

## 1. Introduction

Les « *Statistiques* » sont la science des données, qui est à la fois une science, une méthode et un ensemble de techniques. Ceci implique la collection, la classification, le résumé, l'organisation, l'analyse et l'interprétation d'une information numérique. De ce fait, la science de la Biostatistique englobe à la fois : **(1)** La conception des expériences biologiques ; **(2)** La collecte des informations ; **(3)** La compilation et analyse des données chiffrées de ces expériences ; **(4)** L'interprétation des résultats en vue d'avancer une conclusion.

Cette science est exploitée alors dans plusieurs domaines: **(1)** La santé publique, y compris l'épidémiologie, les services de santé, la nutrition et l'environnement ; **(2)** La conception et analyse d'essais cliniques en médecine ; **(3)** La génomique, génétique des populations et la génétique statistique afin de relier la variation dans le génotype avec une variation dans le phénotype ; **(4)** L'agriculture afin d'améliorer les cultures et les animaux d'élevage ; **(5)** L'écologie en vue de mettre en place des prévisions écologiques ; **(6)** L'analyse de séquences biologiques.

<p>on emploie le terme de statistique pour désigner les données elles-mêmes « un ensemble de données numériques »          ainsi on parle des statistiques démographiques Économiques, sanitaires...les statistiques des étudiants .répartition par âge, sexe, année, spécialité...          généralement ce sont les tableaux et les chiffres qui figurent dans les bilans administratifs...</p>	<p><i>Au sens « scientifique »</i>          la statistique est l'ensemble des méthodes, des techniques à partir desquelles on  <b>Collecte, organise, résume, présente et analyse</b> des données. Ces techniques nous aident à tirer des conclusions et à prendre des <b>décisions judicieuses</b>. La méthode statistique étudie les phénomènes dont la caractéristique principale est la <b>variabilité</b>.</p>
---	---

## 2. LA STATISTIQUE DESCRIPTIVE UNIVARIÉE

### 2.1. Définition :

La statistique descriptive est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. Il est assez compliqué de définir la meilleure description possible d'un phénomène.

Dans le cadre des statistiques, il s'agira de fournir toute l'information disponible sur le phénomène en moins de chiffres et de mots possibles.

Cette démarche a pour but de :

- Résumer et synthétiser l'information contenue dans la série statistique ;
- Mettre en évidence ses propriétés ;
- Suggérer des hypothèses relatives à la population dont est issu l'échantillon.

Les Outils utilisés sont :

- **Les Tableaux ;**
- **Les Graphiques ;**
- **Les indicateurs.**

Le type d'outils utilisés dépend de :

- La nature de la série (uni ou multidimensionnelle) ;
- La nature des variables (quantitatives ou qualitatives).

Si les données ne sont relatives qu'à *une seule* variable, on parle de statistique descriptive « **univariée** ». Dans le cas où l'on s'intéresse à *deux* variables simultanément, on met en oeuvre la statistique descriptive « **bivariée** ». Si l'ensemble de données provient de l'observation de *plusieurs* variables, on doit faire appel aux méthodes de la statistique descriptive « **multivariée** ».

## 2.2. Notions de bases

**2.2.1. Élément :** C'est une unité qui peut être :

- **Un individu** (êtres vivants) : humain, animal, végétal ... ;
- **Un sujet** : modules enseignés en biologie, les nationalités, les métiers ou professions... ;
- **Un objet** : Table, chaise, verrerie de laboratoire ;
- **Une association** (dans les études écologiques en général) : une parcelle d'herbe, une association d'arbustes...

### 2.2.2. Population

C'est un ensemble d'éléments possédant au moins une caractéristique commune et exclusive permettant de l'identifier et la distinguer sans ambiguïté de toutes les autres.

**Exemple**

- Une population algérienne ;
- Une population estudiantine ;
- Une population de plantes médicinales ;
- Une population de poissons d'eau douce.

**2.2.3. Echantillon**

Pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments de la population. En outre, si cette opération est possible il est rarement utile de la faire, car l'analyse d'un groupe restreint d'éléments extraits de la population fournit généralement des résultats de précision satisfaisante. Cette petite partie de la population qu'on va examiner s'appelle « échantillon ».

