Module : Mathématique de Base 4 **Classe :** 2ème année **AU : 2018 / 2019**

PEG : *Statistique descriptive*

Objectifs : à l'issue de ce **PEG** l'étudiant sera capable de :

- √ Acquérir les principales notations et techniques de statistique descriptive :

  Vocabulaire : Population, échantillon, individu

  Variables : variable qualitative, variable quantitative discrète et variable quantitative continue.

  Etude d'une variable : effectif partiel, effectif cumulé, fréquence partielle, fréquence cumulée, classe

  Paramètres de position : mode, médiane, moyenne, quartile.

  Paramètres de dispersion : étendue, variance, écart-type.

- √ Mettre en œuvre ces techniques : produire et interpréter les indicateurs et les graphiques...

Introduction

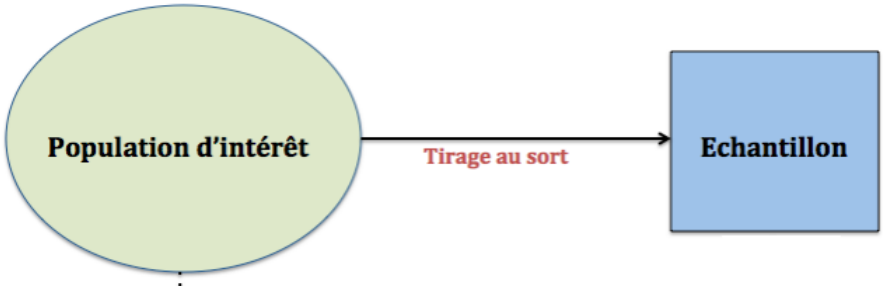
La statistique est à la fois une science, une méthode et un ensemble de techniques dont l'objectif est la compréhension et la gestion des phénomène complexes.

La statistique consiste à :

- ♦ Recueillir des données.
- ♦ Résumer, analyser et présenter ces données.
- ♦ Interpréter les résultats et tirer des conclusion afin d'aider à la prise de décision.
- ♦ En présence de données dépendant du temps, nous essayons de faire de la prévision.

Les statistiques consistent en diverses méthodes de classement des données tels que les tableaux, les histogrammes et les graphiques, permettant d'organiser un grand nombre de données.
Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie et bien sûr les sciences de l'ingénieur.

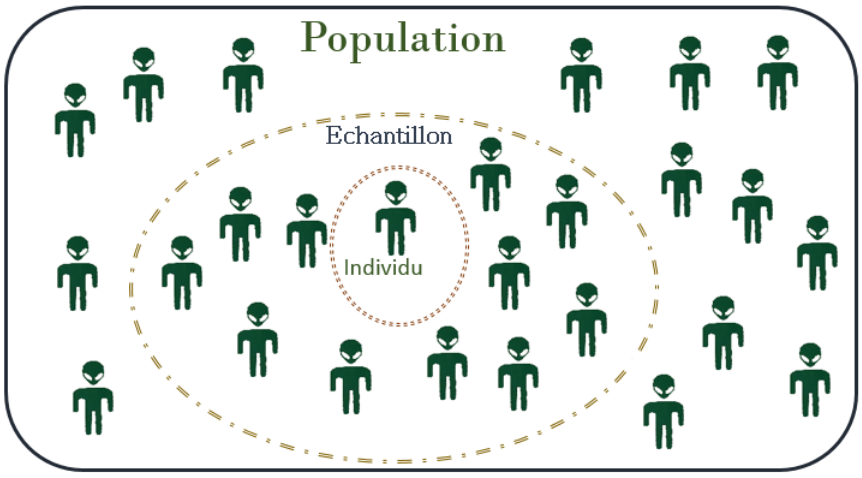
1 Vocabulaire :



Un bureau d'étude et statistique veut faire un sondage sur la population tunisienne. L'étude de toute la **population** est difficile en pratique vue que la taille est grande, en plus coûteux et prend beaucoup de temps, la solution est de prendre un **échantillon** de la population.

- ► **Population :** l'ensemble des individus concernés par l'étude ou le sondage.
  L'effectif total d'une population est noté N.

