CHAPITRE 6

STATISTIQUES

. INTRODUCTION

Le terme *statistiques* provient de l'allemand *Statistik*, emprunté de l'italien *statista* (Homme d'état) ou bien encore du latin *status* qui signifie état. Cette notion renvoie donc aux affaires de l'état, et par extension à la collecte, l'étude, l'analyse, puis l'interprétation, de données pouvant intéresser l'Etat dans son fonctionnement (le nombre d'habitant, répartition par sexe, par âge, par métiers etc.). Elle joue ainsi un rôle de prévisions dans des domaines comme l'économie ou la démographie.

On retrouve des traces de statistiques en Chine plus de 2000 ans avant Jésus-Christ, près de 1700 ans avant J.C en Égypte, et à peine plus tard dans les civilisations précolombiennes. Le premier bureau de statistique a été créé en France en 1800 par Napoléon. Cet organisme a pris en 1946 le nom d'Institut National de la Statistique et des Études Économiques (INSEE).

Les domaines d'applications des statistiques sont, de nos jours, très variés :

- **Economie, assurance, finance :** Etude de marchés, analyse de consommation des ménages, taxation des primes d'assurances etc.
- **Biologie, médecine :** Essais thérapeutiques, épidémiologie, dynamique de population etc.
- **Sciences de la Terre :** Prévisions météo, exploration pétrolière ...
- **Sciences humaines :** Enquête d'opinion, sondages, étude de population...
- **Sciences de l'ingénieur, de l'information** Contrôle qualité, sûreté de fonctionnement, traitement des images, des signaux etc.

Elles interviennent dès lors que les données collectées présentent des incertitudes , des variations. Cela peut provenir aussi bien du fait que les phénomènes observés ne sont pas prévisibles à l'avance, ou bien dès lors qu'on fait une mesure, il y a une marge d'erreur, ou encore du fait qu'on ne peut observer qu'un petit nombre d'individus de la population. Ces données sont ainsi issues de phénomènes aléatoires, et c'est la raison pour laquelle on fait intervenir les calculs de probabilités.

On distingue deux types de statistiques :

- Les statistiques descriptives : Elles consistent à résumer l'information contenu dans un ensemble de données, de façon rapide et claire. Cela passe par des représentations graphiques, des indicateurs de position, de dispersion, pour exprimer des tendances.
- Les statistiques inférentielles : Elles ont pour but de faire des prévisions dans le but de prendre des décisions sur des situations données. On fait des tests, des intervalles de confiances etc.

Dans ce chapitre nous allons seulement aborder des notions de statistique descriptive.

I. VOCABULAIRE DES STATISTIQUES

Définition 1.1

- Faire des statistiques revient à étudier d'abord un ensemble d'objets, d'éléments, qu'on appelle population. Cette population est notée Ω.
- Les éléments de cette population sont appelés individus. On se réfère à un individu de Ω en le notant ω ($\omega \in \Omega$).
- Sur ces individus on étudie un caractère particulier X, appelé aussi variable statistique. On note $X(\Omega)$ l'ensemble des valeurs prises par X.
- On appelle échantillon un sous-ensemble de la population. On appelle taille de l'échantillon le nombre d'éléments de la population dans cet échantillon.

Exemple :

- Différentes populations : La classe d'ECG1 du CIV, les européens, les arbres d'une forêt, les pièces fabriquées par une usine etc. Des individus de ces population seraient, respectivement, un étudiant, un habitant de l'europe, un arbre de la forêt, une pièce fabriquée.
- Différentes variables statistiques : L'âge, le sexe, le nombre de cafés consommés, la taille des arbres , pièce défectueuse etc.

Parmi toutes les variables statistiques, on en distingue deux catégories :

- **Les variables qualitatives.** Elles s'expriment par l'appartenance à une modalité.

$$X = \text{Sexe, Musique écoutée, opinions politiquesn, couleur des yeux}$$

- **Les variables quantitatives.** Ce sont les variables X qui prennent des valeurs numériques. Alors

$$X : \Omega \to \mathbb{R}$$

On dit qu'une variable quantitative est

- Discrète lorsque X prend un nombre fini, ou infini dénombrable, de valeurs.

  $X =$ Notes pendant une compétition, nombre de voiture par jour à un péage, nombre d'enfants par famille.

- Continue sinon.

  $X =$ Temps avant l'arrivée d'un premier client dans un magasin, diamètre d'une pièce de monnaie, salaire

Dans la suite nous n'étudierons que des variables quantitatives, dans un échantillon de taille N.

Remarque :

Bien souvent, la population est trop nombreuses pour pouvoir étudier le caractère sur chacun de ses individus. C'est la raison pour laquelle on considère alors un échantillon.