**Exemple**

- Etude de 20 étudiants pris à partir d'une population de 57.
- Etude de 5 régions prises à partir d'une population de 25.
- Etude de 5 modules pris à partir d'une population de 13.
- Etude de 200 patients pris à partir d'une population de 660.

**2.2.4. Echantillonnage**

C'est l'opération ou la méthode qui consiste à prélever une partie de la population (échantillon). Il existe plusieurs méthodes d'échantillonnage qui varient en fonction de la nature de l'étude envisagée.

Le plus utilisé est l'échantillonnage aléatoire et simple qui est basé sur le principe que tous les éléments de la population ont une probabilité égale (non nulle) de faire partie de l'échantillon. C'est une méthode d'échantillonnage permettant de choisir  $n$  unités parmi les  $N$  de la population de façon aléatoire.

**2.2.5. Caractère****LES CARACTERES**

L'étude portera sur un ou plusieurs caractères présentés par chacun des individus de la population : Taille, poids, taux de glycémie, couleur des yeux, profession, nationalité, groupe sanguin, nombre d'enfants par famille...

**Exemple**

- Couleur des yeux ;
- Poids des souris ;
- Superficie d'une pièce ;
- La température de l'air.

**Un caractère est dit quantitatif** quand ses différentes modalités sont mesurables (exprimées par des chiffres).  
Ex : tailles, poids, nombres d'enfants par famille...

**Un caractère est dit qualitatif** quand ses différentes modalités échappent à la mesure.

Ex : couleur des yeux, profession, nationalité, groupe sanguin...

Le caractère **qualitatif** peut être

**nominal**: le sexe, le groupe sanguin...

**ordinal** : la douleur, le niveau d'instruction...

**2.2.6. Modalité**

Ce sont les diverses situations (cas, état, valeur) susceptibles d'être prises par le caractère. Un caractère peut posséder une ou plusieurs modalités.

**Exemple**

- Couleur des yeux : vert, bleu, noir ;
- Poids des souris (en grammes) : 15, 18, 20, 39 ;
- Superficie d'une pièce (en mètres) : 3, 5, 6 ;
- La température de l'air (en °C) : 8, 16, 27, 30, 38.

**2.2.6.1. Effectif**

Plusieurs définitions sur les effectifs.

- L'effectif de la valeur  $x_i$  est le nombre d'individus de la population ayant cette valeur ou appartenant à cette classe : on le note  $n_i$  ;

- L'effectif total  $N$  est la somme de tous les effectifs, appelé également en Maths « cardinal ou la taille » ;
- En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer l'effectif cumulé croissant en faisant la somme des effectifs de cette valeur et de tous ceux qui la précèdent.

**Exemple**

Dans une promotion de 20 étudiants de Biochimie, voici les notes obtenues au dernier examen de Biostatistique : 10, 14, 12, 15, 7, 8, 10, 11, 12, 18, 2, 4, 12, 13, 14, 15, 19, 11, 9, 0.

On va calculer les effectifs et les effectifs cumulés.

- Premièrement, les effectifs :
- Combien d'étudiants ont eu 10 ? c'est 2.
- Combien d'étudiants ont eu 12 ? c'est 3.

On continue ainsi et on forme le tableau suivant (Tableau I):

**Tableau I**

Notes $x_i$	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Effectifs $n_i$	1	1	1	1	1	1	2	2	3	1	2	2	1	1

Les effectifs cumulés maintenant. On fait la somme des effectifs de la note + la somme des effectifs de toutes les notes qui la précèdent. Ce qui nous donne (Tableau II):

**Tableau II** : Les effectifs  $n_i$  et  $N_i \hat{=}$  relatifs à chaque modalité  $x_i$  (notes)

Notes $x_i$	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Effectifs $n_i$	1	1	1	1	1	1	2	2	3	1	2	2	1	1
Effectifs $N_i \hat{=}$	1	2	3	4	5	6	8	10	13	14	16	18	19	20

**Remarque**

Pour vérifier qu'on ne s'est pas trompé dans le calcul des effectifs cumulés, on vérifie bien que le dernier effectif cumulé correspond bien au nombre d'individus  $N_i$ . Ici, on retrouve bien 20, le nombre d'étudiants de cette promotion de Biochimie.