  Exemples : Population tunisienne, les clients d'une entreprise, les étudiants d'une école...



►**Echantillon :** Une partie de la population qu'on étudie.

Exemples : 5000 familles tunisiennes, 3 classes..

►**Individu :** l'unité d'observation.

Exemples : une famille tunisienne, une personne, un client, un étudiant...

Remarque : Il ne faut absolument pas confondre **échantillon** et **population** ! En statistique descriptive, on décrit un échantillon et non une population !

2 Variables :

Cas d'étude : On souhaite réaliser un sondage à ESPRIT pour connaître le temps de trajet moyen des étudiants qui utilisent les transports en commun pour venir à ESPRIT.

A chaque étudiant sondé, on demande :

- Son âge
- Sa classe
- Son temps moyen passé dans les transports en commun.

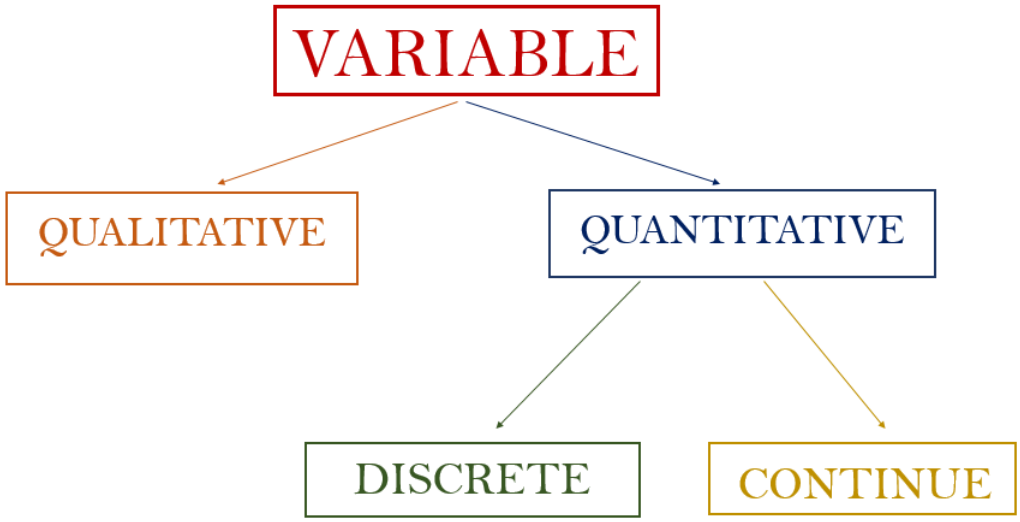
Essayons de définir correctement l'environnement de l'étude :

Population	Échantillon	Individus
Étudiants d'ESPRIT	une partie des étudiants d'ESPRIT	un étudiant d'ESPRIT

On a trois variables

Variable 1	Variable 2	Variable 3
âge	classe	temps moyen

►**Variable :** C'est la (les) caractéristique(s) de l'individu intégrant la population étudiée.



La collecte des informations lors d'une étude nous permet de former :

► **Une série statistique :** est l'ensemble des résultats d'une étude

2.1 Variables qualitatives

Cas d'étude : On considère comme population 100 nouveau-nés et le caractère est le sexe. On indexe les garçons par G et les filles par F. La série statistique : G : 63, F : 37.
Compléter le tableau suivant :

Variable	Type	Modalités
sexe	qualitative	F(fille)/G(garçon)

► **Variables qualitatives** sont des variables à valeurs non numériques qui expriment une qualité, un état, un statut unique, une catégorie..

Exemples :Le sexe, la profession, la couleur des yeux, le département, la classe...

2.2 Variables quantitatives discrètes

Cas d'étude : Une enquête réalisée dans un village porte sur le nombre d'enfants par famille. La série statistique donne

Familles avec 0 enfants	Familles avec 1 enfants	Familles avec 2 enfants	Familles avec 3 enfants
32	18	66	45

Compléter le tableau suivant :

Variable	Type
nombre d'enfant	quantitative discrète

- **Variables quantitatives discrètes** sont des variables à valeurs numériques dont les valeurs possibles sont finies ou dénombrables.

  Exemples : L'âge, le nombre d'enfants...