Définition 1.2

- Soit X une variable discrète. On appelle modalité les différentes valeurs prises par X.

$$X(\Omega) = \{x_1, x_2, ...x_p\} \text{ avec } x_1 < x_2 < ... < x_p$$

- Soit X une variable continue. Ses valeurs sont alors regroupées en classes. S'il y a p classes, on peut les noter

$$[x_1; x_2[, [x_2; x_3[, [x_3; x_4[....[x_{p-1}; x_p[, [x_p; x_{p+1}[\qquad \text{où } x_1 < x_2 ... < x_p < x_{p+1}$$

Remarque :

Dans le cas discret,$X(\Omega) = \{x_i\}_{i\in I}$, il est également possible de ranger les valeurs par classes. Il serait équivalent de compter le nombre d'individus prenant le caractère x_1, x_2,.... que le nombre d'individus dans les classes

$$[x_1; x_2[, [x_2; x_3[, [x_3; x_4[....[x_{p-1}; x_p[, ...$$

Définition 1.3

On considère un échantillon de taille N, et une variable quantitative X sur cet échantillon.

- On appelle effectif, et l'on notera n_i, le nombre d'individus ω de l'échantillon ayant la modalité x_i, ou dans la classe $[x_i; x_{i+1}[$:

$$X(\omega) = x_i \qquad \text{ou} \qquad X(\omega) \in [x_i; x_{i+1}[$$

.

- On appelle effectif cumulé en x_i (ou en $[x_i; x_{i+1}[$), et l'on notera N_i, la somme des effectifs des modalités (ou des classes) qui lui sont inférieures ou égales :

$$N_i = \sum_{k=1}^{i} n_k$$

- On appelle fréquence de x_i (ou de $[x_i; x_{i+1}[$), notée f_i, le nombre

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

- On appelle fréquence cumulée en x_i (ou en $[x_i; x_{i+1}[$), notée F_i, la somme des effectifs des modalités (ou des classes) qui lui sont inférieures ou égales :

$$F_i = \sum_{k=1}^{i} f_k$$

Remarque :

La somme de tous les effectifs doit toujours être égale à la taille de l'échantillon total N :

$$\sum_{i=1}^{p} n_i =$$

La somme des fréquences doit être égale à 1.

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

Définition 1.4

On appelle série statistique d'un échantillon la donnée de la liste des modalités ou des classes de la variable étudiée dans l'échantillon, accompagnée des effectifs ou correspondants. On la note

$$(x_i, n_i)_{i\in[|1;p|]} \qquad \text{ou} \qquad ([x_i, x_{i+1}[, n_i)_{i\in[|1;p|]}$$

Exemples :

E1 – Un concessionnaire d'automobiles neuves a enregistré au cours de ses 40 premières semaines d'opération, le nombre X d'automobiles qu'il a vendu hebdomadairement. Il a obtenu les résultats suivants :

$$5, 7, 2, 6, 3, 4, 8, 5, 4, 3, 9, 6, 5, 7, 6, 8, 3, 4, 4, 0, 8, 6, 7, 1, 5, 5, 4, 6, 6, 10, 9, 8, 1, 5, 5, 6, 7, 8, 5, 5$$

La série statistique est la donnée du tableau suivant :

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	1	3	5	9	7	4	5	2	1

On peut ajouter les données suivantes :

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	1	2	1	3	5	9	7	4	5	2	1
N_i	1	3	4	7	12	21	28	32	37	39	40
f_i (en %)	2,5	5	2,5	7,5	12,5	22,5	17,5	10	12,5	5	2,5
F_i (en %)	2,5	7,5	10	17,5	30	52,5	70	80	92,5	97,5	100

E2 – On relève dans une banque à une date donnée les montants des économies de 1000 clients en euros. Les résultats obtenus sont les suivants :

Montants des économies en euros x_i	Nombre n_i de clients	Effectifs cumulés croissants
[0 ;500[	5	5
[500 ;1000[	12	17
[1000 ;1500[	33	50
[1500 ;2000[	71	121
[2000 ;2500[	119	240
[2500 ;3000[	175	415
[3000 ;3500[	185	600
[3500 ;4000[	158	758
[4000 ;4500[	122	880
[4500 ;5000[	69	949
[5000 ;5500[	35	984
[5500 ;6000[	11	995
[6000 ;6500[	5	1000

II. PARAMÈTRES DE POSITION

II. 1 LE MODE

Définition 2.1

On appelle **mode** d'une série statistique la ou les valeurs du caractère dont l'effectif est le plus élevé.
Dans le cas d'une répartition à l'aide de classes, la classe dont l'effectif est le plus élevé est appelée **classe modale**, le mode étant le centre de la classe.