**2.2.6.2. Fréquence d'une modalité ou d'une classe**

- La fréquence d'une valeur est le quotient de l'effectif de la valeur par l'effectif total ;

- En rangeant les valeurs du caractère dans l'ordre croissant, on peut calculer les fréquences cumulées croissantes en faisant la somme des fréquences de cette valeur et de tous ceux qui la précèdent ;
- Pour les fréquences cumulées croissantes, c'est un peu le même principe que pour les effectifs cumulés croissants.

**Remarque**

Les fréquences sont comprises entre 0 et 1.

**Exemple**

On reprend l'exemple précédent et on applique tout simplement la formule des fréquences pour les calculer (Tableau III).

**Tableau III :** Les fréquences  $f_i$  et  $F_i \hat{=}$  relatives à chaque modalité  $x_i$  (notes)

Notes $x_i$	0	2	4	7	8	9	10	11	12	13	14	15	18	19
Fréquences $f_i$	0,05	0,05	0,05	0,05	0,05	0,05	0,1	0,1	0,15	0,05	0,1	0,1	0,05	0,05
Fréquences $F_i \hat{=}$	0,05	0,1	0,15	0,2	0,25	0,3	0,4	0,5	0,65	0,7	0,8	0,9	0,95	1

**Remarque :**

Pareil, pour vérifier qu'on ne s'est pas trompé dans le calcul des fréquences cumulées, on vérifie bien que la dernière fréquence cumulée vaut bien 1.

**2.3. Nature de caractère**

Le caractère d'une série statistique est l'objet de notre étude sur une population donnée. Ce caractère prend un certain nombre de valeurs ou de modalités.

On distingue deux grandes familles de caractères :

- Les caractères qualitatifs ;
- Les caractères quantitatifs.

Un critère commode est le suivant : la moyenne d'un caractère quantitatif a un sens alors que la moyenne d'un caractère qualitatif est impossible à réaliser ou bien n'a aucune signification.

### 2.3.1. Caractère qualitatif

Un caractère est dit qualitatif lorsque ses modalités ne sont pas mesurables. Le nombre de valeurs que peut prendre la variable est limité. Il existe au sein de ce type deux échelles :

nominale et ordinale.

#### **Echelle nominale**

Chaque modalité est exprimée par un nom ou un code. Les différentes modalités ne sont pas ordonnables.

#### ***Exemple : cas des noms***

- Etat matrimoniale : marié, célibataire, veuf, divorcé ;
- Sexe : féminin, masculin ;
- Profession : enseignant, médecin ;
- Nationalité : Algérienne, Tunisienne ;
- Les différentes séquences nucléotidiques ;
- Les hormones : oestradiol, progestérone.

#### ***Exemple : cas des codes***

- Etat matrimoniale : marié (1), célibataire (2), veuf (3), divorcé (4) ;
- Sexe : féminin (1), masculin (2) ;
- Profession : enseignant (1), médecin (2) ;
- Nationalité : Algérienne (1), Tunisienne (2) ;

#### **Echelle ordinale**

Chaque modalité est explicitement significative du rang pris par chaque individu pour le caractère considéré.

#### ***Exemple :***

- Degré d'intelligence : pas intelligent (0), peu intelligent (1), moyennement intelligent (2), très intelligent (3) ;
- Forme des fruits : petite (1), moyenne (2), grosse (3) ;

- Abondance/Dominance : peu abondant (1), abondant (2), très abondant (3).

### 2.3.2. Caractère quantitatif

Un caractère est quantitatif si ses modalités s'expriment par des nombres. Le nombre de valeurs que peut prendre la variable est illimité.

*Exemple :*

- Latitude, longitude, température, altitude

Il peut être discret ou continu

Enfin, on peut distinguer les caractères quantitatifs discrets et les caractères quantitatifs continus selon que leurs modalités (valeurs) sont définies sur un intervalle continu de l'ensemble des réels (modalités en nombre infini) ou selon qu'elles correspondent à un ensemble fini et dénombrable de valeurs entières ou réelles.

