Université Mohamed Khider Biskra **Département de Mathématiques**
Faculté des SE&SNV **1ière Année MI**
Module: Introduction aux probabilités et statistique descriptive **2022/2023**

Statistique descriptive
Représentation des données

1.Le vocabulaire des statistiques

Le but de l'analyse statistique est d'évaluer une population à partir d'échantillons tirés de cette population

1.1 Population

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie, étude des populations humaines, a occupé une place centrale aux débuts de la statistique, notamment au travers des recensements de population. Mais, en statistique, le terme de population s'applique à tout objet statistique étudié, qu'il s'agisse d'étudiants (d'une université ou d'un pays), de ménages ou de n'importe quel autre ensemble sur lequel on fait des observations statistiques. Nous définissons la notion de population..

Définition 1
On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω.

Exemple
– On considère l'ensemble des étudiants de la section A. On s'intéresse aux nombre de frères et soeurs de chaque étudiant. Dans ce cas

$$\Omega = ensemble\ des\ étudiants.$$

– Si l'on s'intéresse maintenant a la circulation automobile dans une ville, la population est alors constituée de l'ensemble des véhicules susceptibles de circuler dans cette ville à une date donnée. Dans ce cas

$$\Omega = ensemble\ des\ véhicules.$$

1.2 Individu (unité statistique)

Une population est composée d'individus. Les individus qui composent une population statistique sont appelés unités statistiques..

Définition 2
On appelle individu tout élément de la population Ω, il est noté ω (ω dans Ω).

Remarque

- L'ensemble peut être un ensemble de personnes, de choses ou d'animaux...
- L'unité statistique est un objet pour lequel nous sommes intéressés à recueillir de l'information.

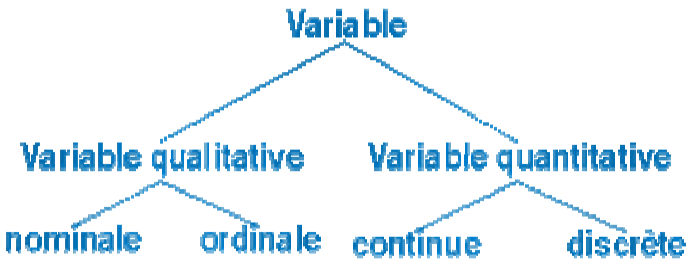
2.Les caractères (Variables) statistiques
La statistique « descriptive », comme son nom l'indique cherche à décrire une population donnée. Nous nous intéressons au caractéristique des unités qui peuvent prendre différentes valeurs.

Définition 3
On appelle caractère (ou variable statistique, dénotée V.S) toute application

$$X : \Omega \longrightarrow C$$

L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré ou observé sur les individus)

Exemple:
Taille, température, nationalité, couleur des yeux, catégorie socioprofessionnelle ...



2.1.Les caractères (Variables) qualitatifs:
Une variable statistique est qualitative si ses valeurs, ou modalités, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne,somme, ... , n'ont pas de sens

- Mesurées dans une échelle **nominale**, les modalités sont exprimables par des noms et ne sont pas hiérarchisées.
- Mesurées dans une échelle **ordinale**: les modalités traduisent le degré d'un état caractérisant un individu sans que ce degré ne puisse être défini par un nombre qui résulte d'une mesure. Les modalités sont alors hiérarchisées.

Exemple(nominale): la couleur du pelage, les groupes sanguins, les différents nucléotides de l'ADN, la présence ou l'absence d'un caractère (dichotomique), etc.

Exemple(ordinale): le stade d'une maladie.

2.2.Les caractères (Variables) quantitatifs:

- s'il ne prend que des valeurs "isolées", on dit qu'il est **discret**.
- s'il prend des valeurs sur un intervalle, on dit qu'il est **continu**.

Exemple (**caractère discret):** le nombre de petits par portée, le nombre de cellules dans une culture, le nombre d'accidents pour une période donnée, etc.

Exemple (**caractère continu)**: le poids, la taille, le taux de glycémie, le rendement, etc.

