Université Frères Mentouri Constantine

Biostatistique

3.3

L 2.Sc.Biologiques

Mr. LATELI Ahcene

2019-2020

Légende

 Entrée du glossaire

 Abréviation

 Référence générale

Table des matières

Partie I : Statistiques descriptives

I

A. Définitions et vocabulaires

Définition

1. **la statistique** est une **méthode** scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.
2. **La statistique** est une méthode qui vise à la description quantitative des ensembles nombreux. (définition donnée par **Gerard-Calot**). C'est une méthode et non une théorie

Le but de la **statistique descriptive** est de structurer et de représenter l'information contenue dans les données.

Définition : La biostatistique

La **biostatistique** c'est la **statistique appliquée** à la **biologie**.

Exemple

1. Étude descriptive des poids des étudiants inscrits en première année de biologie à l'université de Constantine 1.
2. Une étude statistique sur l'ensemble des exploitations agricoles dans la région de Constantine.

Comme toute science, la statistique fait appel à un vocabulaire spécialisé :

Définition : Population

Une population statistique **P** est l'ensemble d'individus définis par une propriété commune donnée. C'est un ensemble généralement très grand

Exemple

- Si l'on veut étudier la durée de vie des ampoules électriques fabriquées par

une compagnie, la population considérée est l'ensemble de toutes les ampoules fabriquées par cette compagnie.

- ensemble de personnes sur lesquelles on mesure la glycémie.
- ensemble de pays pour lesquels on dispose de données géographiques ou économiques, ...

Définition : Échantillon

Un échantillon E est une partie de la population **P**.

Exemple

Pour établir la durée de vie des ampoules électriques produites par une machine, on peut prélever au hasard un certain nombre d'ampoules - un échantillon - parmi toutes les celles produites par cette machine.

Définition : Individu ou unité statistique

Chaque élément de la population ou de l'échantillon appelé **individu** ou **unité statistique**.

Exemple

Dans l'exemple précédant, chaque ampoule constitue un individu ou une unité statistique.

Définition : La taille

Représente le nombre d'individus d'un échantillon ou d'une population. Elle est symbolisée par «n» dans le cas d'un échantillon et par «N» dans le cas d'une population.

Définition : Le caractère

C'est l'aspect particulier que l'on désire étudier.

Exemple

Concernant un groupe de personnes, on peut s'intéresser à leur age, leur sexe leur taille...

Définition : Les modalités

Les différentes manières d'être que peut présenter un caractère.

Exemple

1. Le sexe est un caractère qui présente deux modalités : féminin ou masculin.
2. Quant au nombre d'enfants par famille, les modalités de ce caractère peuvent être 0,1,2,3...,20.

