XXVII. Statistiques

La **statistique** est la branche des mathématiques qui consiste à la collecte, au classement, à l'*analyse* et à l'interprétation de *données* afin d'en tirer des conclusions et de faire des *prévisions*. Parmi les "consommateurs" de statistiques, on trouve les assureurs (risques d'acidents, de maladie des assurés), le domaine médial (épidémiologie, traitements), les démographes (populations et leur dynamique), les économistes (emploi, conjoncture économique), les météorologues...

On peut distinguer deux classes de méthodes statistiques.

- **La statistique descriptive.** Elle a pour but de *résumer* l'information contenue dans les données de façon synthétique et efficace par des représentations graphiques, des indicateurs de position et dispersion,.... Elle permet de dégager les *caractéristiques* essentielles du phénomène étudié.
- **La statistique inférentielle.** Elle a pour but de faire des *prévisions* et de prendre des décisions au vu des observations par des estimations paramétriques, des intervalles de confiance, des tests d'hypothèses. Cela nécessite de définir des modèles probabilistes du phénomène aléatoire.

1 Vocabulaire Général

Définition	Exemples
Une **population** est l'*ensemble* auquel se rapportent les données étudiées.	La population française, la classe ECG1, les poissons d'une rivière,...
Un **individu** est un *élément* de la population.	Un·e français·e, un étudiant d'ECG1, un poisson de la rivière,...
Un **échantillon** est un *sous-ensemble* de la population.	Les françaises, les étudiants d'ECG1 qui font espagnol, les poissons mâles de la rivière,...
Un caractère ou une **variable** (statistique) est une *propriété commune* aux individus d'une population ou d'un échantillon	Le sexe, la taille, la catégorie socio-professionnelle,...
Une **variable** est dite **quantitative** si elle est exprimée par un *nombre*..	Revenu, poids, résultats à un examen...
Une **variable** est dite **qualitative** si elle s'exprime par l'appartenance à une *modalité*.	Couleur des yeux, ville de résidence,...
Une **variable quantitative** est dite **discrète** si elle ne peut prendre que des valeurs isolées (formellement, des valeurs *entières* ou *finies*)	Notes à un examen, nombre d'habitants d'un pays,...
Une **variable quantitative** est dite **continue** lorsque toute les valeurs d'un *intervalle* de $\mathbb{R}$ sont possibles.	La température, le poids (si la précision est suffisante),...

Dans ce chapitre, on se limitera à une seule variable d'où le terme de statistique **univariée**.

2 Étude d'une série statistique

Dans la suite, on se place dans un cadre abstrait.

- On note $\mathcal{P}$ la **population** étudiée.
- On note $\mathcal{E}$ l'**échantillon** étudié (c'est-à-dire $\mathcal{E}$ est un sous-ensemble de $\mathcal{P}$). On suppose que cet échantillon contient n **individus** (pour $n \in \mathbb{N}$).
- On s'intéresse à une seule **variable quantitative discrète** x qui peut prendre les valeurs dans l'ensemble
$$\Omega = \{x_1, x_2, \cdots, x_p\}$$
pour un $p \in \mathbb{N}^*$. Par convention, on note les valeurs prises dans l'ordre croissante, c'est-à-dire $x_1 < x_2 < \cdots < x_p$. Ainsi, p est le nombre de valeurs possibles que peut prendre le caractère étudié.

Pour chaque individu de l'échantillon, la variable x prend une certaine valeur parmi les p valeurs possibles. On note, pour tout $i \in [\![1, p]\!]$, n_i le nombre d'individus de l'échantillon $\mathcal{E}$ pour lesquels x prend la valeur x_i.

Définition 2.1 L'ensemble des couples $(x_i, n_i)_{i \in [\![1,p]\!]}$ est appelé une **série statistique univariée** associée à l'échantillon $\mathcal{E}$.

Exemple 2.2 — Notes d'élèves à un DS. Parmi une classe de 30 élèves, on a récupéré chaque des notes obtenues au dernier DS. Les données brutes obtenues sont :

11 – 16 – 18 – 15 – 16 – 17 – 13 – 14 – 13 – 12 – 10 – 15 – 20 – 13 – 14
15 – 15 – 11 – 17 – 12 – 15 – 13 – 19 – 16 – 12 – 16 – 14 – 15 – 14 – 14

- La **population** est $\mathcal{P}$ = ??
- L'**échantillon** étudié est $\mathcal{E}$ = {la classe}
- La **variable quantitative discrète** est x = la note au DS.
- La variable quantitative discrète peut prendre ses valeurs dans l'ensemble
$$\Omega = \{10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$$
- La **série statistique** est donnée par
$$\{(10,1), (11,2), (12,3), (13,4), (14,5), (15,6), (16,4), (17,2), (18,1), (19,1), (20,1)\}$$
Il peut être plus parlant de représenter la série statistique dans un tableau.