Exemples :

E1 – Dans l'exemple 1 précédent, le mode et égal à 5

E2 – Dans l'exemple 2 précédent, la classe modale est $[3000; 3500[$ et le mode égal à 3250

Remarque :

Un mode n'est pas forcément unique. Si on considère la série statistique suivante

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1

Les modes sont 3 et 8.

II. 2 Moyenne

Définition 2.2

- **1er cas : Cas d'une variable discrète** On appelle **moyenne** de la série statistique $(x_i, n_i)_{1\leq i\leq p}$ (ou $(x_i, f_i)_{1\leq i\leq p}$, et on note $\overline{X}$, le réel

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

- **1er cas : Cas d'une variable continue** On appelle **moyenne** de la série statistique $([x_i; x_{i+1}[, n_i)_{1\leq i\leq p}$ (ou $([x_i; x_{i+1}[, f_i)_{1\leq i\leq p}$, et on note $\overline{X}$, le réel

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

où $c_i = (x_i + x_{i+1})/2$ est le centre de la classe $[x_i; x_{i+1}[$.

Exemples :

E1 – Dans le cas du concessionnaire d'automobiles, le nombre moyen de voitures vendu par semaine est donné par

$$\overline{X} = \frac{1\times 0 + 2\times 1 + 1\times 2 + 3\times 3 + 5\times 4 + 9\times 5 + 7\times 6 + 4\times 7 + 5\times 8 + 2\times 9 + 1\times 10}{40}$$
$$= \frac{216}{40} = 5,4$$

E2 – Dans le cas de la banque, la moyenne est donnée par

$$\overline{X} = \frac{5\times 250 + 12\times 750 + \ldots + 11\times 5750 + 5\times 6250}{1000} = 3243$$

II. 3 Médiane

Définition 2.3

On appelle **médiane** d'une série statistique, le réel, souvent noté M_e ou Med, partageant la série en deux séries d'effectifs égaux.

Méthode :

- **Cas d'une variable discrète.** On considère ici une série statistique $(x_i, n_i)_{1\leq i\leq p}$ d'effectif total N.
  - Si N est impair : La médiane est une valeur de la série : la $\frac{N+1}{2}$-ième.
  - Si N est pair : La médiane tombe entre la $\frac{N}{2}$-ième valeur et la $\frac{N}{2}+1$-ième valeur.
    - Si ces deux valeurs sont égales, alors la médiane est la valeur commune.
    - Sinon, la médiane est la moyenne de ces deux valeurs.
- **Cas d'une variable discrète.** On considère ici une série statistique $([x_i; x_{i+1}[, f_i)_{1\leq i\leq p}$.
  Pour repérer la médiane de la série il suffit :
  - De dresser le tableau des fréquences cumulées.
  - De repérer la classe $[x_m, x_{m+1}[$ pour laquelle les fréquences cumulées atteignent ou dépasse 50%. La

médiane appartient alors à cette classe :

$$\boxed{M_e = x_m + (x_{m+1} - x_m)\left(\frac{0,5 - F_{m-1}}{F_m - F_{m-1}}\right)}$$

Exemples :

E1 – Considérons la série statistique suivante :

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1
N_i	3	4	8	12	14	15

Ici $N = 15$ est impair. Ainsi la médiane est la 8-ième valeur. On regarde dans la ligne des effectifs cumulées où se trouve la 8-ième valeur. On trouve alors $\boxed{M_e = 3.}$

E2 – Considérons à présent la série statistique suivante :

x_i	1	2	3	8	9	10
n_i	3	2	4	4	4	3
N_i	3	5	9	13	17	20

Ici $N = 20$ est pair. On repère où se trouve les 10 et 11-ième valeurs. Elles sont toutes les deux égales à 8, ainsi $\boxed{M_e = 8.}$

E3 – Reprenons l'exemple de la banque :

