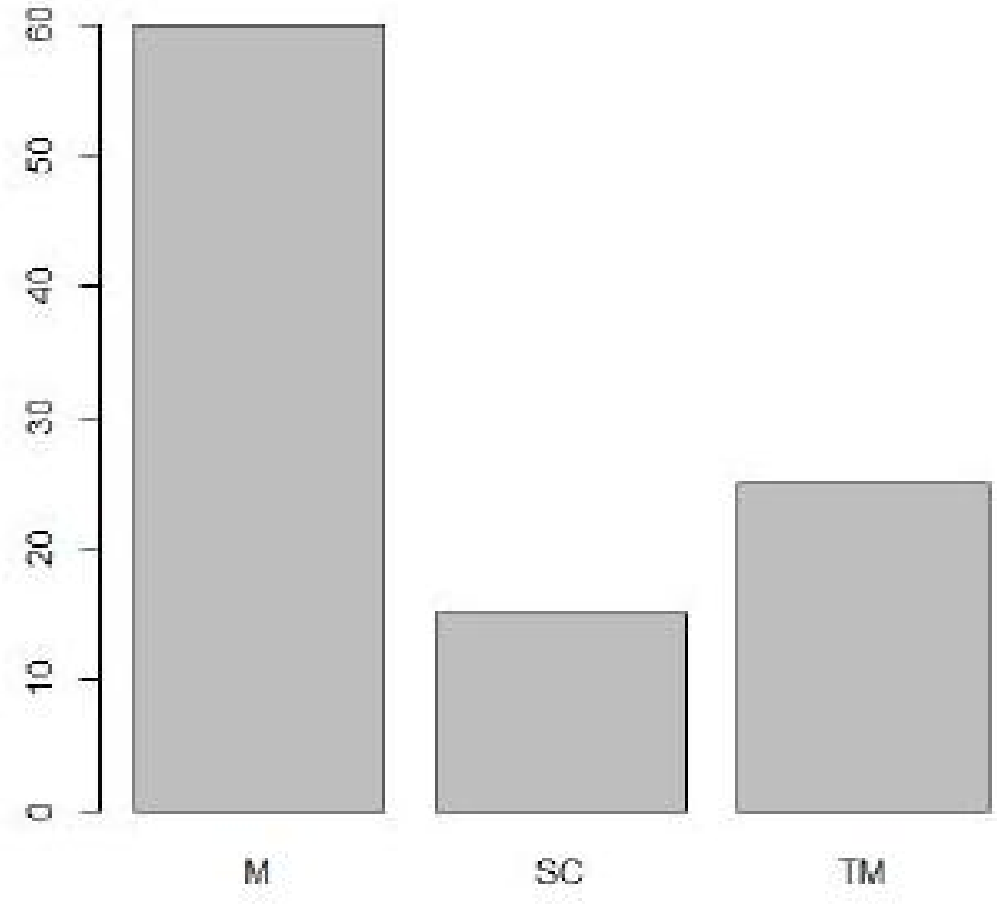
Diagramme en barres

On appelle diagramme à bandes un graphique qui, à chaque modalité de la variable qualitative associe un rectangle de base constante dont la hauteur est proportionnelle à l'effectif.

Exemple 5

Une enquête auprès de 100 étudiants de 2emme année ST, sur les spécialités de leurs BAC, a fourni les résultats suivant :

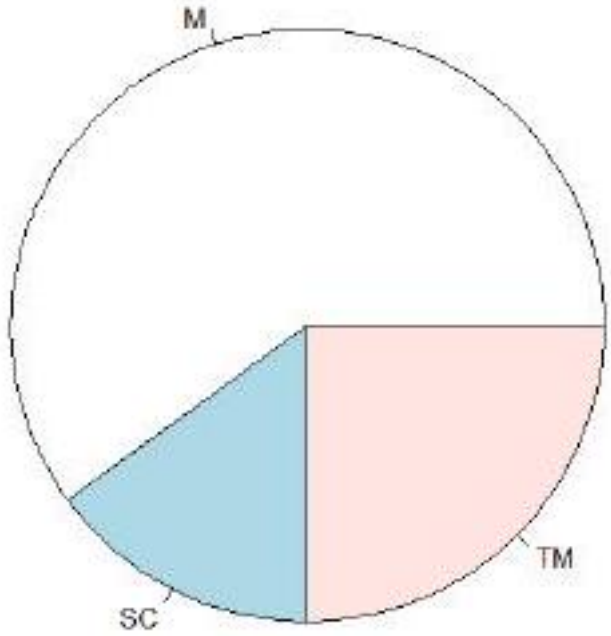
BAC	Nombre d'étudiants n_i
Mathématiques (M)	60
Technique Maths (TM)	25
sciences naturelles (SC)	15



