Historiquement, à l'origine, la statistique rassemblait des renseignements sur les populations. Les premiers recensements permettaient de connaître le nombre d'habitants d'un pays et leur répartition par sexe, par âge, par catégorie socioprofessionnelle, selon l'état matrimonial, selon le lieu de résidence, …

Les méthodes statistiques sont aujourd'hui utilisées dans tous les domaines : économie (évaluation des ressources et des stocks), médecine (évaluation de l'efficacité d'un médicament, l'état sanitaire d'une population), industrie (contrôle de qualité), agronomie (recherche d'engrais spécifique, sélection de variétés,…), démographie, éducation, psychologie et tellement d'autres.

La statistique descriptive est un ensemble de méthodes permettant de décrire, résumer et synthétiser l'information contenue dans des données nombreuses. Pour ce faire, les outils utilisés sont : les tableaux, les graphiques et les indicateurs (paramètres).

Remarque : Ne pas confondre les statistiques avec la statistique. Les statistiques sont :

- les données chiffrées ou autres renseignements en grands nombres sur lesquels on va appliquer les méthodes statistiques ou
- les résultats numériques auxquels conduit l'application des méthodes statistiques.

Vocabulaire (Terminologie)

- **Population et individus**

La population ou univers est l'ensemble que l'on observe (ensemble que l'on étudie, ensemble sur lequel l'étude statistique est faite) et dont chaque élément est appelé individu ou unité statistique.

Le terme population se rapporte à des ensembles de toute nature :

- Ensembles d'êtres humains :
  - ✓ La population de l'Algérie à la date du dernier recensement,
  - ✓ La population des étudiants de l'université d'Alger année 2020/2021,
  - ✓ Le personnel de l'entreprise X au 30 décembre 2018.
- Ensemble d'êtres vivants :
  - ✓ La population des poissons d'un lac,
  - ✓ Population des plantes.
- Stock d'objets concrets :
  - ✓ Le parc automobile Algérien au 1er janvier 2016,
  - ✓ L'ensemble des livres de statistique descriptive en Algérie.
- Flux : Les demandes d'emplois déposées à l'ANEM (Agence Nationale de l'Emploi) au cours du mois de janvier 2019.
- Ensembles non concrets :
  - ✓ Les accidents de la route dus à l'excès de vitesse enregistrés sur une période.
  - ✓ Les intentions de vote à telle élection recueillies lors d'une enquête.

- **Echantillon**

Un échantillon de taille n est une partie (sous-ensemble) de n individus de la population.

On étudie un échantillon de la population notamment lorsque celle-ci est impossible à étudier dans son ensemble; c'est le cas pour les sondages ou pour des mesures rendant inutilisable les objets étudiés, par exemple la durée de vie de piles électriques d'un certain type.

- **Caractère**

Un caractère ou une variable est une propriété (phénomène) observée dans une population donnée ou dans un échantillon considéré, par exemple : la taille, le poids, le nombre d'enfants dans une famille, le sexe, l'état matrimonial, la catégorie socioprofessionnelle.

➢ **Modalités d'un caractère**

Les modalités d'un caractère correspondent aux différentes situations que peut prendre ce caractère (valeur possibles du caractère)

Par exemple :

- ✓ Le caractère sexe a pour modalités : Féminin (F) ou Masculin (M),
- ✓ Le caractère groupe sanguin présente les modalités : O, A, B, AB,
- ✓ Le caractère nombre d'absences d'un étudiant a pour modalités : 0, 1,2,…
- ✓ Le caractère durée de vie d'une pile électrique prend ses valeurs dans $[0, +\infty[$.

➢ **Types de caractères**

On distingue deux types de caractères :

1. **Caractère qualitatif :** Un caractère est qualitatif s'il n'est pas mesurable (il est lié à une observation ne pouvant faire l'objet d'une mesure). Ses diverses modalités sont simplement constatées, repérées par un mot traduisant un état (ses modalités ne sont pas des nombres).
   Exemples **:** le sexe (F,M), la couleur des yeux (vert, bleu, noisette, marron), la situation matrimoniale (célibataire, marié, divorcé, veuf), groupe sanguin (0, A, B, AB), état de santé (très bonne, bonne, moyenne, mauvaise santé), degré de satisfaction relativement à un produit (non satisfait, satisfait, très satisfait).

- ➢ Un caractère qualitatif est dit ordinal si ses modalités sont ordonnées (il existe une hiérarchie naturelle entre les modalités).
  Exemples : l'état de santé, la qualité de l'eau, la mention du bac obtenu.
- ➢ Un caractère qualitatif est dit nominal s'il n'existe pas de relation d'ordre sur les modalités.
  Exemples : le sexe, la couleur des yeux.