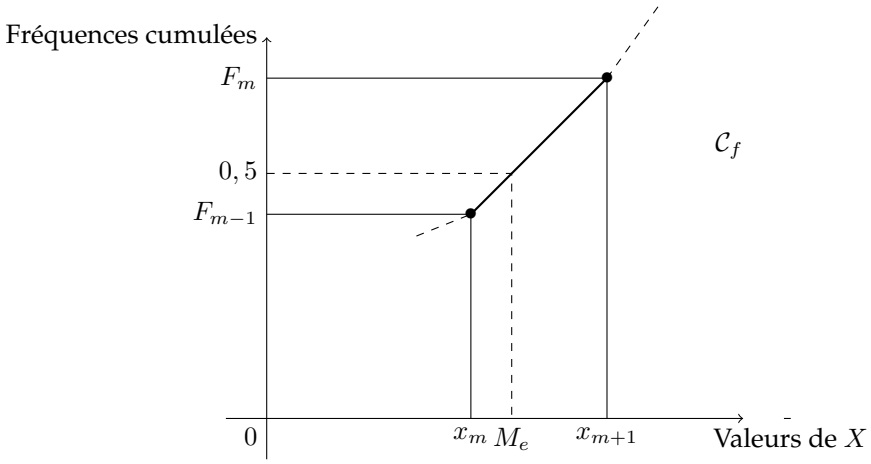
x_i	n_i	N_i	Fréquences cumulées F_i (%)
[0;500[	5	5	0,55
[500;1000[	12	17	1,7
[1000;1500[	33	50	5
[1500;2000[	71	121	12,1
[2000;2500[	119	240	24
[2500;3000[	175	415	41,5
[3000;3500[	185	600	60
[3500;4000[	158	758	75,8
[4000;4500[	122	880	88
[4500;5000[	69	949	94,9
[5000;5500[	35	984	98,4
[5500;6000[	11	995	99,5
[6000;6500[	5	1000	100

On repère la classe contenant au moins $\%50$ des effectifs, il s'agit de la classe $[3000; 3500[$. (Ici $x_m = 3000$, $x_{n+1} = 3500$, $F_m = 60$, $F_{m-1} = 41,5$. La formule précédente donne

$$M_e = 3000 + \left(\frac{0,5 - 0,415}{0,6 - 0,415}\right)(3500 - 3000) = 3000 + 500\left(\frac{50 - 41,5}{60 - 41,5}\right) = 3000 + 500 \times \frac{8,5}{18,5} \simeq 3229,73$$

Explications : D'où vient la formule $\boxed{M_e = x_m + (x_{m+1} - x_m)\left(\frac{0,5 - F_{m-1}}{F_m - F_{m-1}}\right)}$?

Il s'agit d'une méthode *d'interpolation linéaire* . On sait que la médiane appartient à la classe $[x_m; x_{m+1}[$ et on fait l'hypothèse que les valeurs dans la classe sont **uniformément répartis :**



Cela se traduit par le fait qu'entre les points $(x_m; F_{m-1})$ et $(x_{m+1}; F_m)$, l'évolution est linéaire.

Rappel : L'équation d'une droite est de la forme $y = ax+b$, où a est le coefficient directeur, donné par l'expression

$$a = \frac{\text{différence en ordonnées}}{\text{différence en abscisses}}$$

Ici $a = \dfrac{F_m - F_{m-1}}{x_{m+1} - x_m}$ ainsi l'équation de la droite a pour équation

$$y = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x + b$$

Cette droite passe par le point $(x_m; F_{m-1})$ donc

$$b = F_{m-1} - \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x_m$$

En remplaçant on trouve pour équation

$$y = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x + F_{m-1} - \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) x_m = \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) (x - x_m) + F_{m-1}$$

La médiane M_e a pour image $0,5$, donc en insérant ces valeurs dans l'équation de la droite, on trouve

$$\begin{aligned}
0,5 &= \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) (M_e - x_m) + F_{m-1} \\
\Leftrightarrow 0,5 - F_{m-1} &= \left(\frac{F_m - F_{m-1}}{x_{m+1} - x_m}\right) (M_e - x_m) \\
\Leftrightarrow M_e - x_m &= (0,5 - F_{m-1}) \left(\frac{x_{m+1} - x_m}{F_m - F_{m-1}}\right) \\
\Leftrightarrow M_e &= x_m + \left(\frac{0,5 - F_{m-1}}{F_m - F_{m-1}}\right) (x_{m+1} - x_m)
\end{aligned}$$

Remarques :

R1 – La médiane correspond à la valeur telle que au moins 50% est valeurs de la série lui sont inférieures.

R2 – Il est possible, dans le cas d'une variable continue, de travailler avec les effectifs plutôt qu'avec les fréquences. On remplace alors dans la formule les fréquences cumulées par les effectifs cumulées N_m et N_{m-1} et le 0,5 par $\dfrac{N}{2}$.

R3 – On doit bien distinguer la médiane M_e et la moyenne $\overline{X}$ d'une population. Le calcul de la moyenne fait intervenir toutes les données ce qui n'est pas le cas pour la détermination de la médiane. De plus, la moyenne est sensible aux variations des valeurs extrêmes de la série statistique, ce qui n'est pas le cas de la médiane.