3 Représentation des données:
Il existe plusieurs niveaux de description statistique : la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques et des résumés numériques fournis par un petit nombre de paramètres caractéristiques.

3.1 Tableaux statistiques

- **Fréquences cumulées:** On appelle fréquences cumulées ou fréquences relatives cumulées en x_i, le nombre $f_i cum$ tel que

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

- **Effectifs cumulées:**On appelle effectifs cumulées ou fréquences relatives cumulées en x_i, le nombre n_{cum} tel que

$$n_{icum} = \sum_{p=1}^{i} n_p$$

Remarque: On peut noter que $\sum_{i=1}^{k} n_i = n$, taille de l'échantillon et $\sum_{i=1}^{k} f_i = 1$.

3.1.2 Tableau statistique d'une variable qualitative

En présence d'une variable qualitative X pouvant prendre K modalités $x_1, x_2, ..., x_K$, on commence par réaliser un tri à plat, c'est à dire faire l'inventaire des modalités ou valeurs rencontrées dans la série, avec les effectifs correspondants. On calcule ensuite les **proportions** (ou **fréquence**) de chaque modalité en divisant l'effectif de chaque modalité par l'effectif total,. On construit donc un tableau de la forme :

$$f_k = \frac{n_k}{n}$$

Modalités	**Effectifs**	**fréquences**
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k

Exemple1:si on a un tableau de la forme

Numéro de Cliente	Signalétique
1	M.
2	Mme
3	Mlle
$\vdots$	$\vdots$
627630	Mme

$\Longrightarrow$

Signalétique	Nombre de Clientes	Proportions
M.	60985	0,0972
Mme	424641	0,6766
Mlle	142004	0,2262
Total	627630	1

On va par un tri plat construire un tableau de la forme:

Variable Signalétique

3.1.2 Tableau statistique d'une variable quantitative:

Le tableau brut se présente sous la forme suivante:

Individu	**Variable**
1	x_1
2	x_2
$\vdots$	$\vdots$
n	x_n

Objectif: créer un tableau plus synthétique.

Cas des variables discrètes : on étudie une variable discrète X à p modalités dans une population de taille n.

Modalités	x_1	x_2	$\cdots$	x_p
Effectifs	n_1	n_2	$\cdots$	n_p
Fréquences: $f_i = \frac{n_k}{n}$	f_1	f_2	$\cdots$	f_p

Exemple: La cécidomyie du hêtre provoque sur les feuilles de cet arbre des galles dont la distribution de fréquences observées est la suivante:

x_i	0	1	2	3	4	5	6	7	8
n_i	182	98	46	28	12	5	2	3	0
$f_i = \frac{n_k}{n}$	0.485	0.261	0.123	0.075	0.032	0.013	0.005	0.006	0
$f_i cum$	0.485	0.746	0.869	0.944	0.976	0.989	0.994	1	1

avec:
x_i :le ombre de galles par feuille
n_i : nombre de feuilles portant x_i galles

Cas des variables continues : on regroupe les individus par classes. On décompose l'intervalle des valeurs possibles en une partition d'intervalles.
Soit p le nombre d'intervalles. Les données se présentent sous la forme suivante:

X	n_i	X_i	$N_i \nearrow$	$F_i \nearrow$	$N_i \searrow$
$[a_0\ ,\ a_1]$	n_1	$\frac{a_0+a_1}{2}$	$N_1 = 0$	$F_1 = N_1/n$	n
$[a_1\ ,\ a_2]$	n_2	$\frac{a_1+a_2}{2}$	$N_2 = 0 + n_1$	$F_2 = N_2/n$	$n - n_1$
$[a_2\ ,\ a_3]$	n_3	$\frac{a_2+a_3}{2}$	$N_3 = 0 + n_1 + n_2$	$F_3 = N_3/n$	$n - n_1 - n_2$
$\vdots$					
$[a_{i-1}\ ,\ a_i]$	n_i	$\frac{a_{i-1}+a_i}{2}$	$N_i = 0 + n_1 + ... + n_i$	$F_i = N_i/n$	$n - n_1 - ... - n_i$
$\vdots$					
$[a_{m-1}\ ,\ a_m]$	n_m	$\frac{a_{m-1}+a_m}{2}$	$N_m = 0 + n_1 + ... + n_{m-1}$	$F_m = N_m/n$	$n - n_1 - ... - n_{m-1}$
Σ	n	–	n	1	0