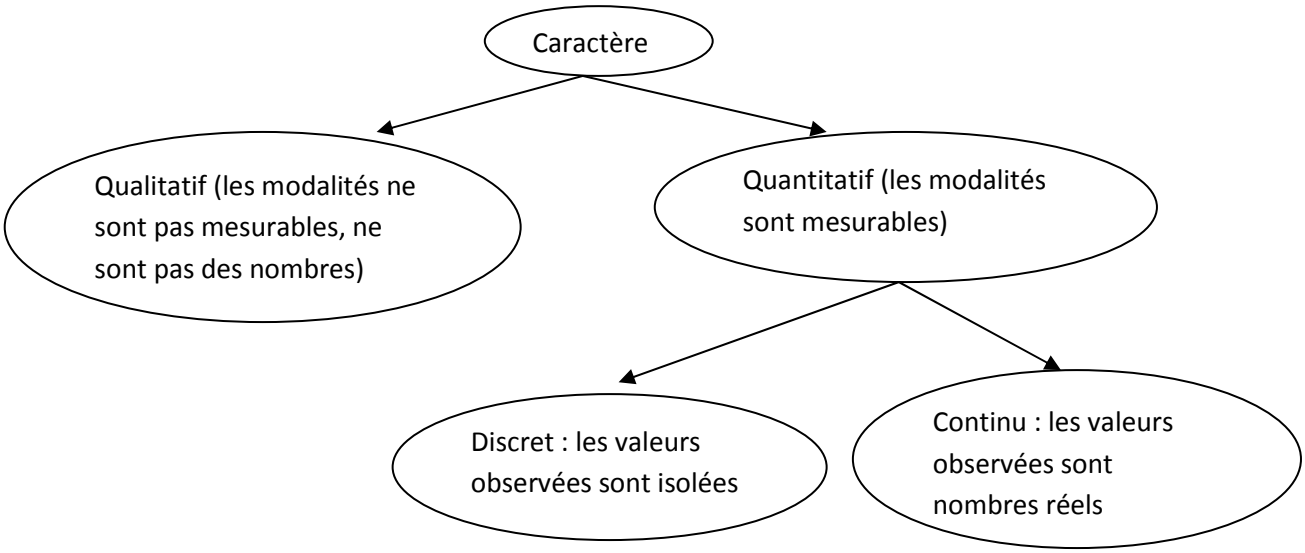
2. **Caractère quantitatif :** un caractère est dit quantitatif si on peut le mesurer ou le compter. Le caractère quantitatif prend alors le nom de variable statistique et ses différentes modalités sont les valeurs possibles de la variable.
   Exemples : le poids, le nombre d'enfants.

- ➢ Un caractère quantitatif est dit discret s'il ne peut prendre que des valeurs isolées (l'ensemble des valeurs possibles est fini ou infini dénombrable).
  Exemples : le nombre de voitures par ménage, le nombre d'absences d'un étudiant : 0, 1, 2,3,…

- ➢ Un caractère quantitatif est dit continu s'il peut prendre toute valeur d'un intervalle réel.

Exemples : le poids d'un individu, la taille, l'âge, la durée de vie d'une pile électrique.

En général les grandeurs liées à l'espace (longueur, surface), au temps (âge vitesse), à la masse (poids, teneur), à la monnaie (salaire, chiffre d'affaires) sont considérées comme des variables statistiques continues.



- **Série statistique**

Pour une variable quantitative, notée X, si à chacun des éléments de l'échantillon de taille n, on fait correspondre la valeur du caractère étudié, on obtient un ensemble de nombres appelé série statistique $X_1, X_2, \dots, X_n$ où X_i est la valeur du caractère pour le $i$ème individu de l'échantillon.

- Les valeurs de la série statistique peuvent être rangées par ordre de grandeur par exemple croissante On obtient alors une série ordonnée $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ où $X_{(1)} = X_{min}$ est la plus petite valeur observée (valeur minimale) et $X_{(n)} = X_{max}$ est la plus grande valeur observée (valeur maximale). Les valeurs X_{min} et X_{max} sont appelées les valeurs extrêmes.

- Etendue de la série : La différence e entre les deux valeurs extrêmes est appelée étendue de la série

$$e = X_{max} - X_{min}.$$

- **Classes**

Dans le cas d'une variable statistique continue, les individus sont regroupés en classes, définies par leurs bornes (extrémités). La $i$ème classe formée est de la forme $[e_{i-1}, e_i[$ où e_{i-1} et e_i sont respectivement la borne inférieure et la borne supérieure de la classe. Pour la dernière classe, l'extrémité droite peut être comprise dans la classe.