II. 3 Quartiles et Déciles

Définition 2.4

- On appelle *premier quartile*, et on le note q_1 ou Q_1, la valeur de la série telle que au moins 25% des valeurs de l'effectif total lui sont inférieures ou égales.
- On appelle *troisième quartile*, et on le note q_3 ou Q_3, la valeur de la série telle que au moins 75% des valeurs de l'effectif total lui sont inférieures ou égales.

Définition 2.5

- On appelle *k-ième décile*, et on le note d_k ou D_k, la valeur de la série telle que au moins $k \times 10\%$ des valeurs de l'effectif total lui sont inférieures ou égales.

Méthode :

Les calculs des quartiles et déciles se font de manière analogue à ceux sur la médiane.

Exemples :

E1 – Reprenons la série statistique suivante :

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1
N_i	3	4	8	12	14	15

Ici $N = 15$ est impair, $\frac{N}{10} = 1,5$,$\frac{N}{4} = 3,25$, $\frac{3N}{4} = 9,75$. . Ainsi

- le premier décile est la 2-ième valeur : $\boxed{d_1 = 1}$.
- le premier quartile est la 4-ième valeur : $\boxed{Q_1 = 2.}$
- le troisième quartile est la 10-ième valeur : $\boxed{Q_3 = 8}$.

E2 – Reprenons l'exemple de la banque :

x_i	n_i	N_i	Fréquences cumulées F_i (%)
[0 ;500[	5	5	0,55
[500 ;1000[	12	17	1,7
[1000 ;1500[	33	50	5
[1500 ;2000[	71	121	12,1
[2000 ;2500[	119	240	24
[2500 ;3000[	175	415	41,5
[3000 ;3500[	185	600	60
[3500 ;4000[	158	758	75,8
[4000 ;4500[	122	880	88
[4500 ;5000[	69	949	94,9
[5000 ;5500[	35	984	98,4
[5500 ;6000[	11	995	99,5
[6000 ;6500[	5	1000	100

On repère les classes contenant au moins 25% et 75% des effectifs : il s'agit des classes $[2500; 3000[$ et $[3500; 4000[$.

$$Q_1 = 2500 + \left(\frac{0,25 - 0,24}{0,415 - 0,24}\right) 500 \simeq 2528,57$$

$$Q_3 = 3500 + \left(\frac{0,75 - 0,6}{0,758 - 0,6}\right) 500 \simeq 3974,68$$

III. Paramètres de dispersion

Considérons deux élèves A et B qui ont obtenue en mathématiques les notes suivantes :

$$7, 8, 11, 12, 13, 13, 13 \text{ pour A} \qquad 4, 7, 9, 12, 13, 13, 19 \text{ pour B}$$

Il est facile de vérifier que les séries de notes de A et B ont la même médiane (12), la même moyenne (11) et le même mode (13). Pourtant, ces deux séries statistiques sont bien différentes car un simple coup d'?il permet de voir que les notes de B sont plus dispersées.

Les paramètres de positions ne peuvent donc pas, à eux seuls, donner un aperçu de la répartition des valeurs. Pour être plus précis dans notre étude, on introduit plusieurs outils qui permettent de rendre compte de la dispersion, et ainsi décrire plus précisément la population.

III. 1 L'étendue

Définition 3.1

On appelle *étendue* d'une série statistique, l'écart entre la plus grande et la plus petite valeur.

Exemple :

Dans le cas des séries de notes de précédente, les étendues sont $e_A = 13 - 7 = 6$ et $e_B = 19 - 4 = 15$. On en conclut que les notes de B sont plus étendue que celles de A.

III. 2 L'écart inter-quartiles

Les quartiles peuvent être vus comme des indicateurs de dispersion, ils renseignent tout de suite de la répartition des valeurs autours de la médiane. Pour illustrer autrement cette dispersion, on peut utiliser :

Définition 3.2

On appelle *écart interquartile* la longueur de l'intervalle $[Q_1; Q_3]$ c'est à dire $Q_3 - Q_1$.

Remarque :

L'intervalle $[Q_1; Q_3]$ contient environ 50% des valeurs, dont la médiane. Plus l'écart interquartile est élevé, plus les valeurs peuvent dispersées.

III. 3 Variance et écart-type

Définition 3.3

- Si X est discrète : La variance de la série $(x_i, n_i)_{1 \leq i \leq p}$ est la quantité

$$V(X) = \frac{1}{N} \sum_{i=1}^{p} n_i (x_i - \overline{X})^2 = \sum_{i=1}^{p} f_i (x_i - \overline{X})^2$$

- Si X est continue : La variance de la série $([x_i; x_{i+1}[, n_i)_{1 \leq i \leq p}$ est la quantité

$$V(X) = \frac{1}{N} \sum_{i=1}^{p} c_i (x_i - \overline{X})^2$$

où $c_i = \dfrac{x_{i+1} + x_i}{2}$ est le centre de la classe $[x_i; x_{i+1}[$.