2.3 Variables quantitatives continues

Cas d'étude : On revient au sondage du "temps de trajet moyen" calculer en minutes, on obtient le tableau suivant :

Temps	entre 0 et 15 min	entre 15 et 30 min	entre 30 et 45 min	entre 45 et 60 min
Nbre d'étudiants	10	35	64	41

Compléter le tableau suivant :

Variable	Type
temps de trajet moyen	quantitative continue

- **Variables quantitatives continues** sont des variables à valeurs numériques, qui peuvent prendre toutes les valeurs possibles d'un intervalle.

  Exemples : taille, salaire, âge, ...

3 Étude d'une variable quantitative discrète

Pour cette section on considère les notations suivantes :

Ω : échantillon d'étude. $N = Card(\Omega)$: Nombre d'individus dans l'étude

et la variable quantitative discrète X définie par :

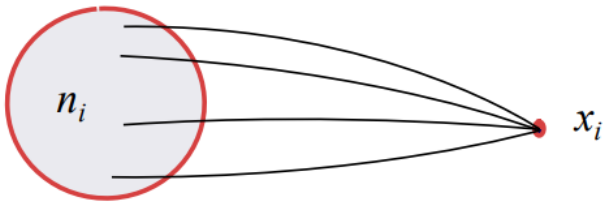
$$X \ : \ \Omega \ \rightarrow \ \{x_1, x_2, \ldots, x_k\}$$

3.1 Effectif partiel / Effectif cumulé :

Cas d'étude : Une enquête réalisée dans un village porte sur le nombre d'enfants à charge par famille. La série statistique est résumée dans le tableau suivant :

x_i	0	1	2	3	4	5	6
n_i	15	66	45	14	20	11	29

- **Effectif partiel :** n_i le nombre d'individus qui ont le même x_i.



Exemples : 45 est le nombre de familles qui ont 2 enfants.

Question : Quel est le nombre de familles qui ont un nombre d'enfant inférieur à 3 ?

$$15 + 66 + 45 + 14 = 140$$

- **Effectif cumulé :** N_i est le nombre d'individus dont la valeur du caractère est inférieur ou égale à x_i

$$N_i = \sum_{j \leq i} n_j$$

Question : Quel est l'effectif total ?

$$N = 15 + 66 + 45 + 14 + 20 + 11 + 29 = 200$$

- **Effectif total :** N est le nombre total des individus

$$N = \sum_{i} n_i = \sum_{i=1}^{n} n_i \qquad \text{pour } n \text{ valeurs possibles de } x_i$$

Exercice : On reste toujours avec le même cas d'étude, compléter le tableau

x_i	0	1	2	3	4	5	6
N_i	15	81	126	140	160	171	200

3.2 Fréquence partielle / fréquence cumulée :

Cas d'étude : On reste toujours avec le même cas d'étude,

x_i	0	1	2	3	4	5	6
n_i	15	66	45	14	20	11	29

Question : Quel est le pourcentage de familles qui ont un nombre d'enfant égale à

- 3 ? $\Rightarrow \dfrac{14}{200} = 0.07$
- 5 ? $\Rightarrow \dfrac{11}{200} = 0.055$

▶ **Fréquence parielle :** f_i le rapport de l'effectif partiel n_i sur l'effectif total N.

$$f_i = \frac{n_i}{N} \qquad \text{et} \qquad \sum_i f_i = 1$$

Exercice : On reste toujours avec le même cas d'étude, compléter le tableau

x_i	0	1	2	3	4	5	6
f_i	$\dfrac{15}{200} = 0.075$	$\dfrac{66}{200} = 0.33$	$\dfrac{45}{200} = 0.225$	$\dfrac{14}{200} = 0.07$	$\dfrac{20}{200} = 0.1$	$\dfrac{11}{200} = 0.055$	$\dfrac{29}{200} = 0.145$
f_i	7.5%	33%	22.5%	7%	10%	5.5%	14.5%

Cas d'étude : On reste toujours avec le même cas d'étude,

Question : Quel est le pourcentage de familles qui ont un nombre d'enfant inférieur à