Diagrammes à secteurs

On appelle diagramme à secteurs un graphique qui divise un disque en secteurs angulaires dont les angles au centre sont proportionnels aux effectifs de chaque modalité. Pour une modalité donnée, d'effectif n_i l'angle au centre α_i, correspondant est donné (en degré) par :

$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$$



1.3 Fonction de répartition

Nous avons déjà abordé les distributions cumulées d'une variable statistique. Nous allons dans cette partie exploiter ses valeurs cumulées pour introduire la notion de la fonction de répartition. Cette notion ne concerne que les variables quantitatives. Soit la fonction $F_x : \mathbb{R} \longrightarrow [0,1]$ définie $\text{par}\, F(x) :=$ pourcentage des individus dont la valeur du caractère est $< x$, d'où

$$F(x) = \sum_{x_i < x} f_i$$

de plus, pour tout $i \in 1, \ldots, p$, on a

$$F(x_i) = \tilde{f}_i$$

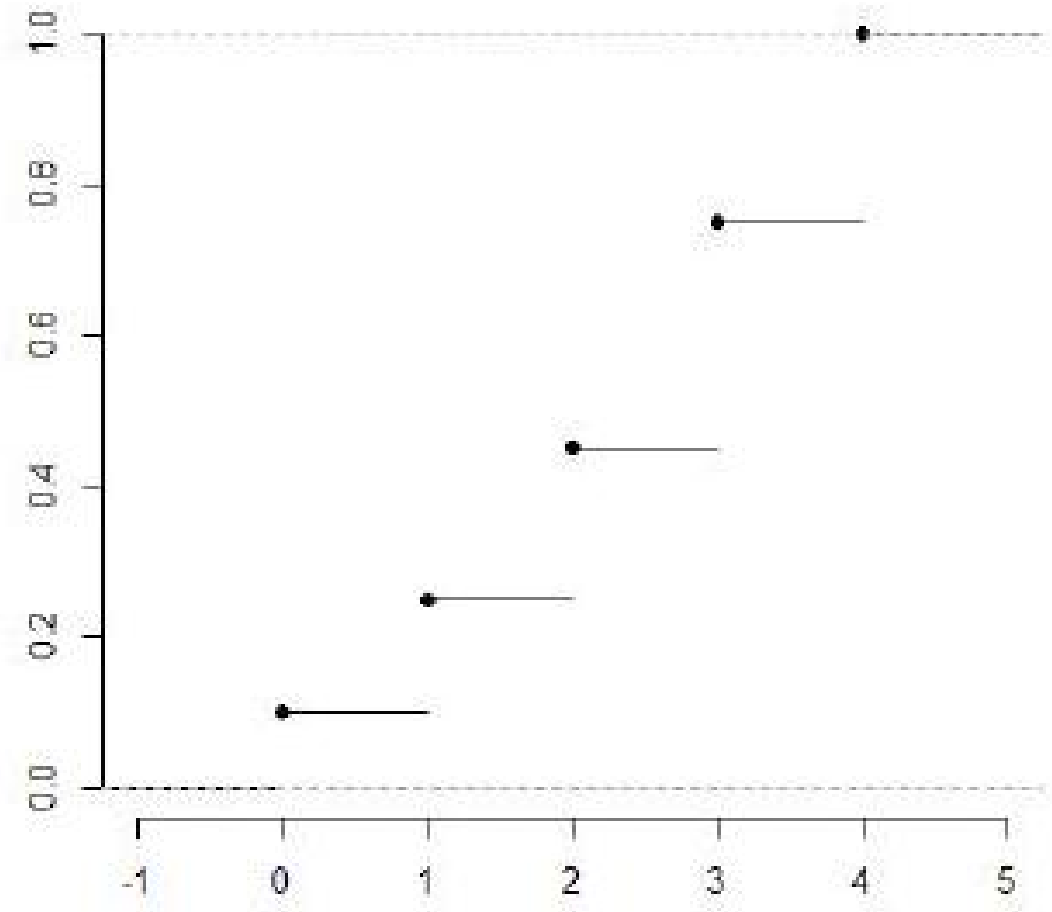
Exemple 6

Reprenons le tableau précedent pour le cas discret

x_i	0	1	2	3	4
n_i	2	3	4	6	5
f_i	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{4}{20}$	$\frac{6}{20}$	$\frac{5}{20}$
$\tilde{n}_i$	2	5	9	15	20
$\tilde{f}_i$	$\frac{2}{20}$	$\frac{5}{20}$	$\frac{9}{20}$	$\frac{15}{20}$	$\frac{20}{20} = 1$

— Pour $x = 2$, alors $F(2) = \sum_{x_i<2} f_i, = \frac{5}{20}$
— Pour $x = 6$, alors $F(6) = \sum_{x<6} f_i, = 1$.