Valeur x_i	10	11	12	13	14	15	16	17	18	19	20
Effectif n_i	1	2	3	4	5	6	4	2	1	1	1

Exemple 2.3 — Nombre de personnes vivant dans le foyer. Lors d'une enquête on a interrogé 50 employés d'une entreprise afin de connaître le nombre de personnes vivant avec eux dans leur foyer. Les données brutes obtenues sont :

0 3 1 4 3 0 4 1 3 1 5 2 4 2 3 3 2 5 5 2 4 2 2 2 4
1 1 2 3 5 1 0 3 3 4 5 1 2 1 2 3 2 2 2 4 0 3 0 2 2

- La **population** est $\mathcal{P}$ = ??
- L'**échantillon** étudié est $\mathcal{E}$ = {les employés}
- La **variable quantitative discrète** est x = nbre de personnes vivant avec eux.
- La variable quantitative discrète peut prendre ses valeurs dans l'ensemble
$$\Omega = \{0, 1, 2, 3, 4, 5\}$$
- La **série statistique** est donnée par
$$\{(0,5), (1,8), (2,15), (3,10), (4,7), (5,5)\}$$
Il peut être plus parlant de représenter la série statistique dans un tableau.

Valeur x_i	0	1	2	3	4	5
Effectif n_i	5	8	15	10	7	5

2.1 Fréquences, Fréquences cumulées

Définition 2.4 Soit x une variable statistique prenant les valeurs $x_1 < x_2 < \cdots < x_p$ et $\mathcal{E}$ un échantillon de n individus.

- L'**effectif** de x_i est le nombre n_i de fois où la valeur x_i est prise dans l'échantillon.
- La **fréquence** de x_i est
$$f_i = \frac{n_i}{n}.$$
- La **fréquence cumulée** jusqu'à x_i est
$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^{i} f_j = \frac{1}{n}\sum_{j=1}^{i} n_j.$$

Exemple 2.5 Reprenons la série statistique de l'Exemple 2.3. Compléter le tableau de cette série statistique pour y faire apparaître les fréquences et les fréquences cumulées.

Valeur x_i	0	1	2	3	4	5
Effectif n_i	5	8	15	10	7	5
Fréquence f_i	$\frac{1}{10} = 0.1$	$\frac{4}{25} = 0.16$	$\frac{3}{10} = 0.3$	$\frac{1}{5} = 0.2$	$\frac{7}{50} = 0.14$	$\frac{1}{10} = 0.1$
Fréq. cum. F_i	0.1	0.26	0.56	0.76	0.90	1

Exemple 2.6 Reprenons la série statistique de l'Exemple 2.2. Compléter le tableau de cette série statistique pour y faire apparaître les fréquences et les fréquences cumulées.

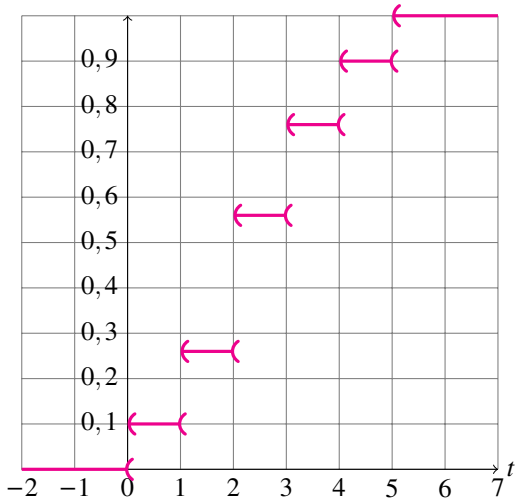
Valeur x_i	10	11	12	13	14	15	16	17	18	19	20
Effectif n_i	1	2	3	4	5	6	4	2	1	1	1
Fréquence f_i	$\frac{1}{30} \approx 0.33$	$\frac{1}{15}$	$\frac{1}{10}$	$\frac{2}{15}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{30}$
Fréq. cum. F_i	$\frac{1}{30}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{3}{4}$	...	...	...	...	1

2.2 Fonction de répartition

Définition 2.7 La **fonction de répartition** $F_x : \mathbb{R} \to \mathbb{R}$ d'une série statistique est la fonction en escalier correspondant aux fréquences cumulées F_i, c'est-à-dire,

$$\forall t < x_1,\ F_x(t) = 0, \qquad \forall i \in [\![1, p-1]\!], \forall t \in [x_i, x_{i+1}[,\ F_x(t) = F_i, \qquad \forall t \geq x_p,\ F_x(t) = 1$$

Exemple 2.8 Tracer la fonction de répartition associée à la série statistique de l'Exemple 2.3.