Classes	**Effectifs**	**Centres de classe**	**Fréquences:** $f_i = \frac{n_k}{n}$
$[e_0, e_1[$	n_1	c_1	f_1
$[e_1, e_2[$	n_2	c_2	f_2
$[e_2, e_3[$	$\vdots$	$\vdots$	$\vdots$
$[e_3, e_4[$	n_p	c_p	f_p

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau de fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe
Diverses formules empiriques permettent d'établir le nombre de classes pour un échantillon de taille n.

- La règle de **STURGE** : Nombre de classes: $k = 1 + 3,332\,(\log n)$
- La règle de **YULE** : Nombre de classes: $k = 2,5\,(\sqrt[4]{n})$
- L'intervalle entre chaque classe est obtenu ensuite de la manière suivante:

Intervalle de classe: $C = (X_{max} - X_{min})/k$

avec X_{max} et X_{min}, respectivement la plus grande et la plus petite valeur de X dans la série statistique..

Exemple3:
Dans le cadre de l'étude de la population de gélinottes huppées (Bonasa umbellus), les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante:
$n = 50$ avec $X_{max} = 174$ et $X_{min} = 140$

Nombre de classes:

$$\underline{k} = 1 + 3.332(\log n) = 1 + 3,332\,(\log 50) = 6,60$$

Intervalle de classe:

$$c = (X_{max} - X_{min})/Nombredeclasses = \frac{174 - 140}{6.6} = 5,15 \approx 5$$

Caractère X : x_i : longueur de la rectrice bornes des classes	[140-145[	[145-150[	[150-155[	[155-160[	[160-165[	[165-170[	[170-175[
Valeur médiane des classes, x_i'	142,5	147,5	152,5	157,5	162,5	167 ,5	172,5
n_i : nombre d'individu par classe de taille x_i	1	1	9	17	16	3	3
f_i : fréquence relative	0,02	0,02	0,18	0,34	0,32	0,06	0,06
$f_i cum.$: fréquence relative cumulée	0,02	0,04	0,22	0,56	0,88	0,94	1

3.3 Représentations graphiques

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution. Elles facilitent l'interprétation des données recueillies.

- **3.3.1 Cas d'une variable qualitative:** peut se représenter à l'aide de trois types de diagrammes:

1. Un diagramme en camembert (ou diagramme à secteurs)
2. Tuyaux d'orgue.(Un diagramme en bâtons)
3. Le diagramme à barre (Diagramme en blocs).

Un diagramme en camembert (ou diagramme à secteurs):
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.

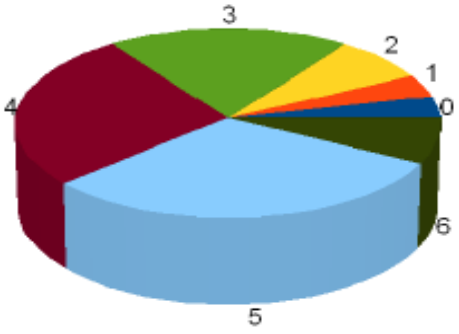
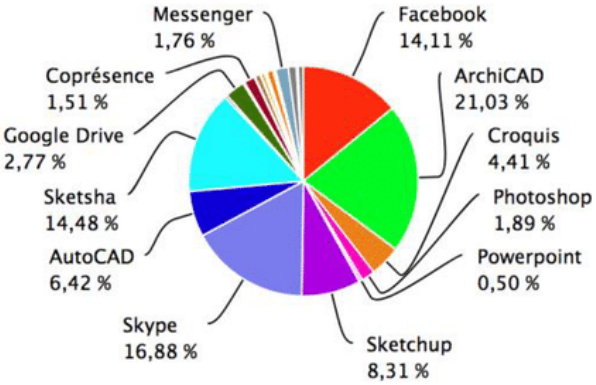
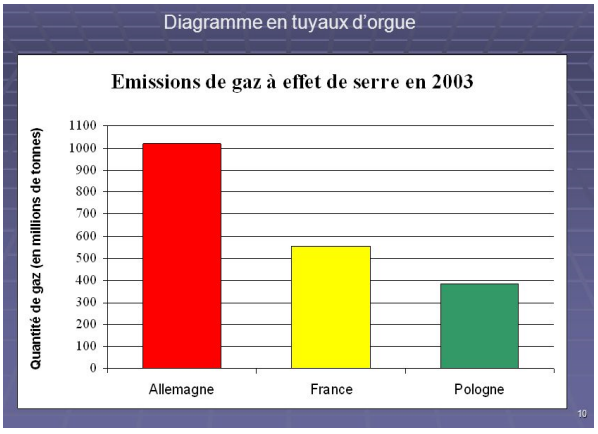