La représentation graphique de la fonction de répartition est donnée comme suite



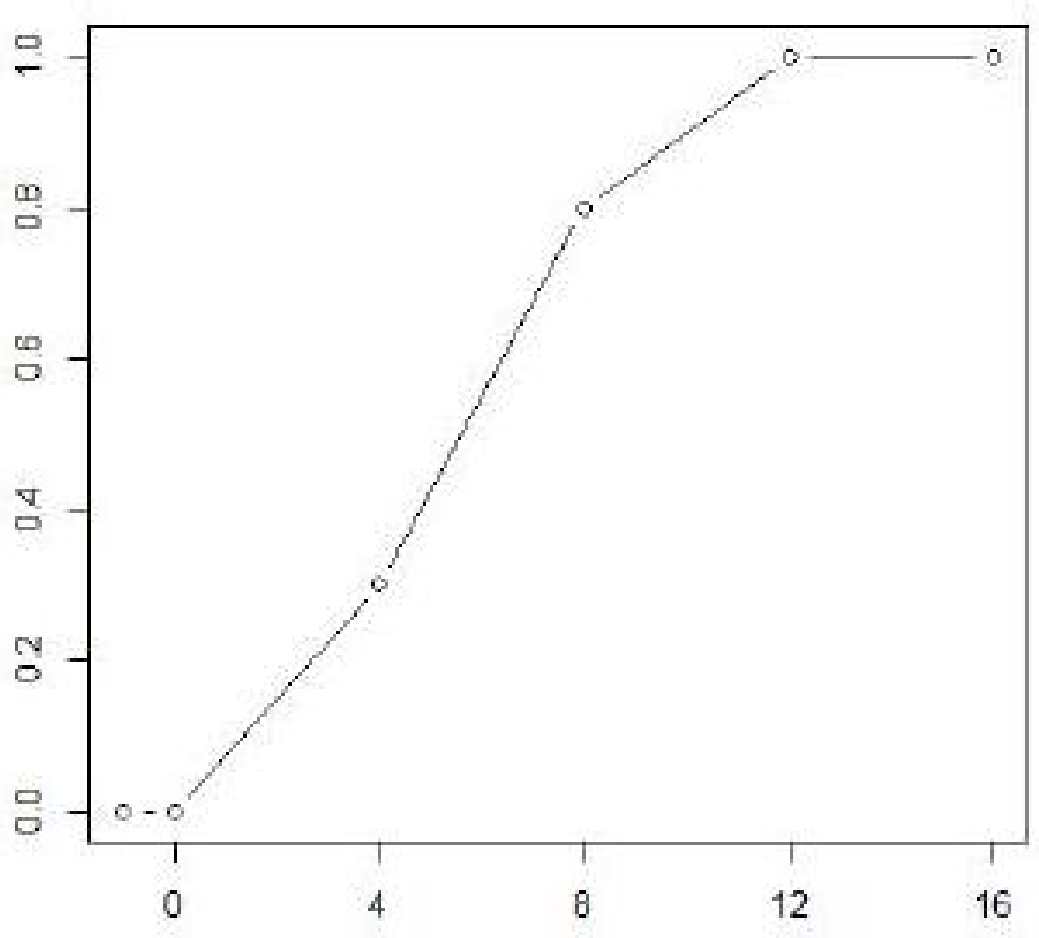
Exemple 7

Reprenons le tableau précedent pour le cas continue

X	$[0;4[$	$[4;8[$	$[8;12[$
C_i	2	6	10
n_i	30	50	20
f_i	$\frac{30}{100} = 0.3$	$\frac{30}{100} = 0.5$	$\frac{30}{100} = 0.2$
$\tilde{n}_i$	30	80	100
$\tilde{f}_i$	$\frac{30}{100} = 0.3$	$\frac{80}{100} = 0.8$	$\frac{100}{100} = 1$

— Pour $x = 0$, alors $F(0) = \sum_{x_i<0} f_i, = 0$,
— Pour $x = 8$, alors $F(8) = \sum_{x_i<8} f_i, = 0.8$
— Pour $x = 16$, alors $F(16) = \sum_{x_i<16} f_i, = 1$

La représentation graphique de la fonction de répartition est donnée comme suite



CHAPITRE

2

STATISTIQUE DESCRIPTIVE À UNE VARIABLE

Nous avons vu que les représentations graphiques sous forme de tableaux ou sous forme de courbes donnent déjà une première analyse des phénomènes étudiés. Cependant il est plus commode de caractériser une série statistique au moyen des paramètres représentatifs de