On peut remarquer que

- Pour tout $t \in \mathbb{R}$, $F_x(t) \in [0, 1]$
- $\lim\limits_{t \to -\infty} F_x(t) = 0$
- $\lim\limits_{t \to +\infty} F_x(t) = 1$
- La fonction F_x est croissante mais pas strictement croissante

2.3 Diagramme en barres, Diagramme des fréquences cumulées

Exemple 2.9 Reprenons la série statistique de l'Exemple 2.3.

Valeur x_i	0	1	2	3	4	5
Effectif n_i	5	8	15	10	7	5
Fréquence f_i	0.1	0.16	0.3	0.2	0.14	0.1
Fréq. cum. F_i	0.1	0.26	0.56	0.76	0.90	1

Représentons les données grâce à des **diagrammes en barre/en bâtons**. Par exemple, le diagramme en bâtons ci-dessous à gauche (resp. à droite) représente (sous forme de barres) les effectifs (resp. les fréquences) pour chaque valeur de la variable statistique étudiée. Les deux diagrammes ont la même allure mais avec des échelles différentes.

Diagramme en barre des effectifs

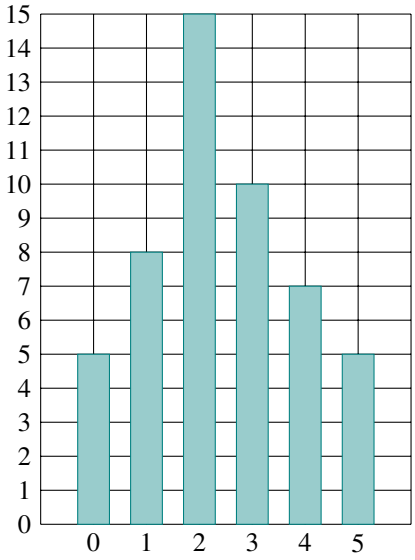
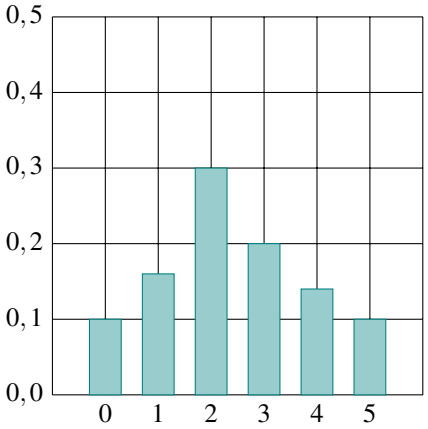
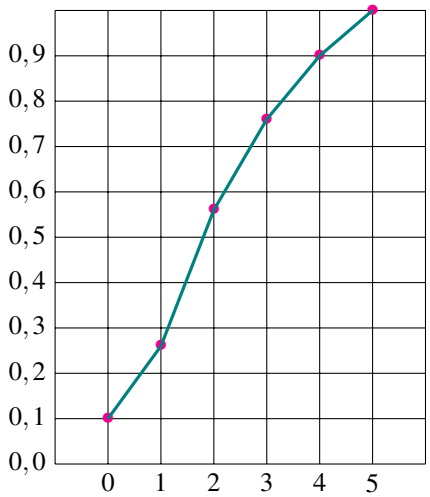


Diagramme en barre des fréquences



On peut aussi représenter les données grâce à un **diagramme des fréquences cumulées**. Sur ce graphique, on représente la courbe de la fonction qui passe par tous les points (x_i, F_i) (pour tout $i \in [\![1, p]\!]$) et qui relie ces points par une droite. Ce diagramme vise à montrer l'évolution du cumul de la fréquence lorsque la variable statistique étudiée croît.