Le centre x_i et l'amplitude a_i de la classe $[e_{i-1}, e_i[$ sont définis comme suit :

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

$$a_i = e_i - e_{i-1}.$$

Si la série statistique a été partagée en k classes de même amplitude a, on a la relation $\boldsymbol{ka \geq e}$**.**

Le choix du nombre k de classes est fonction du nombre d'observations n. Le nombre k indiqué de classes pour une série de n observations est donné approximativement par

Formule 1 : $k \simeq \sqrt{n}$,

Formule 2 : $k \simeq E(5log_{10}n)$, où $E(x)$ est la partie entière de x et $log_{10}n = \frac{\ln(n)}{\ln(10)}$ (ln étant le logarithme népérien),

Formule 3 (formule de Struges) : $k \simeq 1 + 3.3222\ log_{10}n$,

Formule 4 (formule de Yule) : $k \simeq 2.5\sqrt[4]{n} = 2.5\ n^{\frac{1}{4}}$.

Il s'agit ensuite de choisir les amplitudes des classes. On les choisit généralement égales, d'amplitude approximativement égale à $a = {}^{e}/_{k}$.

Exemple : Une série de 50 valeurs de longueurs a pour valeurs extrêmes : $X_{max} = 44.2\ cm$, $X_{min} = 12.5\ cm$.

Etendue de la série : $e = X_{max} - X_{min} = 31.7$.

Nombre de classes à retenir : $k \simeq \sqrt{n} = \sqrt{50} = 7.07,$ on prend $k = 7$.

Amplitude des classes : $a \geq {}^{e}/_{k} = {}^{31.7}/_{7} = 4.528$, on prend par exemple $a = 5$.

Les classes seront donc (en cm)

$[12.5\ ,17.5[$, $[17.5, 22.5[$, $[22.5\ ,27.5[$, $[27.5, 32.5[$, $[32.5, 37.5[$, $[37.5, 42.5[$ $[42.5\ ,47.5[$ ([42.5,47.5])

de centres $x_1 = 15,\ x_2 = 20\ , x_3 = 25,\ x_4 = 30,\ x_5 = 35,\ x_6 = 40,\ x_7 = 45.$

On peut aussi choisir comme classes :

$[12\ ,17[$, $[17, 22[$, $[22\ ,27[$, $[27, 32[$, $[32\ ,37[$, $[37, 42[$ $[42\ ,47[$

ou

$[10\ ,15[$ $[15\ ,20[$ $[20\ ,25[$ $[25\ ,30[$ $[30\ ,35[$ $[35\ ,40[$ $[40\ ,45[$.
12.5 17.5 22.5 27.5 32.5 37.5 42.5

Dans tout les cas, il faut que la valeur minimum de la série appartienne à la première classe et la valeur maximum appartienne à la dernière classe.

Remarque

Les données se présentent, au départ, sous forme d'une liste éventuellement très longue et sans autre classement que l'ordre d'arrivée des informations. Aussi, pour faciliter leur lecteur, est-on amené à les présenter de manière synthétique sous forme de tableau ou de graphique.

Tableau statistique :

- **Effectif total** : L'effectif total est le nombre d'individus appartenant à la population (ou bien la taille de l'échantillon), noté n.
- **Effectif** (ou fréquence absolue) : L'effectif d'une modalité est le nombre d'individus présentant cette modalité et l'effectif d'une classe est le nombre d'observations dans cette classe.

  L'effectif de la $i$ème modalité ($i$ème classe) est noté n_i. L'effectif total n est tel que

  $$n = n_1 + \cdots + n_k = \sum_{i=1}^{k} n_i,$$

  k étant le nombre de modalités (classes).

- **Fréquence** (ou fréquence relative) : La fréquence d'une modalité est la proportion d'individus présentant cette modalité et la fréquence d'une classe est la proportion d'individus appartenant à cette classe.

  La fréquence f_i de la $i$ème modalité (de la $i$ème classe) est donnée par

  $$f_i = \frac{n_i}{n}, \qquad 0 \le f_i \le 1, \;\; i = 1, \dots, k.$$

  Notons que

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

On synthétise les modalités (classes), les effectifs et les fréquences sous forme de tableaux statistiques de la forme

Modalité x_i (Classe $[e_{i-1}, e_i[$)	Effectif n_i	Fréquence f_i
x_1	n_1	$f_1 = \frac{n_1}{n}$
⋮	⋮	.
x_i	n_i	. $f_i = \frac{n_i}{n}$
⋮	⋮	.
x_k	n_k	. $f_k = \frac{n_k}{n}$
somme $\sum$	n	1

Exemple1 : On a relevé le groupe sanguin sur un échantillon de 10 patients. Les résultats sont les suivants :