#### a. Les caractères quantitatifs discrets

Sont des caractères dont les modalités sont des nombres isolés, pas nécessairement entiers.

*Exemple :*

- Nombre de pièces d'un immeuble ;
- Nombre d'enfants d'une famille ;
- Nombre des doigts

#### b. Les caractères quantitatifs continus

Sont des caractères dont les modalités sont définies sur un intervalle (continu) de valeur donné appelé domaine de variation et défini par les valeurs minimales et maximales.

*Exemple :*

- Notes des étudiants, la taille, le poids, l'âge.

## 2.4. Tableaux statistiques

Un tableau statistique constitue un résumé ou une synthèse numérique des résultats d'une distribution statistique, on distingue trois formes de tableaux statistiques qui sont fonction de l'objectif envisagé et de la nature du caractère étudié.



**2.4.1. Tableaux élémentaire (brut)**

<b>Eléments <math>i</math></b>	<b>Modalités <math>x_i</math></b>
1	$x_1$
2	$x_2$
3	$x_3$
·	·
·	·
·	·
$k$	$x_k$

Après la collecte des données, celles-ci apparaissent de façon brute. Sous cette forme, elles sont peu informatives. Il nous faut donc des moyens pour en extraire un maximum d'informations.

**Cas qualitatif**

Dans une étude d'extraction de substances végétales, on fait un sondage pour savoir ce que chacun des 25 étudiants/es a trouvé le plus intéressant. On utilise la règle suivante

<b>Réponse</b>	<b>X</b>
Les huiles essentielles	1
Les flavonoïdes	2
Les polyphénols totaux	3
Les tannins	4
Autres	5

Les résultats obtenus sont les suivants

<b>E</b>	<b>R</b>	<b>E</b>	<b>R</b>	<b>E</b>	<b>R</b>	<b>E</b>	<b>R</b>	<b>E</b>	<b>R</b>
E1	3	E6	2	E11	2	E16	4	E21	2
E2	3	E7	4	E12	2	E17	3	E22	2
E3	5	E8	3	E13	3	E18	4	E23	3
E4	4	E9	4	E14	5	E19	4	E24	5
E5	2	E10	5	E15	5	E20	3	E25	1

**Cas quantitatif**

Voici les résultats du dosage des polyphénols obtenus par 25 étudiants de Biochimie par des méthodes d'extraction différentes (unité en  $\mu\text{g/g} * 100$ )

Tableau des résultats des doses en polyphénols obtenus par des méthodes d'extraction différente (E : Etudiant, D : Doses en polyphénols)

<b>E</b>	<b>D</b>	<b>E</b>	<b>D</b>	<b>E</b>	<b>D</b>	<b>E</b>	<b>D</b>	<b>E</b>	<b>D</b>
E1	68	E6	65	E11	78	E16	55	E21	88
E2	74	E7	52	E12	66	E17	61	E22	68
E3	42	E8	41	E13	49	E18	72	E23	90
E4	47	E9	57	E14	59	E19	56	E24	63
E5	50	E10	65	E15	60	E20	79	E25	69

**2.4.2. Tableau de dénombrement**

Modalités $x_i$	Effectifs $n_i$	Fréquences relatives $f_i$
$x_1$	$n_1$	$f_1$
$x_2$	$n_2$	$f_2$
$x_3$	$n_3$	$f_3$
·	·	·
·	·	·
·	·	·
$x_k$	$n_k$	$f_k$
Total	$n$	1

**2.4.3. Tableau de distribution des fréquences**

Il est de la forme

Classes	Centre $c_i$	Effectifs $n_i$	Fréquences relatives $f_i$
$[e_{i-1}, e_{i-1} + a_i[$	$c_1$	$n_1$	$f_1$
$[e_{i-1} + a_i, e_{i-1} + 2a_i[$	$c_2$	$n_2$	$f_2$
$[e_{i-1} + 2a_i, e_{i-1} + 3a_i[$	$c_3$	$n_3$	$f_3$
·	·	·	·
·	·	·	·
·	·	·	·
	$c_k$	$n_k$	$f_k$
Total	-	$n$	1