Remarques :

R1 – La variance représente donc la moyenne des carrés des écarts à la moyenne.

R2 – Le calcul de la variance ne peut être fait avant celui de la moyenne.

Cette formule n'est pas très commode car elle nécessite p soustractions, p mises au carrés, p multiplications et $p-1$ additions. Dans la pratique on utilise plutôt la formule suivante

Proposition 3.4 — Formule de Koenig-Huygens

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - (\overline{X})^2 = \sum_{i=1}^{p} f_i x_i^2 - (\overline{X})^2$$

Démonstration. □

Remarque :

Pour une variable continue, il suffit de remplacer x_i par c_i dans la formule.

Définition 3.5

L' écart-type d'une série statistique est la racine carrée de la variance.

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

Remarque :

L'écart-type est la caractéristique de dispersion la plus souvent utilisée. On peut montrer que sous certaines conditions

- l'intervalle $[\overline{X} - \sigma(X), \overline{X} + \sigma(X)]$ contient environ 68% des valeurs de la série.
- l'intervalle $[\overline{X} - 2\sigma(X), \overline{X} + 2\sigma(X)]$ contient environ 95% des valeurs de la série.

Exemple :

Dans l'exemple de la banque, on peut rassembler les données dans le tableau suivant :

x_i	n_i	c_i	$n_i x_i$	$n_i x_i^2$
[0 ;500[	5	250	1250	312500
[500 ;1000[	12	750	9000	6750000
[1000 ;1500[	33	1250	41250	51562500
[1500 ;2000[	71	1750	124250	217437500
[2000 ;2500[	119	2250	267750	602437500
[2500 ;3000[	175	2750	481250	1323437500
[3000 ;3500[	185	3750	601250	1954062500
[3500 ;4000[	158	3750	592500	2221875000
[4000 ;4500[	122	4250	518500	2203625000
[4500 ;5000[	69	4750	327750	1556812500
[5000 ;5500[	35	5250	183750	964687500
[5500 ;6000[	11	5750	63250	363687500
[6000 ;6500[	5	6250	31250	195312500
TOTAL	1000		3243000	11662000000

On sait déjà que $\overline{X} = 3243$, mais cela se retrouve grâce à la quatrième colonne

$$\overline{X} = \frac{3243000}{1000} = 3243$$

La somme des valeurs dans la dernière colonne permet de trouver la variance.

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - (\overline{X})^2 = \frac{11662000000}{1000} - (3243)^2 = 1144951$$

On en déduit l'écart-type

$$\sigma(X) = \sqrt{1144951} \simeq 1070,02$$

On peut interpréter ce résultat de la manière suivante

- l'intervalle $[\overline{X} - \sigma(X), \overline{X} + \sigma(X)] = [3243 - 1070,02; 3243 + 1070,02] = [2172,98; 4313,02]$ contient environ 68% des valeurs de la série.
- l'intervalle $[\overline{X} - 2\sigma(X), \overline{X} + 2\sigma(X)] = [3243 - 2 \times 1070,02; 3243 + 2 \times 1070,02] = [1102,96; 5383,04]$ contient environ 95% des valeurs de la série.

Proposition 3.6 — Transformation affine

Soit une série statistique de moyenne $\overline{X}$, de médiane M_e, de variance $V(X)$ et d'écart-type $\sigma(X)$. Soit $a \in \mathbb{R}^*$, $b \in \mathbb{R}$.
Alors la variable $aX + b$

- a pour moyenne $a\overline{X} + b$.
- a pour médiane $aM_e + b$.
- a pour variance $a^2V(X)$.
- a pour écart-type $|a|\sigma(X)$.

Méthode :

On souhaite calculer les caractéristique de la série $(x_i, n_i)_{1 \leq i \leq p}$. On s'aperçoit qu'il existe $a \neq 0$ et $b \in \mathbb{R}$ tel que les caractéristiques de la série $(ax_i + b, n_i)_{1 \leq i \leq p}$ sont plus simples à calculer.

- On calcule les caractéristiques de $(ax_i + b, n_i)_{1 \leq i \leq p}$. Elles sont égales à $a\overline{X} + b, aM_e + b, a^2V(X), |a|\sigma(X)$.
- On retrouve les caractéristiques $\overline{X}, M_e, V(X), \sigma(X)$ par des opérations élémentaires.