Diagramme des fréquences cumulées



3 Indicateurs de tendance centrale

3.1 Moyenne

Définition 3.1 La **moyenne** de la série statistique

Valeur x_i	x_1	x_2	…	x_p	Total
Effectif n_i	n_1	n_2	…	n_p	N

est

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

où

- pour tout $i \in [\![1, p]\!]$, $f_i = \dfrac{n_i}{N}$ est la fréquence de la valeur x_i,
- et $n = n_1 + n_2 + \cdots + n_p$ est le nombre total d'individus dans l'échantillon.

Exemple 3.2 Reprenons la série statistique de l'Exemple 2.3.

Valeur x_i	0	1	2	3	4	5
Effectif n_i	5	8	15	10	7	5

Sa moyenne est donnée par

$$\bar{x} = \frac{5\times 0 + 8\times 1 + 15\times 2 + 10\times 3 + 7\times 4 + 5\times 5}{5+8+15+10+7+5} = 2.42$$

Dans l'entreprise, il y a en moyenne 2.42 personnes vivant avec les employés dans leurs foyers.

Exemple 3.3 Reprenons la série statistique de l'Exemple 2.2.

Valeur x_i	10	11	12	13	14	15	16	17	18	19	20
Effectif n_i	1	2	3	4	5	6	4	2	1	1	1

Quelle est la note moyenne obtenue par les élèves ?

La moyenne de cette série statistique est donnée par :

$$\bar{x} = \frac{1\times 10 + 2\times 11 + 3\times 12 + 4\times 13 + 5\times 14 + 6\times 15 + 4\times 16 + 2\times 17 + 1\times 18 + 1\times 19 + 1\times 20}{1+2+3+4+5+6+4+2+1+1+1}$$
$$= 14.5$$

Donc la note moyenne obtenue par les élèves est de 14.5.

Un des problèmes de la moyenne en tant qu'indicateur est qu'elle est très sensible aux valeurs extrêmes. Si on calcule les moyennes de ces deux séries statistiques suivantes,

Valeur x_i	1	2	3	300
Effectif n_i	2	4	2	0

Valeur x_i	1	2	3	300
Effectif n_i	2	4	1	1

on obtient

$$\bar{x} = 2 \quad \text{et} \quad \bar{x} = \frac{313}{8} = 39.125$$

Proposition 3.4 — Transformation affine. Soit x une variable statistique. Soient a et b deux réels et soit $y = ax + b$ une autre variable statistique, obtenue à partir de x par transformation affine. Alors $\bar{y} = a\bar{x} + b$.

Proposition 3.5 — Associativité. On suppose que l'échantillon $\mathcal{E}$ est séparé en deux échantillons $\mathcal{E}_a$ et $\mathcal{E}_b$ ($\mathcal{E} = \mathcal{E}_a \cup \mathcal{E}_b$). On note N_a l'effectif de $\mathcal{E}_a$ et N_b l'effectif de $\mathcal{E}_b$. On note $\bar{x}_a$ la moyenne de x sur l'échantillon $\mathcal{E}_a$ et $\bar{x}_b$ la moyenne de x sur l'échantillon $\mathcal{E}_b$. Alors, la moyenne globale $\bar{x}$ sur l'échantillon total est :

$$\bar{x} = \frac{N_a\bar{x}_a + N_b\bar{x}_b}{N_a + N_b}.$$

Cette formule permet

- soit de calculer la moyenne globale sur une population constituée de plusieurs groupes dont on a déjà calculé la moyenne,
- soit de calculer facilement la moyenne en cas d'ajout d'une observation.

Exemple 3.6 Le salaire moyen des huit 8 employés d'une entreprise est 30 000 euros. L'entreprise recrute deux nouveaux employés qualifiés dont le revenu moyen est de 100 000 euros. Calculer le nouveau revenu moyen de l'entreprise.

- Échantillon a

  Effectif $N_a = 8$ et Moyenne $\bar{x}_a = 30000$
- Échantillon b

  Effectif $N_b = 2$ et Moyenne $\bar{x}_b = 100000$

La moyenne de cette nouvelle série statistique est donnée par

$$\bar{x} = \frac{8 \times 30000 + 2 \times 1000000}{8 + 2} = 44000$$

Donc le nouveau revenu moyen de l'entreprise est de 44 000 euros.

3.2 Médiane

Définition 3.7 La **médiane** d'une variable statistique x sur un échantillon de données *rangées dans l'ordre croissant* est la valeur de x séparant les données de la série en deux sous-ensembles de tailles égales. Plus précisément, si n est la taille de l'échantillon,

- Si n est impair, la médiane est la valeur de rang $\frac{n+1}{2}$.
- Si n est pair, il y a deux valeurs centrales et la médiane est la moyenne entre ces deux valeurs.