*Exemples*

**a. Tableau statistique relatif à un caractère qualitatif**

Analyse de sang pour 100 étudiants

Groupe sanguin $x_i$	Effectifs $n_i$	Fréquences relatives $f_i$
A	40	0.40
B	43	0.43
Ab	12	0.12
O	05	0.05
Total	100	1

**b. Tableau statistique relatif à un caractère quantitatif**

• **Cas discret**

On observe 20 lots au laboratoire, on a le nombre de lapins dans chacun : 10, 1, 0, 0, 0, 10, 12, 18, 5, 5, 12, 10, 12, 12, 0, 10, 15, 10, 20, 20

$x_i$	0	1	5	10	12	15	18	20	T
$n_i$	4	1	2	5	4	1	1	2	20
<b>fi</b>	0.20	0.05	0.10	0.25	0.20	0.05	0.05	0.10	1

**Nombre de classes**

Il existe plusieurs formules pour le calcul du nombre de classes. Pour les applications de ce cours, nous avons opté pour la **Règle de STURGE - La formule mathématique de HUNTSBERGER**

$$k = 1 + 3.3 \log N .$$

**Intervalle de classe (amplitude)**

$$a_i = \frac{x_{\max} - x_{\min}}{k} ,$$

**Centre d'une classe**

$$c_i = \frac{e_{i-1} + e_i}{2}$$

**Application**

• **Cas continu**

On s'intéresse à la taille (cm) de 20 étudiants, les résultats obtenus sont

La taille de 20 étudiants (cm)

140	144	150	156
142	146	152	157
143	147	153	158
143	148	154	159
144	150	155	163

Dans ce cas, on doit regrouper cette série en classes

*Nombre de classe :*

$$k = 1 + 3.3 \log n = 1 + 3.3 * \log 20 = 1 + 3.3 * 1.30 = 5.29 \approx 5$$

*Intervalle de classe = amplitude*

$$a_i = \frac{v_{\max} - v_{\min}}{k} = \frac{163 - 140}{5} = 4.6 \approx 5$$

*Centres des classes*

$$x_1 = \frac{e_{i-1} + e_i}{2} = \frac{140 + 145}{2} = 142.5$$

$$x_2 = \frac{e_{i-1} + e_i}{2} = \frac{145 + 150}{2} = 147.5$$

$$x_3 = \frac{e_{i-1} + e_i}{2} = \frac{150 + 155}{2} = 152.5$$

$$x_4 = \frac{e_{i-1} + e_i}{2} = \frac{155 + 160}{2} = 157.5$$

$$x_5 = \frac{e_{i-1} + e_i}{2} = \frac{160 + 165}{2} = 162.5$$

Tableau de la distribution des fréquences des tailles des étudiants

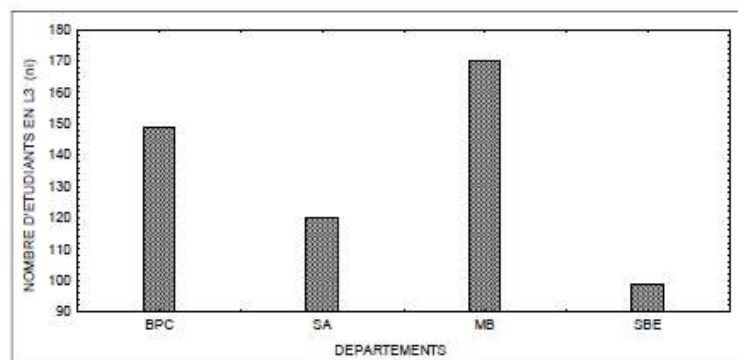
Classes	$x_i$	$n_i$	$f_i$
[140-145[	142.5	6	0.30
[145-150[	147.5	3	0.15
[150-155[	152.5	5	0.25
[155-160[	157.5	5	0.25
[160-165[	162.5	1	0.05
Total	-	20	1