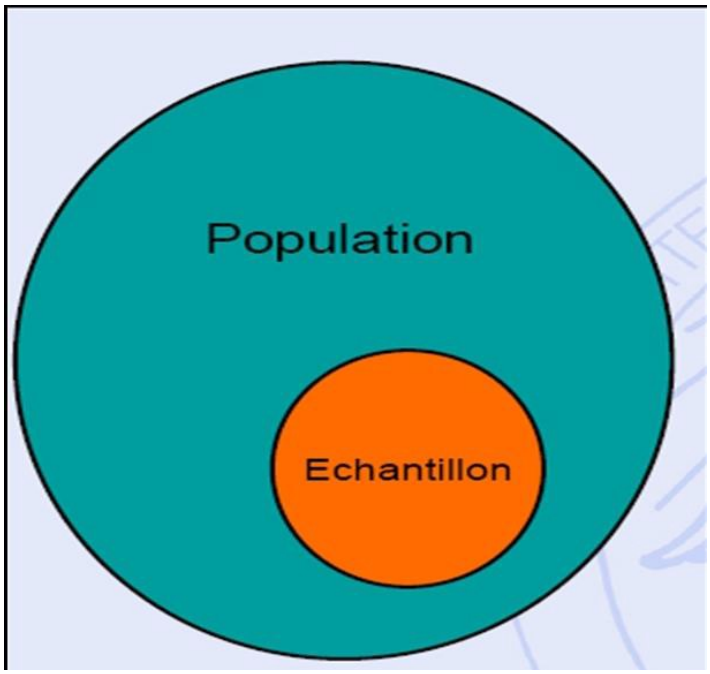


Figure 1 : Un échantillon E est une partie de la population P

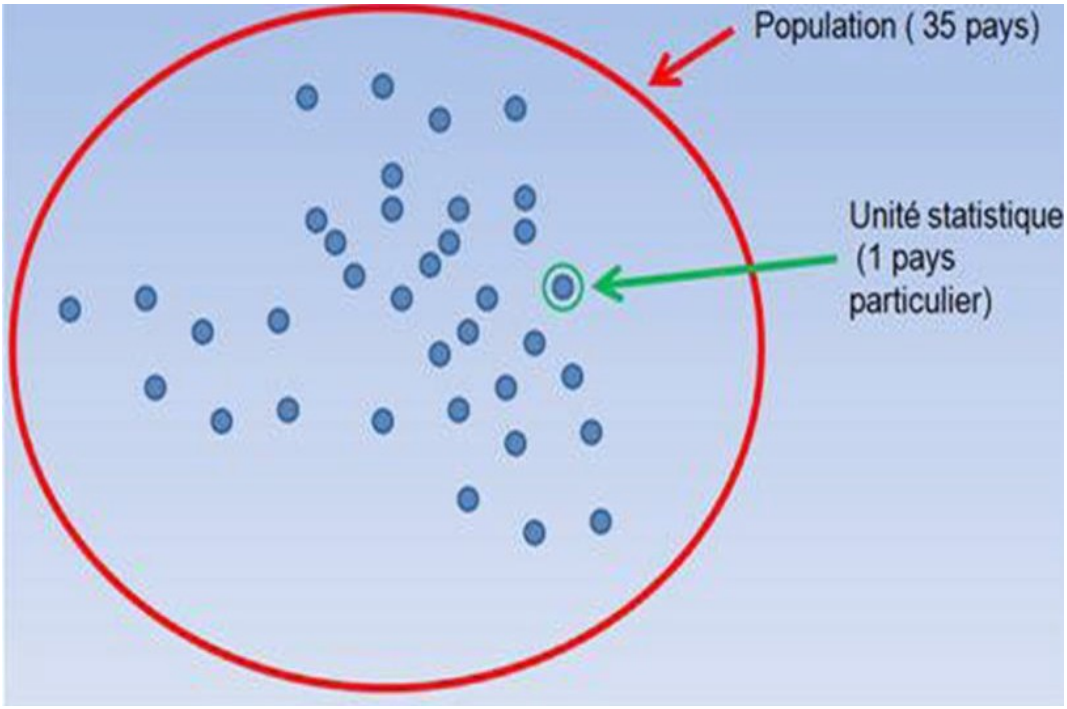


Figure 2 : Ex: dans une population P de 35 pays, un pays particulier est l'unité statistique = l'individu ou l'élément

Définition : Caractère qualitatif

Ses modalités ne s'expriment pas par un nombre

Exemple

la religion, le sexe, l'opinion...

Définition : Caractère quantitatif

Ses modalités sont numériques.

Définition : La variable statistique

chaque attribut (ou caractère ou caractéristique) a des modalités, ou peut s'exprimer selon une mesure, celles-ci varient d'un individu à l'autre ou

d'un groupe d'individus à un autre groupe d'individus. La variable statistique est le nom que l'on donne à ces caractères (attributs, caractéristiques).

Explicitation de variable en biologie

Caractéristique mesurable ou observable sur un élément (variable propre) ou dans son environnement (variable associée).

Définition

On distingue **quatre types de variables** selon leurs valeurs possibles :

- **Les variables qualitatives nominales** sont celles dont les valeurs sont des attributs sans ordre naturel (ex. sexe, csp).
- **Les variables qualitatives ordinales** sont celles dont les valeurs sont des attributs avec un ordre naturel (ex. activité faible, moyenne, forte).
- **Les variables quantitatives discrètes** sont celles dont les valeurs sont le résultat d'un dénombrement (ex. nombre d'enfants dans une famille, pouls, nombre d'oiseaux nichant dans une falaise).
- **Les variables quantitatives continues** sont celles dont les valeurs sont des mesures (ex. taille, poids).