l'ensemble du phénomène. Ces paramètres sont des valeurs numériques qui permettent d'approcher les séries statistiques avec plus ou moins de précisions. On distingue les caractéristiques de tendance centrale et les caractéristiques de dispersion.

2.1 Caractéristique de tendance centrale ou de position

2.1.1 Médiane

Définition 9

La médiane est la valeur de la variable statistique qui partage autant d'individus à gauche qu'à droite dans l'échantillon. Elle correspond au milieu de la distribution. Notée par Me.

Cas d'un caractère discret

On suppose que les n données de la série statistique sont rangées dans l'ordre croissant tel que :

- Si n est **impair** avec $n = 2k+1 (k \in \mathbb{N})$ alors Me est le terme de rang $k+1$.
- Si n est **pair** avec $n = 2k (k \in \mathbb{N})$ alors Me est la demi somme des termes de rang k et $k+1$

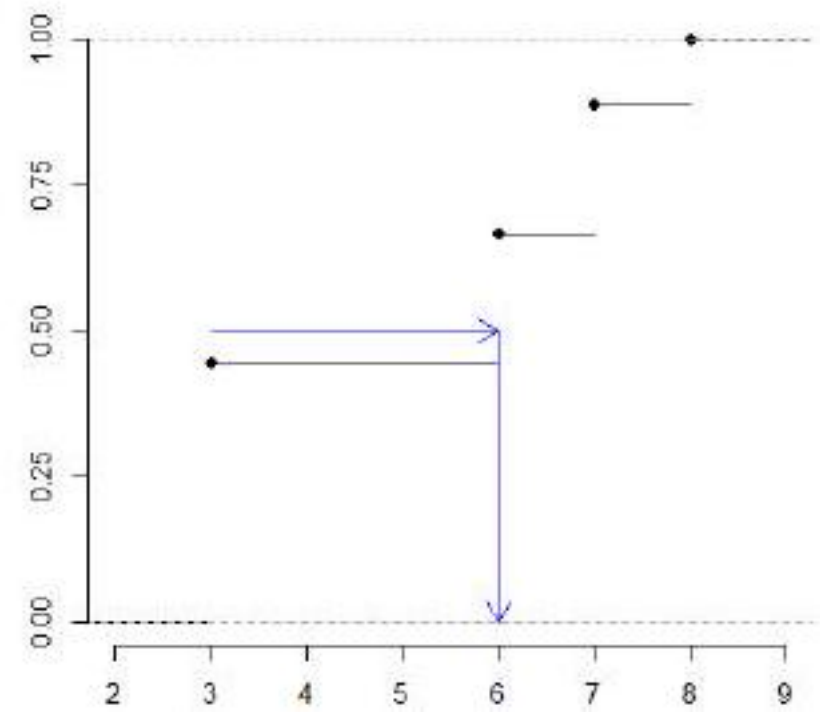
Exemple 8

- Pour la série $3,3,3,3,6,6,7,7,8$, on a $n=9$ d'où $Me=6$ (du rang 5 ou 5ème valeur).