- 2 ?$\Rightarrow 0.075 + 0.33 + 0.225 = 0.63 \Rightarrow 63\%$
- 4 ? $\Rightarrow 0.075 + 0.33 + 0.225 + 0.07 = 0.7 \Rightarrow 70\%$

▶ **Fréquence cumulée :** F_i

$$F_i = f_1 + f_2 + \cdots + f_i = \sum_{j \leq i} f_i = \sum_{j=1}^{i} f_j = \frac{N_i}{N}$$

Exercice : On reste toujours avec le même cas d'étude, compléter le tableau

x_i	0	1	2	3	4	5	6
F_i	7.5%	40.5%	63%	70%	80%	85.5%	100%

▶ **Fréquence totale :** F

$$F = \sum_i f_i = \frac{N}{N} = 1$$

3.3 Paramètre de position

Cas d'étude : On reste toujours avec le même cas d'étude,

x_i	0	1	2	3	4	5	6
n_i	15	66	45	14	20	11	29

Question : Quel est le nombre d'enfant par famille tel que l'effectif est le plus important ?
1 enfant par famille

▶ **Le mode** , noté M_0, est la valeur qui a le plus grand effectif partiel ou la plus grande fréquence partielle.

Question : Quel est la valeur x_i telle que au moins 50% des familles ont un nombre d'enfant inférieur à x_i ?
$x_i = 2$

▶ **La médiane** , noté M_e, est le nombre qui permet de couper la série(ordonnée par ordre croissant) en deux groupes de même effectif. Elle correspond à x_i tel que la fréquence cumulée est égale à 50% ou immédiatement $> 50\%$.

▶ **La moyenne** (moyenne arithmétique)

$$\overline{x} = \frac{1}{N}\sum_i n_i x_i = \sum_i f_i x_i$$

Exercice :Déterminer la médiane et calculer la moyenne de nombre d'enfant au village.

$M_e = 2$ et $\overline{x} = \dfrac{15\times 0 + 66\times 1 + 45\times 2 + 14\times 3 + 20\times 4 + 11\times 5 + 29\times 6}{200} = 2.535$

3.4 Paramètre de dispersion

▶ **L'étendue**

$$e = \max_i(x_i) - \min_i(x_i)$$

▶ **La variance**

$$V = \sum_i f_i(\overline{x} - x_i)^2 = \sum_i (f_i x_i^2 - \overline{x}^2)$$

▶ **L'écart-type**

$$\sigma = \sqrt{V}$$

Exercice : On reste toujours avec le même cas d'étude,

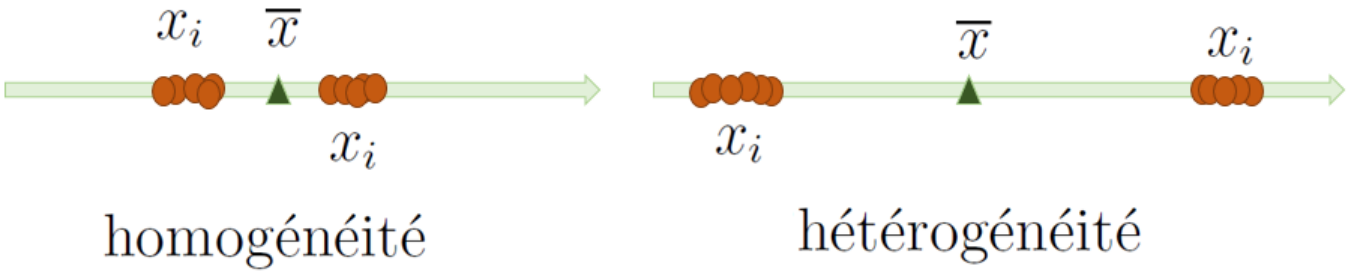
x_i	0	1	2	3	4	5	6
n_i	15	66	45	14	20	11	29

1. Calculer l'étendue. $e = 6$
2. Calculer la variance.
   $V = 0.075 * (2.535 - 0)^2 + 0.33 * (2.535 - 1)^2 + 0.225 * (2.535 - 2)^2 + 0.07 * (2.535 - 3)^2 + 0.1 * (2.535 - 4)^2 + 0.055 * (2.535 - 5)^2 + 0.145 * (2.535 - 6)^2 = 3.628$
3. Calculer l'écart-type. $\sigma = \sqrt{V} = 1.904$

Remarque : l'écart-type σ présente la distance entre les x_i et la moyenne $\overline{x}$, alors,

Plus σ est petite plus les x_i sont proches de la moyenne $\Rightarrow$ plus d'homogénéité.