O	A	AB	O	B	A	A	AB	AB	AB

Caractère : groupe sanguin

Nature : caractère qualitatif

Modalités : O, A, B, AB

Modalité x_i	Effectif n_i	Fréquence f_i
O	2	0.2
A	3	0.3
B	1	0.1
AB	4	0.4
$\sum$	$n = 10$	1

Exemple 2 : On a comptabilisé le nombre d'absences de chacun des étudiants d'un groupe de 32. On a obtenu la série statistique suivante :

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

Série ordonnée : 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 4

Etendue de la série : 4-0=4

Caractère étudié : nombre d'absences

Nature : quantitatif discret

Modalités : $0, 1, 2, 3, 4$

Modalité x_i	Effectif n_i	Fréquence f_i
0	5	0.15625
1	10	0.3125
2	11	0.31375
3	5	0.15625
4	1	0.03125
$\sum$	32	1

Exemple3 : Le taux de glycémie déterminé chez 32 sujets et donné ci-dessous en g/l.

Série ordonnée : 0.85 0.95 1.00 1.06 1.13

0.87 0.97 1.01 1.07 1.14

0.90 0.97 1.03 1.08 1.14

0.93 0.98 1.03 1.08 1.15

0.94 0.98 1.03 1.10 1.17

0.94 0.99 1.04 1.10 1.19

1.11 1.20

Caractère : taux de glycémie

Nature : quantitatif continu

Valeurs extrêmes : $X_{min} = 0.85$, $X_{max} = 1.20$

Etendue de la série : $e = X_{max} - X_{min} = 0.35$

Pour $n = 32$, le nombre de classes k à retenir est 6 ($k = \sqrt{32} = 5.65, on\ prend\ donc\ 6$) et l'amplitude a de chaque classe est égale à $a = 0.06$ ($\frac{e}{k} = \frac{0.35}{6} = 0.0583,\ a \geq \frac{e}{k}$)

Classe	Effectif n_i	Fréquence f_i
[0.85, 0.91 [	3	0.09375
[0.91, 0.97 [	4	0.125
[0.97, 1.03 [	7	0.21875
[1.03, 1.09 [	8	0.25
[1.09, 1.15 [	6	0.1875
[1.15, 1.21 [	4	
$\sum$	32	1

25% des patients ont un taux de glycémie compris entre 1.03 et 1.09 g/l.

Dans le cas d'un caractère quantitatif (variable statistique), on a les notions d'effectif cumulé et de fréquence cumulée.

- **Effectif cumulé** : On appelle effectif cumulé de la $i$ème modalité (ou de la $i$ème classe) le nombre N_i défini par

$$N_i = \sum_{j=1}^{i} n_j .$$

On note aussi n_{icum} , $\tilde{n}_i$.

On a $N_1 = n_1$ et $N_k = n$ (k est le nombre de modalités ou de classes) .

Les N_i sont appelés effectifs cumulés ascendants

- **Fréquence cumulée** : On appelle fréquence cumulée de la $i$ème modalité (ou de la $i$ème classe) le nombre F_i défini par

$$F_i = \frac{N_i}{n} = \frac{1}{n}\sum_{j=1}^{i} n_j = \sum_{j=1}^{i} \frac{n_j}{n} = \sum_{j=1}^{i} f_j .$$

On note aussi f_{icum} , $\tilde{f}_i$.
On a $F_1 = f_1$ et $F_k = 1$.
Les F_i sont appelées fréquences cumulées ascendantes.

Remarque : On définit aussi les effectifs cumulés descendants N_i' par $N_i' = \sum_{j=i}^{k} n_j$ et les fréquences cumulées descendantes F_i' par $F_i' = \sum_{j=i}^{k} f_j$.

Exemple : Nombre d'absences

Modalité x_i	Effectif n_i	Fréquence f_i	Effectif cumulé N_i	Fréquence cumulée F_i	N_i'	F_i'
0	5	0.15625	5	0.15625	32	1
1	10	0.3125	15	0.46875	27	0.84375
2	11	0.31375	26	0.8125	17	0.53125
3	5	0.15625	31	0.96875	6	0.1875
4	1	0.03125	32	1	1	0.03125
$\sum$	32	1				

26 étudiants ont au plus 2 absences, 46.875% des étudiants ont au plus 1 absence

6 étudiants ont au moins 3 absences, 53.125% des étudiant ont au moins 2 absences

Exemple : Taux de glycémie

Classe	Effectif n_i	Fréquence f_i	N_i	F_i
[0.85, 0.91 [	3	0.09375	3	0.09375
[0.91, 0.97 [	4	0.125	7	0.21875
[0.97, 1.03 [	7	0.21875	14	0.4375
[1.03, 1.09 [	8	0.25	22	0.6875
[1.09, 1.15 [	6	0.1875	28	0.875
[1.15, 1.21 [	4	0.125	32	1
$\sum$	32	1		