Exemple 3.8 Reprenons la série statistique de l'Exemple 2.2.

Valeur x_i	10	11	12	13	14	15	16	17	18	19	20
Effectif n_i	1	2	3	4	5	6	4	2	1	1	1

Quelle est la médiane ?

Ici, la taille de l'échantillon est

$$n = 30.$$

La médiane est plus facile à calculer avec les données (rangées dans l'ordre croissant) plutôt qu'avec le tableau.

$$10 - 11 - 11 - 12 - 12 - 12 - 13 - 13 - 13 - 13 - 14 - 14 - 14 - 14 - \underline{14}$$
$$\underline{15} - 15 - 15 - 15 - 15 - 15 - 16 - 16 - 16 - 16 - 17 - 17 - 18 - 19 - 20$$

La médiane est donnée par

$$m_x = \frac{14 + 15}{2} = 14.5$$

Exemple 3.9 Donner la médiane de la série statistique suivante

$$0 - 0 - 1 - 1 - 2 - 2 - 4$$

La médiane de cette série est

$$m_x = 1.$$

Exemple 3.10 Donner la médiane de la série statistique suivante

$$0 - 3 - 2 - 1 - 4 - 2 - 1 - 0$$

Il s'agit d'abord de classer les données dans l'ordre croissant.

$$0 - 0 - 1 - \underline{1} - \underline{2} - 2 - 3 - 4$$

La médiane de cette série vaut alors

$$m_x = \frac{1+2}{2} = \frac{3}{2}$$

Contrairement à la moyenne, la médiane n'est pas sensible à la présence de valeurs extrêmes.

$1 - 1 - 2 - 2 - 2 - 2 - 3 - 3$ $\qquad$ $1 - 1 - 2 - 2 - 2 - 2 - 3 - 300$

on obtient, pour les deux séries, la même médiane qui vaut 2

Proposition 3.11 Soit x une variable statistique. Soient a et b deux réels et soit $y = ax + b$ une autre variable statistique, obtenue à partir de x par transformation affine. Notons m_x la médiane de x et m_y la médiane de y. Alors

$$m_y = am_x + b.$$

4 Indicateur de dispersion

4.1 Étendue

Définition 4.1 L'**étendue** d'une série statistique est la différence entre la plus grande valeur et la plus petite.

L'étendue est donc sensible aux valeurs aberrantes.

Valeur x_i	1	2	3
Effectif n_i	2	4	2

Valeur x_i	1	2	3	300
Effectif n_i	2	4	1	1

on obtient

$\text{etendue}_1 = 3 - 1 = 2$ et $\text{etendue}_2 = 300 - 1 = 299$

4.2 Quantiles

Définition 4.2 La notion de quantile généralise celle de médiane.

- Le **quantile d'ordre p** ($p \in [0,1]$) Q_p d'une variable statistique x est la plus petite valeur de la série telle que au moins p des valeurs prises par x sont inférieures ou égales à Q_p.
- Les **quartiles** sont les quantiles d'ordre $\frac{1}{4}$, $\frac{1}{2}$ et $\frac{3}{4}$.
- Les **déciles** sont les quantiles d'ordre $\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}$.
- L'**écart interquartile** est $Q_3 - Q_1$.

Exemple 4.3 Étudions la série statistique suivante.

Valeur x_i	0	1	2	5	6	7	10	Total
Effectif n_i	4	2	3	1	5	4	1	20

Pour s'aider, on peut aussi représenter les données sous forme brute (rangées dans l'ordre croissant).

$$0-0-0-0-\underline{1}-1-2-2-2-\underline{5}$$
$$\underline{6}-6-6-6-\underline{6}-7-7-7-7-10$$

Alors,

- $Q_1 = 1$
- La médiane vaut $m_x = \frac{5+6}{2} = 5,5$
- $Q_3 = 6$
- L'écart interquartile est $6-1=5$.

4.3 Variance et écart-type

Définition 4.4 La **variance** d'une série statistique x :

Valeur x_i	x_1	x_2	...	x_p	Total
Effectif n_i	n_1	n_2	...	n_p	n

est

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

où $\bar{x}$ est la moyenne de x et $f_i = \frac{n_i}{n}$ est la fréquence de la valeur x_i.