diagramme secteurs



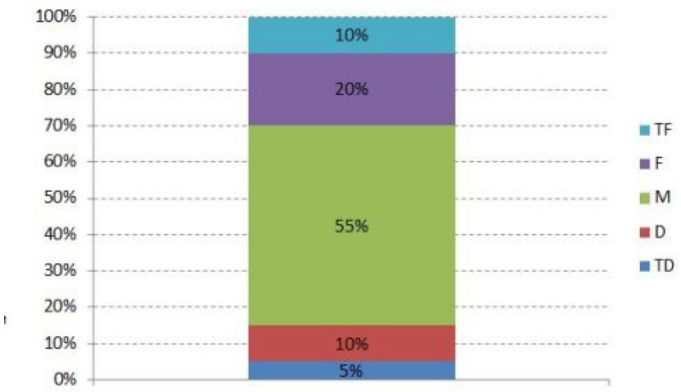
les Tuyaux d'orgue:

- Nous portons en abscisses les modalités, de façon arbitraire.
- Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité



.

Le diagramme à barre(Diagramme en barres empilées ou **Diagramme en blocs**)



3.3.2 Cas d'une d'une variable quantitative:

Cas d'une variable quantitative: peut se représenter à l'aide de trois types de diagrammes:

1. Diagramme en bâton si **la variable est discrète**

2. la courbe en escalier (**variable discrète)**

3. l'histogramme des densités si la distribution est **continue**

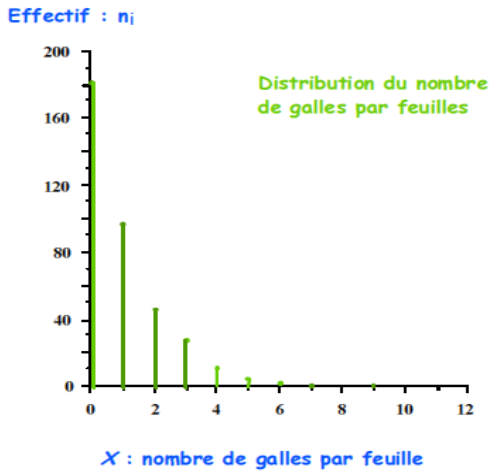
4. Polygonne de fréquence si **la variable est continue**

5. la courbe des fréquences cumulées (ou des effectifs cumulés)..

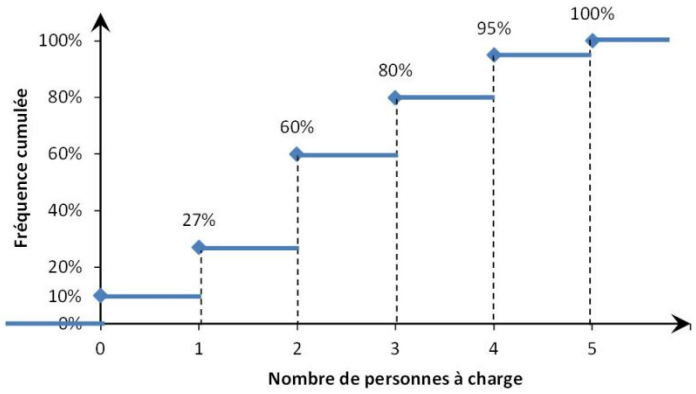
3.3.2.1 Cas d'une d'une variable quantitative discrète:.