Plus σ est grande plus x_i sont loin de la moyenne $\Rightarrow$ plus d'hétérogénéité.



4 Étude d'une variable quantitative continue

4.1 Classes

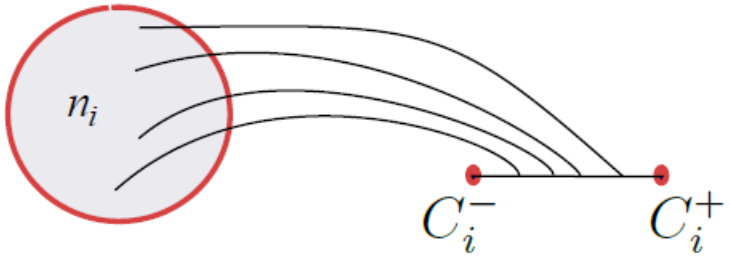
Cas d'étude : On mesure la taille en centimètres de 50 élèves d'une école, on obtient le tableau suivant :

x_i	152	152	153	153	154	154	154	155	155	156	156	156	156	156	164	167
159	159	160	160	160	161	160	160	161	162	157	157	157	158	158	165	162
168	168	168	169	169	170	171	171	171	171	162	163	164	164	164	166	152

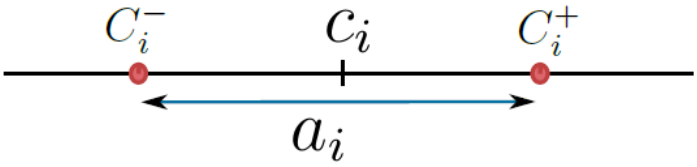
Question : Si on veut regrouper les valeurs en des intervalles de largeur $5cm$, quels sont les intervalles possible ?(Intervalle de type $[a, b[$)
$[152, 157[$, $[157, 162[$, $[162, 167[$, $[167, 172[$

On peut alors exprimer la série statistique

$[C_i^-$, $C_i^+[$	$[152, 157[$	$[157, 162[$	$[162, 167[$	$[167, 172[$
n_i	15	14	10	11



▶ **Classe de valeurs** C_i est tout intervalle de type $C_i = [C_i^-, C_i^+[$.



C_i^+ : Borne supérieure

a_i : Amplitude ou longueur,

$$a_i = C_i^+ - C_i^-$$

C_i^- : Borne inférieure

c_i : Centre

$$c_i = \frac{C_i^+ + C_i^-}{2}$$

Exercice : On reste toujours avec le même cas d'étude, compléter le tableau

$[C_i^-$, $C_i^+[$	$[152, 157[$	$[157, 162[$	$[162, 167[$	$[167, 172[$
n_i	15	14	10	11
c_i	154.5	159.5	164.5	169.5
a_i	5	5	5	5

Question : Quel est le nombre d'élèves qui ont une taille entre $152cm$ et $162cm$?$\Rightarrow 15+14 = 29$

▶ **Effectif partiel de** C_i **:** est le nombre n_i d'individus dont la valeur du caractère est dans C_i

- **Effectif cumulé de C_i :** est le nombre N_i d'individus dont la valeur du caractère est dans $\cup_{j\leq i} C_j$

$$N_i = \sum_{j\leq i} n_j$$

- **Effectif total :**

$$N = \sum_i n_i$$

Exercice : On reste toujours avec le même cas d'étude, calculer N et compléter le tableau

$[C_i^-\ ,\ C_i^+[$	$[152, 157[$	$[157, 162[$	$[162, 167[$	$[167, 172[$
n_i	15	14	10	11
N_i	15	29	39	50

Question : Quel est le pourcentage d'élèves qui ont une taille entre $162cm$ et $167cm$?
$\Rightarrow \dfrac{14}{50} = 0.28 = 28\%$