## 2.5. Représentation graphique

### 2.5.1. Cas qualitatif

#### 2.5.1.1. Diagramme en tuyaux d'orgue

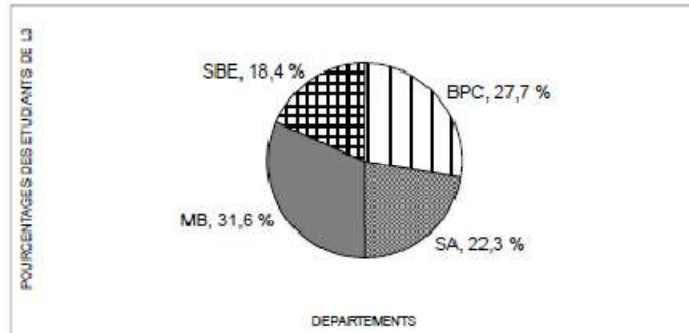
Le diagramme en bâtonnets (ou tuyaux d'orgue) est une représentation graphique de la distribution de fréquences d'une variable qualitative. Les «bâtonnets» sont bien séparés pour indiquer les différentes catégories. La hauteur d'un bâtonnet est proportionnelle à la fréquence de la catégorie correspondante.



**Figure 1** : Tuyaux d'orgue illustrant le nombre d'étudiants en L3 par département (SA : Science Alimentaire, MB : Microbiologie, SBE : Science Biologique de l'Environnement, BPC : Biologie Physico-Chimique).

**2.5.1.2. Le camembert**

Dans le diagramme circulaire, chaque secteur a une surface proportionnelle à la fréquence de chaque modalité (Figure 2).



**Figure 2 :** Le camembert illustrant le nombre d'étudiant en L3 en fonction des départements (SA : Science Alimentaire, MB : Microbiologie, SBE : Science Biologique de l'Environnement, BPC : Biologie Physico-Chimique).

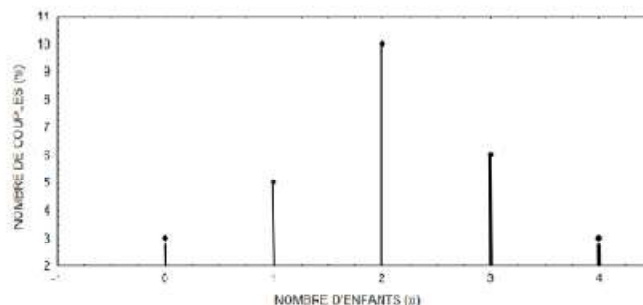
**2.5.2. Cas quantitatif**

Il existe deux types de représentation graphique d'une distribution statistique à caractère quantitatif :

- Le diagramme différentiel correspond à une représentation des effectifs ou des fréquences.
- Le diagramme intégral correspond à une représentation des effectifs cumulés, ou des fréquences cumulées.

➤ **Variable statistique discrète.**

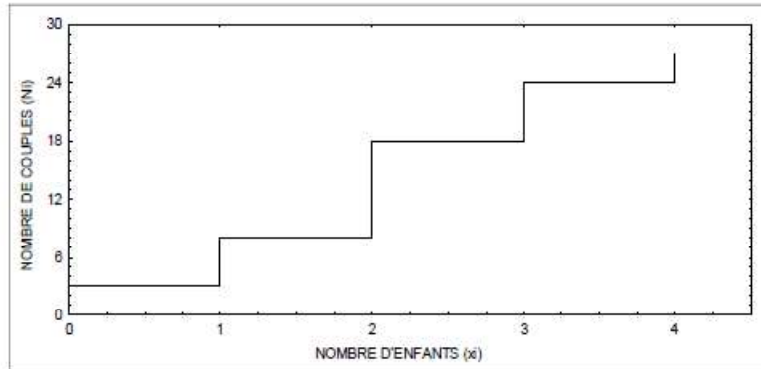
- 1. Diagramme différentiel :** *Diagramme en bâtons* : est réalisé en fonction des effectifs ou des fréquences.



**Figure 3 :** Diagramme différentiel illustrant le nombre de couples en fonction du nombre d'enfants.

## 2. Diagramme intégral

*Courbe en escalier* : est réalisée en fonction des effectifs cumulés ou des fréquences cumulées. Dans cette représentation les effectifs ou les fréquences des diverses valeurs de la variable statistique correspondent aux hauteurs des marches de la courbe.