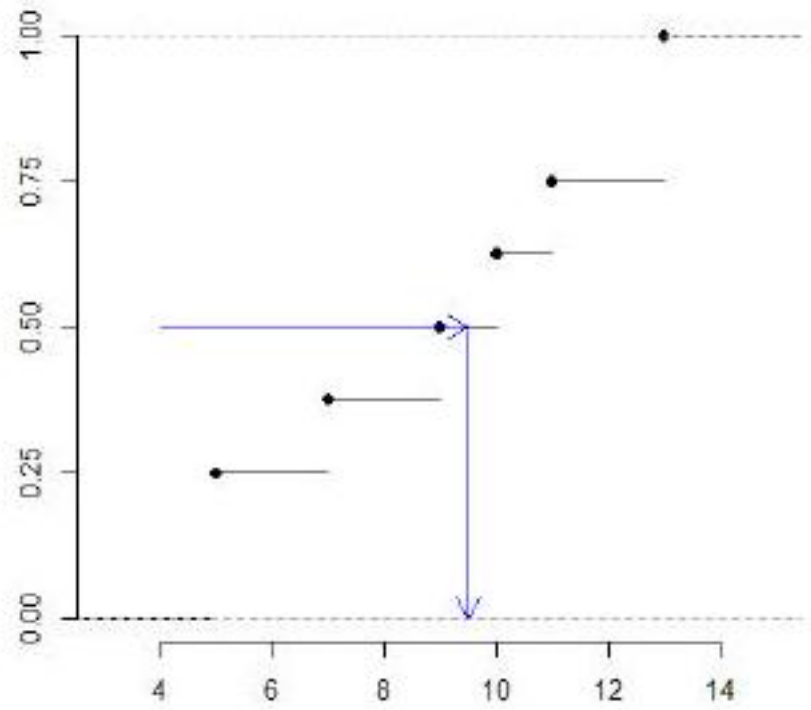
— Pour la série $5,5,7,9,10,11,13,13$, on a $n=8$ d'où $Me=\frac{9+10}{2}=9,5$

On peut aussi déterminer graphiquement la médiane au moyen de la courbe des fréquences cumulées, car la médiane Me est la valeur de la la variable statistique telle que l'ordonnée de la courbe cumulative soit égale à $\frac{1}{2}$, autrement dis,

$$F(Me)=\frac{1}{2}$$



FI G URE 2.1– Le premier cas.



F I G URE 2.2– Le deuxième cas.

Cas d'une variable continue

On commence d'abord par déterminer la classe médiane $[a_j, a_{j+1}[$, comme dans le cas discret, en utilisant les effectifs cumulés croissants (la classe comprenant la fréquence cumulée 50%).
On détermine ensuite la valeur de la médiane par interpolation linéaire suivant