Exemple :

Considérons la série statistique suivante :

x_i	-150	-155	-160	-165	-170	-180	-185
n_i	2	1	4	6	4	2	1

Il semble que la série soit centrée autours de -165. On peut étudier les caractéristique de la variable $Y = -X - 165$:

x_i	-15	-10	-5	0	5	10	15
n_i	2	1	4	6	4	2	1

- La moyenne est alors

$$\overline{Y} = \frac{-15 \times 2 + (-10) \times 1 + ... + 10 \times 2 + 15 \times 1}{20} = \frac{-5}{20} = -0.25$$

  On a $X = -Y - 165$ donc $\overline{X} = -(-0.25) - 165 = -164,75$.
- Comme 20 est pair, la médiane de Y est entre la 10ième et la 11ième valeur $M = 0$. Ainsi la médiane de X est $M_e = -0 - 165 = -165$.
- La variance de Y est donnée par la formule de Koenig-Huygens

$$V(Y) = \frac{(-15)^2 \times 2 + (-10)^2 \times 1 + ... + 10^2 \times 2 + 15^2}{20} = \frac{1175}{20} = \frac{235}{4}$$

  et l'écart-type par

$$\sigma(Y) = \sqrt{\frac{235}{4}} = \frac{\sqrt{235}}{2}$$

  Ainsi la variance et l'écart-type de X sont

$$V(X) = (-1)^2V(Y) = \frac{235}{4} \qquad \sigma(X) = |-1|\sigma(Y) = \frac{\sqrt{235}}{2}$$

IV. Représentations graphiques

IV. 1 Représentation de variables discrètes

1. Diagramme en bâton

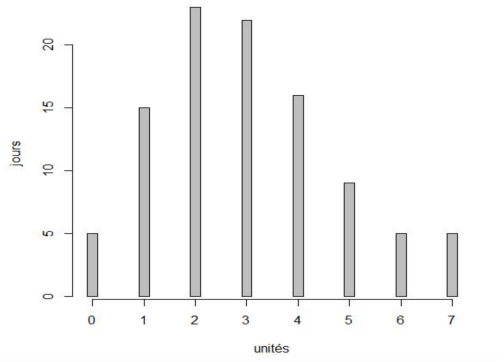
C'est une représentation graphique particulièrement indiquée pour la représentation de variable discrètes. La série $(x_i, n_i)_{1\leq i\leq p}$ se représente ainsi dans un repère orthogonal. On indique en abscisses les valeurs x_i et les n_i en ordonnées. Pour chacune des valeurs, on trace un bâton de hauteur n_i .

Exemple :

Le responsable des ventes d'un magasin notes, au cours de 100 jours ouvrables, la demande journalière pour un de ses produit :

x_i	0	1	2	3	4	5	6	≥ 7
n_i	5	15	23	22	16	9	5	5
N_i	5	20	43	65	81	90	95	100
f_i (en %)	5	15	23	22	16	9	5	5
F_i(en %)	5	20	43	65	81	9	95	1

Le diagramme en bâtons correspondant est le suivant :



Remarque :

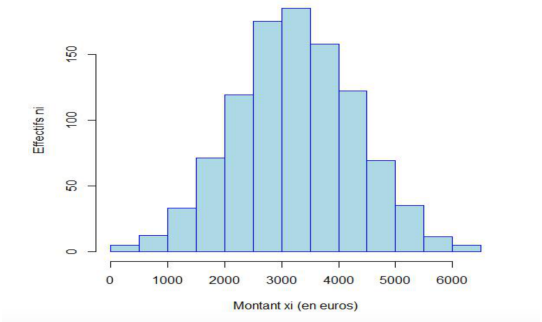
On peut mettre les fréquences en ordonnées.

2. Histogramme

C'est une représentation graphique particulièrement indiquée pour la représentation de variable continues. La série $([x_i; x_{i+1}[, n_i)_{1\leq i\leq p}$ se représente ainsi dans un repère orthogonal. On indique en abscisses les valeurs x_i et les n_i en ordonnées. On trace les rectangles de bases $[x_i; x_{i+1}]$ et d'aire proportionnelle à n_i.

Exemple :

Dans l'exemple de la banque, un histogramme est donné par

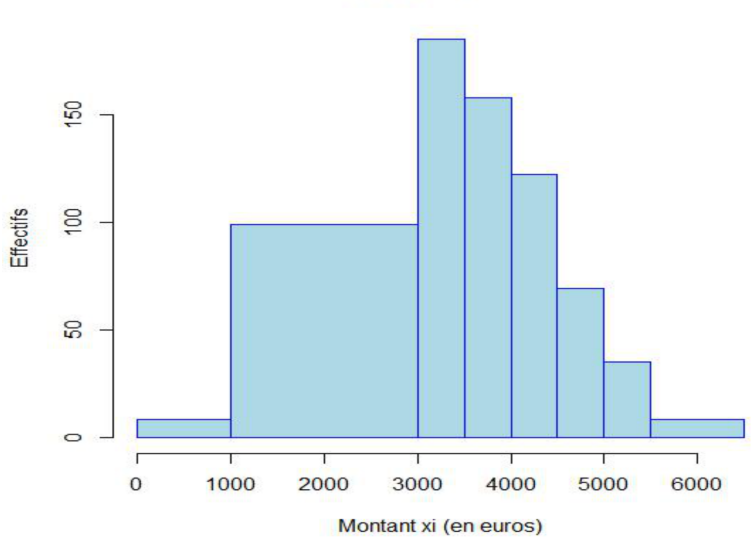


Remarque :