**Figure 4** : Diagramme intégral illustrant le nombre de couples en fonction du nombre d'enfants.

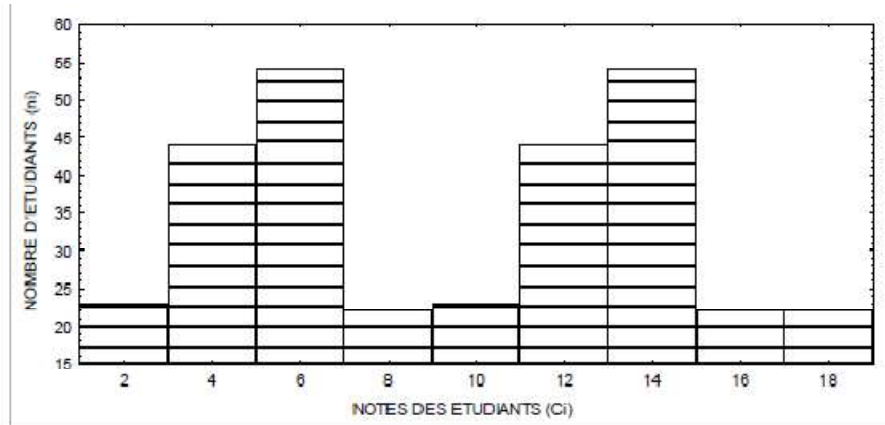
### ➤ Variable statistique continue

#### 1. Histogramme et polygone des effectifs ou des fréquences

*L'histogramme* est une représentation graphique (en tuyaux d'orgue) de la distribution des effectifs ou des fréquences d'une variable quantitative. Souvent, les «tuyaux» sont accolés pour montrer la continuité de la variable. La hauteur du tuyau est proportionnelle à l'effectif ou la fréquence de la classe correspondante (Figure 5).

*Le polygone des effectifs ou des fréquences* : est une autre représentation graphique (en ligne brisée) de la distribution des effectifs ou des fréquences d'une variable quantitative.

Pour tracer le polygone, on joint les points milieu du sommet des rectangles adjacents par un segment de droite. Le polygone est fermé aux deux bouts en le prolongeant sur l'axe horizontal.

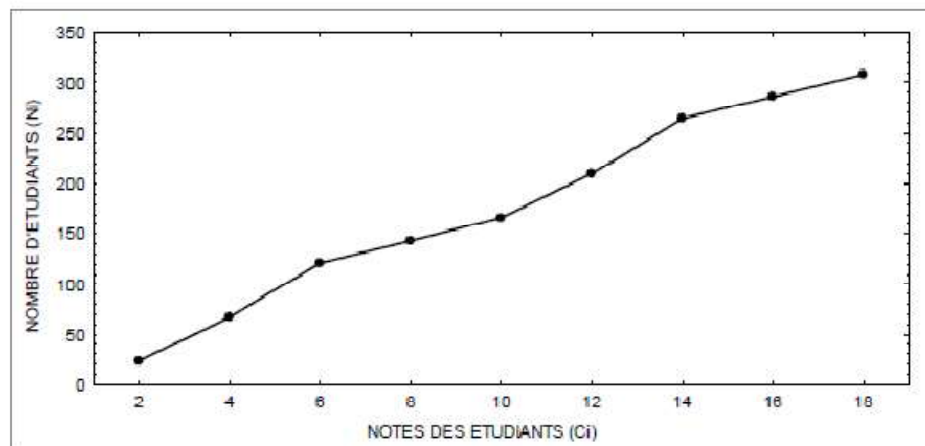


**Figure 5** : Histogramme illustrant le nombre d'étudiants en fonction de leurs notes.

## 2. Diagramme intégral

Représentation graphique intégrale en fonction des effectifs cumulés ou des fréquences cumulées appelée parfois ogive (Figure 6).

Une telle figure fournit des données descriptives intéressantes telles que la valeur médiane, i.e. le point séparant le groupe en deux parties égales.



**Figure 6** : Diagramme intégral illustrant le nombre d'étudiants en fonction de leurs notes.