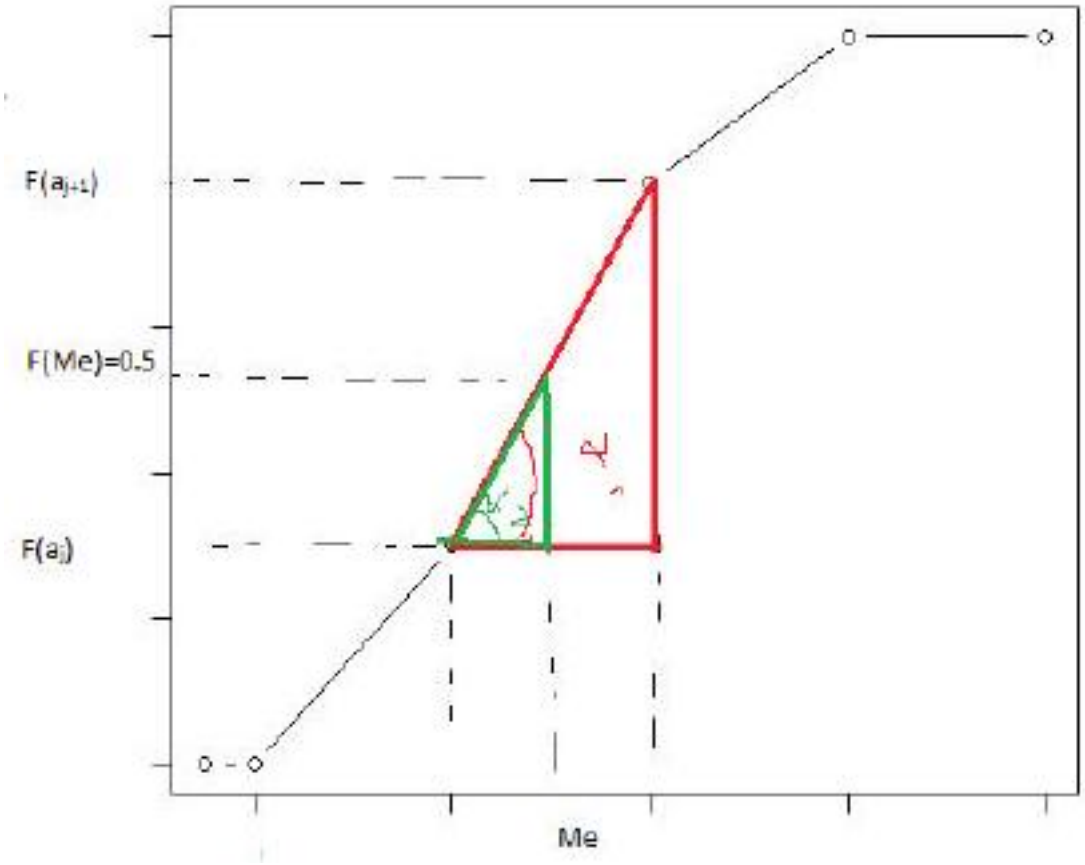
$$\tan\alpha_1 \approx \tan\alpha_2$$

alors,

$$\frac{F(a_{j+1}) - F(a_j)}{a_{j+1} - a_j} \approx \frac{F(Me) - F(a_j)}{Me - a_j}$$

où encore

$$\frac{F(a_{j+1}) - F(a_j)}{a_{j+1} - a_j} \approx \frac{0.5 - F(a_j)}{Me - a_j}$$



a_j a_{j+1}

Soit $F(a_{j+1}) = F_{j+1}, F(a_j) = F_j$ et $l_{j+1} = a_{j+1} - a_j$. Sachant que, $F(a_{j+1}) - F(a_j) = f_{j+1}$, alors

$$Me \approx a_j + \frac{0.5 - F_j}{f_{j+1}} \times l_{j+1}$$

Exemple 9

Soit le tableau statistique suivant :

X	$[0;4[$	$[4;8[$	$[8;12[$
f_i	0.1	0.6	0.3
$\tilde{f}_i$	0.1	0.7	1

On commence d'abord par déterminer la classe médiane, comme dans le cas discret, en utilisant les effectifs cumulés croissants par exemple. La classe médiane est donc ici $[4\,;\,8[$. En appliquant l'approximation précédente, on aura

$$Me \approx 4 + \frac{0.5 - 0.1}{0.6} \times 4 \approx 6.7$$

2.1.2 Mode

Définition 10

Le mode d'une variable statistique discrète X est la valeur x_i correspondant à l'effectif le plus élevé ; il est noté Mo.

Remarque

— Le mode peut être calculé pour tous les types de variable, quantitative et qualitative.

— Le mode n'est pas nécessairement unique.