Diagramme en bâtons:
des effectifs ou des fréquences: La différence avec le cas qualitatif consiste en ce que les

abscisses ici sont les valeurs de la variable statistique.**(voir exemple2)**

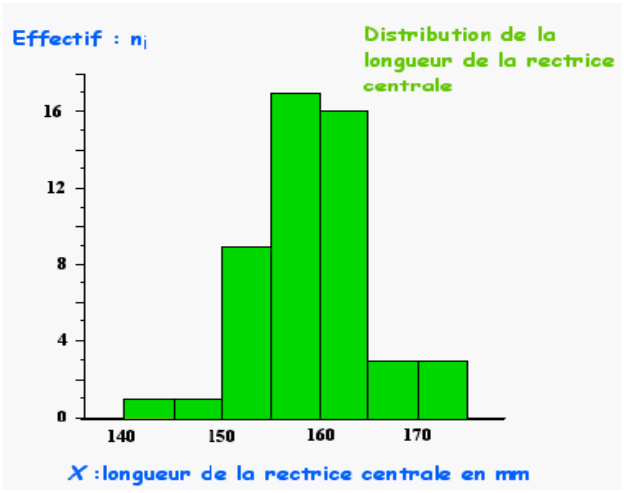


la courbe en escalier:

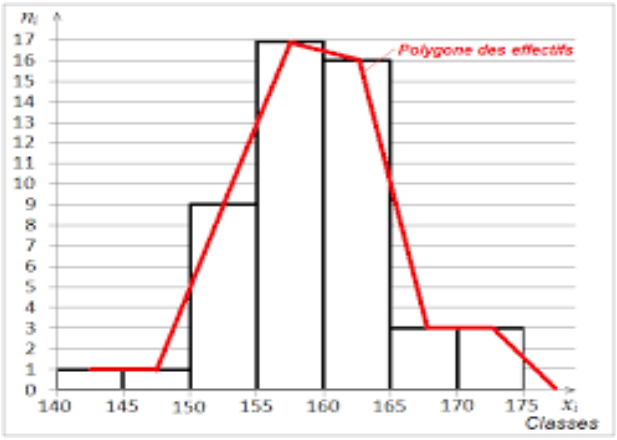


La courbe en escalier

3.3.2.2 Cas d'une d'une variable quantitative continue:
l'histogramme des densités si la distribution est **continue (exemple)**



Polygonne des effectifs: d'une **variable continue:** On obtient le polygone des effectifs (ou des fréquences) en reliant les milieux des bases supérieures des rectangles.



le rouge:Polygonne des effectifs

la courbe des effectifs cumulées croissant et décroissant sont présentés

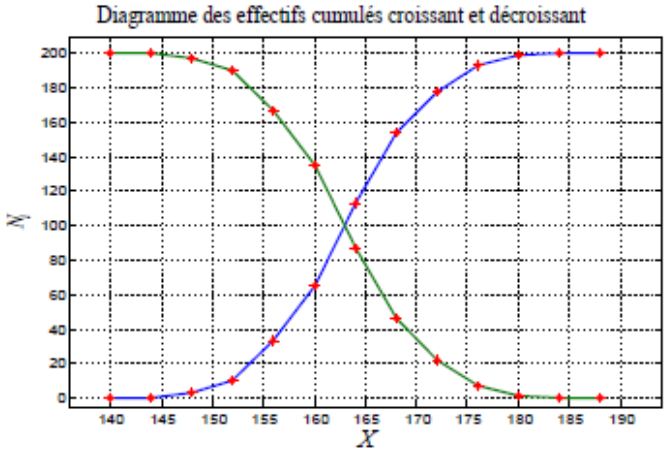


Diagramme des effectifs cumulés croissant et décroissant
200
180
160
140
120
100
80
60
40
20
0
N
140
145
150
155
160
165
170
175
180
185
190
X