- **Fréquence partielle de C_i :** est le pourcentage d'individus dont la valeur du caractère est dans C_i

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

- **Fréquence cumulée de C_i :** est le pourcentage d'individus dont la valeur du caractère est dans $\cup_{j\leq i} C_j$

$$F_i = \sum_{j\leq i} f_j$$

Exercice : On reste toujours avec le même cas d'étude, compléter le tableau

$[C_i^-\ ,\ C_i^+[$	$[152, 157[$	$[157, 162[$	$[162, 167[$	$[167, 172[$
n_i	15	14	10	11
f_i	$\dfrac{15}{50} = 0.3 = 30\%$	$\dfrac{14}{50} = 0.28 = 28\%$	$\dfrac{10}{50} = 0.2 = 20\%$	$\dfrac{11}{50} = 0.22 = 22\%$
F_i	$0.3 = 30\%$	$0.58 = 58\%$	$0.78 = 78\%$	$1 = 100\%$

4.2 Paramètres de tendance centrale

- **La moyenne**

$$\overline{x} = \sum_i f_i c_i$$

Exercice : On reste toujours avec le même cas d'étude, calculer la moyenne de la série statistique.

$$\overline{x} = 0.3 * 154.5 + 0.28 * 159.5 + 0.2 * 164.5 + 0.22 * 169.5 = 161.5$$

Quelle est la classe avec l'effectif le plus grand ?$\Rightarrow [152, 157[$

- **Classe modale** est la classe avec l'effectif le plus grand.

On note C_k la classe modale, $C_k = [C_k^-, C_k^+[$, on définit,

- **Le mode** est la quantité :

$$M_0 = C_k^- + \frac{\Delta_1}{\Delta_1 + \Delta_2} a_k$$

Avec :

$$\begin{array}{rcl|rcl} \Delta_1 & = & n_k - n_{k-1} & \Delta_2 & = & n_k - n_{k+1} \\ \text{ou} & = & f_k - f_{k-1} & \text{ou} & = & f_k - f_{k+1} \end{array}$$

- **Classe médiane** est la classe contenant 50% d'effectif total .Elle correspond à C_i tel que la fréquence cumulée est égale à 50% ou immédiatement $> 50\%$.

On note C_l la classe médiane, $C_l = [C_l^-, C_l^+[$, on définit,

- **La médiane** est la quantité :

$$M_e = C_l^- + \frac{a_l}{f_{l+1}}(0.5 - F_l)$$

Exercice : On reste toujours avec le même cas d'étude, calculer le mode

$$\begin{array}{rclcrclcrcl} C_k & = & [152, 157[& , & C_k^- & = & 152 & , & C_k^+ & = & 157 \\ n_k & = & 15 & , & n_{k-1} & = & 0 & , & n_{k+1} & = & 14 \\ \Delta_1 & = & 15 - 0 = 15 & , & \Delta_2 & = & 15 - 14 = 1 & , & a_k & = & 5 \end{array}$$

On aura

$$M_0 = 152 + \frac{15}{15 + 1} * 5 = 156.6875$$

et la médiane

$$\begin{array}{rclcrclcrcl} C_l & = & [157, 162[& , & C_l^- & = & 157 & , & C_l^+ & = & 162 \\ f_{l+1} & = & 0.2 & , & F_l & = & 0.58 & , & a_l & = & 5 \end{array}$$

On aura

$$M_e = 157 + \frac{5}{0.2} * (0.5 - 0.58) = 155$$

4.3 Paramètres de dispersion

- **La variance**

$$V = \sum_i f_i(\overline{x} - c_i)^2 = \sum_i (f_i c_i^2 - \overline{x}^2)$$

- **L'écart-type**

$$\sigma = \sqrt{V}$$

Exercice : On reste toujours avec le même cas d'étude, calculer la variance $V = 0.3*(161.5-154.5)^2 + 0.28*(161.5-159.5)^2 + 0.2*(161.5-164.5)^2 + 0.22*(161.5-169.5)^2 = 31.21$
et l'écart-type $\sigma = \sqrt{V} = 5.58$