Proposition 4.5 — Formule de Koenig. La **variance** d'une série statistique x vaut aussi :

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

Définition 4.6 L'**écart-type** de x est :

$$s_x = \sqrt{s_x^2}.$$

Exemple 4.7 Étudions la série statistique suivante.

Valeur x_i	0	1	2	3	4	5	Total
Effectif n_i	5	8	15	10	7	5	50

- Sa moyenne vaut

$$\bar{x} = \frac{1}{50}(0\times5+1\times8+2\times15+3\times10+4\times7+5\times5) = \frac{121}{50} = 2,42$$

- Sa variance vaut

$$s_x^2 = \frac{1}{50}\left(0^2\times5+1^2\times8+2^2\times15+3^2\times10+4^2\times7+5^2\times5\right)-(2,42)^2 = 7,9-4,1764 = 2,0436$$

- Son écart-type vaut

$$s_x = \sqrt{s_x^2} \approx 1,43$$

5 Comparer des séries statistiques

5.1 Boîte à moustaches

Une **boîte à moustaches** (box plot en anglais) permet de représenter le minimum, Q_1, la médiane, Q_3 et le maximum d'une série statistique.

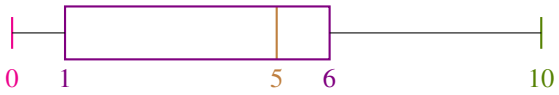
Exemple 5.1 Étudions la série statistique suivante.

Valeur x_i	0	1	2	5	6	7	10	Total
Effectif n_i	4	2	3	2	4	4	1	20

Représentons les données brutes rangées dans l'ordre croissant.

$$0-0-0-0-\underline{1}-1-2-2-2-\underline{5}$$
$$\underline{5}-6-6-6-\underline{6}-7-7-7-7-10$$

- Minimum = 0
- Premier quartile = 1
- Médiane = 5
- Troisième quartile = 6
- Maximum = 10



Exemple 5.2 Étudions la série statistique suivante.

Valeur x_i	0	1	3	5	6	8	10	Total
Effectif n_i	2	3	5	3	4	1	2	20

Représentons les données brutes rangées dans l'ordre croissant.

$$0-0-1-1-\underline{1}-3-3-3-3-\underline{3}$$
$$\underline{5}-5-5-6-\underline{6}-6-6-8-10-10$$

- Minimum = 0
- Premier quartile = 1
- Médiane = 4
- Troisième quartile = 6
- Maximum = 10

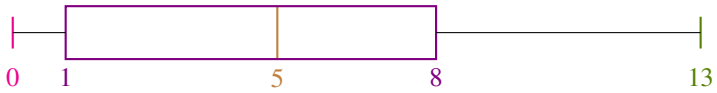
Exemple 5.3 Étudions la série statistique suivante.

Valeur x_i	0	1	2	5	7	8	10	13	Total
Effectif n_i	2	4	3	2	3	2	1	3	20

Représentons les données brutes rangées dans l'ordre croissant.

$$0-0-1-1-\underline{1}-1-2\ -2\ -2\ -\underline{5}$$
$$\underline{5}-7-7-7-\underline{8}-8-10-13-13-13$$

- Minimum = 0
- Premier quartile = 1
- Médiane = 5
- Troisième quartile = 8
- Maximum = 13



5.2 Moyenne & Écart-type

Une autre façon de comparer les séries statistiques est de comparer leurs moyenne et écart-type.

Exemple 5.4

Étudions la série statistique suivante.

Valeur x_i	0	1	2	5	6	7	10	Total
Effectif n_i	4	2	3	2	4	4	1	20

- Moyenne : $\bar{x} = 4$
- Variance : $s_x^2 = 25,2 - 4^2 = 9,2$
- Écart-type : $s_x = \sqrt{9,2} \approx 3,03$

Exemple 5.5

Étudions la série statistique suivante.

Valeur x_i	0	1	3	5	6	8	10	Total
Effectif n_i	2	3	5	3	4	1	2	20

- Moyenne : $\bar{x} = 4,25$
- Variance : $s_x^2 = 8,49$
- Écart-type : $s_x = 2,91$

Exemple 5.6

Étudions la série statistique suivante.

Valeur x_i	0	1	2	5	7	8	10	13	Total
Effectif n_i	2	4	3	2	3	2	1	3	20

- Moyenne : $\bar{x} = 5,3$
- Variance : $s_x^2 = 19,3$
- Écart-type : $s_x \approx 4,39$