2.1.3 Moyenne

Définition 11

La moyenne arithmétique simple d'une série statistique de modalité x_i est défini par

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_p}{n} = \frac{\sum_{i=1}^{p} x_i}{n}$$

La moyenne arithmétique pondérée d'une série statistique (x_i, n_i) est défini par :

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_p x_p}{n} = \frac{\sum_{i=1}^{p} n_i x_i}{n}$$

Exemple 10

La moyenne arithmétique de $11, 10, 12$ et 13 est

$$\bar{x} = \frac{10 + 11 + 12 + 13}{4} = 11.5$$

Exemple 11

La moyenne arithmétique pondérée de 11,11 , $11, 10, 10, 10, 10, 12, 12$ et 13 est

$$\bar{x} = \frac{4 \times 10 + 3 \times 11 + 2 \times 12 + 1 \times 13}{10} = 11$$

Remarque

— Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.

— on peut calculer $\bar{x}$ a partir des fréquences relatives car :

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

2.1.4 Quartiles

Définition 12

Le premier quartile Q_1 est la plus petite valeur du caractère telle qu'au moins 25% des termes de la série aient une valeur qui lui soit inférieure ou égale.
Le troisième quartile Q_3 est la plus petite valeur du caractère telle qu'au moins 75% des termes de la série aient une valeur qui lui soit inférieure ou égale.

Dans le cas d'une série à caractère discret, les quartiles s'obtiennent en ordonnant les valeurs dans l'ordre croissant puis :

— Si N est multiple de 4 alors Q_1 est la valeur de rang $\frac{N}{4}$ et Q_3 est la valeur de rang $\frac{3N}{4}$

— Si N n'est pas multiple de 4 alors Q_1 est la valeur de rang immédiatement supérieur à $\frac{N}{4}$ et Q_3 est la valeur de rang immédiatement supérieur à $\frac{3N}{4}$.
Dans le cas d'une série à caractère continu, les quartiles peuvent s'obtenir à partir du polygone des fréquences cumulées croissantes où Q_1 est la valeur correspondant à la fréquence cumulée croissante égale $0,25\,(F(Q_1) = 0.25)$ et Q_3 est la valeur correspondant à la fréquence cumulée croissante égale $0,75\,(F(Q_3) = 0.75)$. Comme pour le cas de la médiane pour le cas continu.

Intervalle interquartile

Lintervalle interquartile est la différence entre les valeurs du troisième et du premier quartile : $Q_3 - Q_1$

2.2 Paramètres de dispersion

Les caractéristiques de dispersion indiquent la position des observations au tour de la moyenne.

2.2.1 Etendue

Définition 13

L'étendue est la différence entre la plus grande valeur du caractère et la plus petite.

Remarque : L'étendue est très sensible aux valeurs extrêmes.

Exemple 12

Soit la série statistique suivante : $3, 6, 8, 9, 10, 14, 15, 6, 7, 11, 2, 5, 9$. L'étendue de cette série est : $15 - 2 = 13$.

Remarque
Contrairement à l'étendue, l'écart interquartile élimine les valeurs extrêmes, ce peut être un avantage. En revanche il ne prend en compte que 50% de l'effectif, ce peut être un inconvénient.

2.2.2 Variance et écart-type

Pour mesurer la dispersion d'une série, on peut s'intéresser à la moyenne des distances des valeurs à la moyenne. On utilise plutôt les carrés des distances qui facilitent les calculs.

Définition 14

On appelle **variance** d'une série quelconque à caractère quantitatif discret le nombre :

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

On appelle **écart-type** de cette série le nombre $\sigma = \sqrt{V}$.

Remarque

- Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.
- On peut calculer la variance de la façon suivante :

$$V = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$$

Exemple 13

Soit la variable statistique donnée par le tableau suivant

X	$[0;4[$	$[4;8[$	$[8;12[$
C_i	2	6	10
n_i	30	50	20

La moyenne est

$$\bar{x} = \frac{\sum_{i=1}^{p} n_i x_i}{n} = \frac{1}{100}[30 \times 2 + 50 \times 6 + 20 \times 10] = 5.6$$

La variance est

$$V = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2 = \frac{1}{100}\left[30 \times 2^2 + 50 \times 6^2 + 20 \times 10^2\right] - 5.6^2 = 38.12 - 31.36 = 6.76$$