4.4 Les quartiles

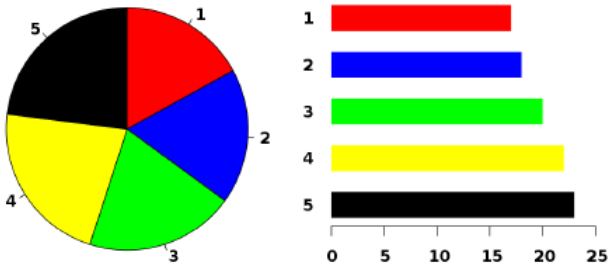
L'idée générale est de partager la population en quatre parties de même effectif. On appelle :

- 1^{er} **quartile :** Q_1 correspond à x_i tel que la fréquence est égale à 25% ou immédiatement $> 25\%$.
- 2^{eme} **quartile :** Q_2 correspond au médiane.
- 3^{eme} **quartile :** Q_3 correspond à x_i tel que la fréquence est égale à 75% ou immédiatement $> 75\%$.

5 Interprétation graphique

5.1 Variables qualitatives

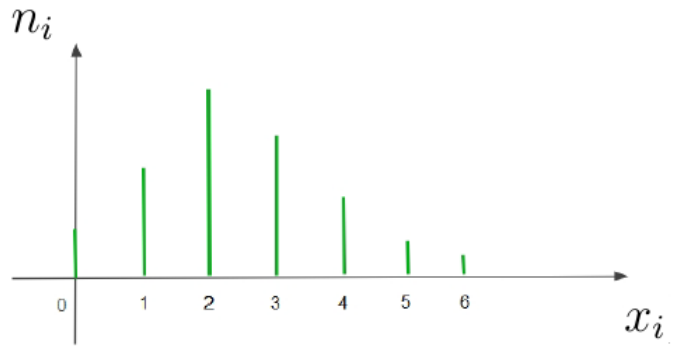
- **Diagramme par secteur** (diagramme circulaire)



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

5.2 Variables quantitatives discrètes

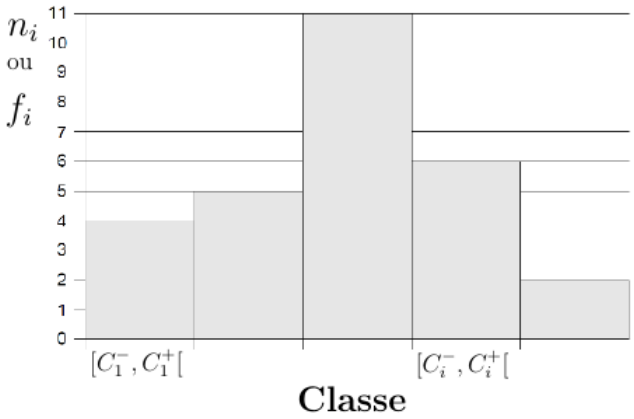
- **Diagramme à bâtons**



À chaque valeur correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés

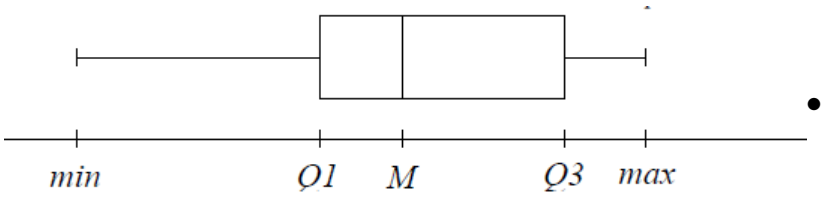
5.3 Variables quantitatives continues

▶ **Histogramme des fréquences (ou effectifs)**



On reporte les classes sur l'axe des abscisses et, au-dessus de chacune d'elles, on trace un rectangle dont l'aire est proportionnelle à la fréquence f_i (ou l'effectif n_i) associée.

▶ **Diagramme en boîtes**



- 25% de la population admet une valeur du caractère entre la valeur minimale et Q_1.
- 25% de la population admet une valeur du caractère entre Q_1 et M.
- 25% de la population admet une valeur du caractère entre M et Q_3.

Références :

"http ://math.univ-lyon1.fr/ chekroun/Files/chekroun-statistiques.pdf"