Il est possible de changer l'amplitude des classes

Montant des économies (en euros) x_i	Nombre n_i de clients	Effectifs cumulés croissants
$[0; 1000[$	17	17
$[1000; 3000[$	398	415
$[3000; 3500[$	185	600
$[3500; 4000[$	158	758
$[4000; 4500[$	122	880
$[4500; 5000[$	69	949
$[5000; 5500[$	35	984
$[5500; 6500[$	16	1000
TOTAL	1000	–

On obtient l'histogramme suivant



Remarques :

R1 – L'histogramme est obtenu en juxtaposant des rectangles qui ont pour bases les différentes classes, et dont les **surfaces** sont proportionnelles aux fréquences.

R2 – Si les classes ont toutes la même amplitudes, les **hauteurs** sont proportionnelles aux fréquences, et dont aux effectifs.

R3 – Il est possible de mettre les fréquences en ordonnées.

3. Courbe des effectifs cumulés croissants

Obtenir la courbe des fréquences cumulées, indiquées pour les variables continues, consiste à placer les différents points

$$(x_1; 0) \qquad (x_i; F_{i-1}) \text{ pour } i \text{ allant de 1 à p}$$

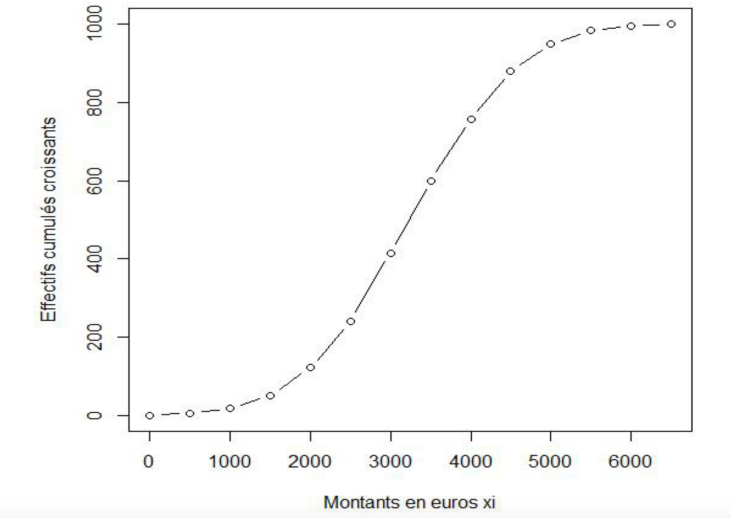
puis à les relier par des segments.

Exemple :

Toujours dans l'exemple de la banque :

x_i	n_i	N_i	Fréquences cumulées F_i (%)
[0 ;500[	5	5	0,55
[500 ;1000[	12	17	1,7
[1000 ;1500[	33	50	5
[1500 ;2000[	71	121	12,1
[2000 ;2500[	119	240	24
[2500 ;3000[	175	415	41,5
[3000 ;3500[	185	600	60
[3500 ;4000[	158	758	75,8
[4000 ;4500[	122	880	88
[4500 ;5000[	69	949	94,9
[5000 ;5500[	35	984	98,4
[5500 ;6000[	11	995	99,5
[6000 ;6500[	5	1000	100

On obtient la courbe suivante



Remarques :

R1 – On trace de la même manière la courbe des fréquences cumulées croissantes.

R2 – Ce n'est pas vraiment une "courbe", mais une succession de segments.

4. Diagramme en boîte

Le diagramme en boîte, (aussi appelé *boîte de Tukey* ou encore *boîte à moustaches* est un diagramme qui regroupe les valeurs extrêmes, les quartiles Q_1 et Q_3 , ainsi que la médiane M_e.

Exemple :

Reprenons la série statistique suivante :

x_i	1	2	3	8	9	10
n_i	3	1	4	4	2	1
N_i	3	4	8	12	14	15

Nous avions trouver $M_e = 3$, $Q_1 = 2$, $Q_3 = 8$. Le diagramme en boîte se représente ainsi :