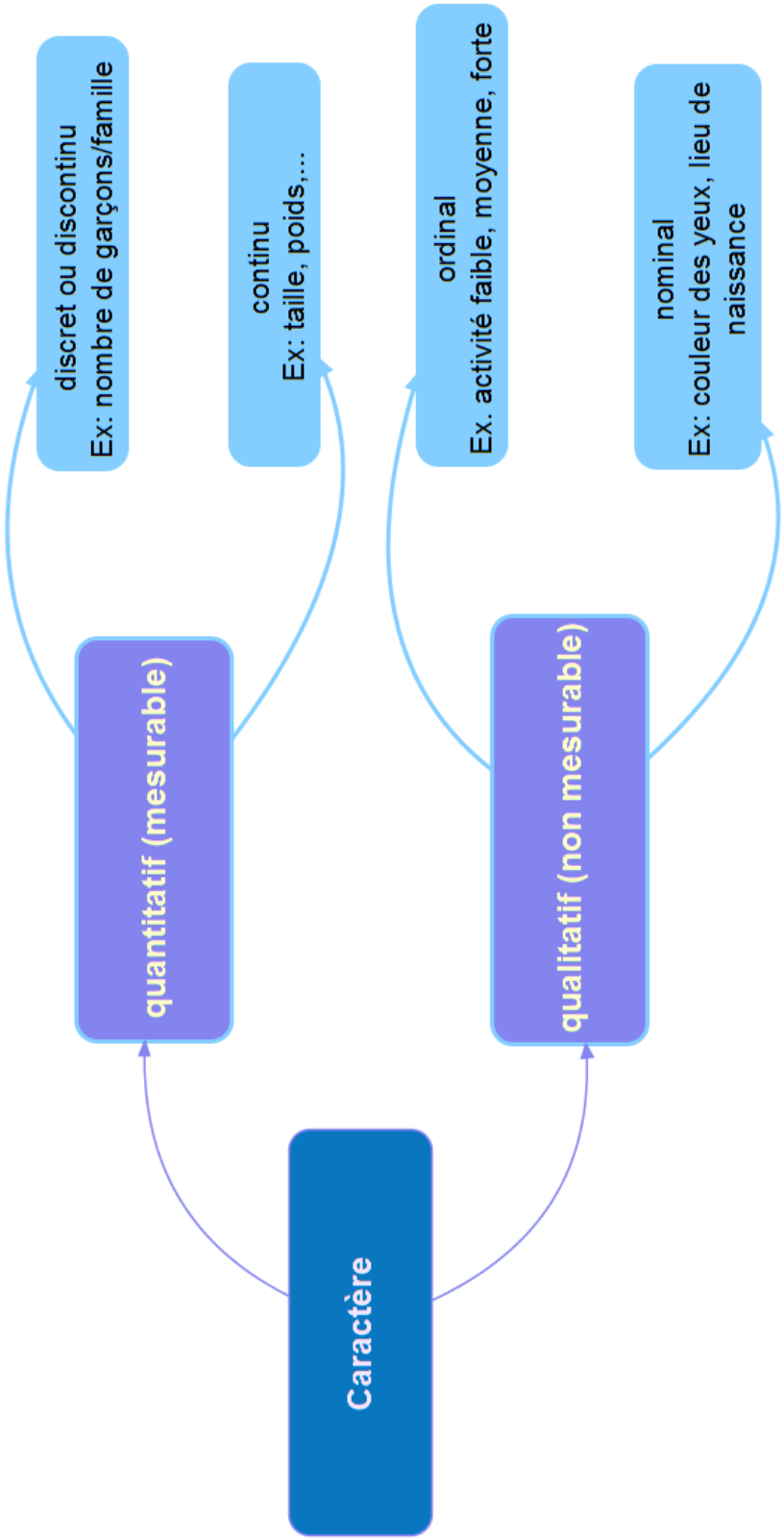


Figure 3 : Typologie des caractères pour une approche statistique

B. Statistique descriptive univariée (Statistique descriptive à un caractère)

1. Présentation d'une DS à 1 caractère

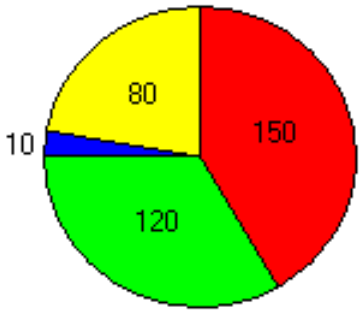
Deux types de présentations: **en tableau** et **représentation graphique**

- Une **présentation en tableau** est une présentation des effectifs et ou/ fréquences des individus en fonction des modalités(valeurs ou intervalles de classes du caractère étudié).
- La **représentation graphique** dépend du type de caractère et d'effectifs (cumulés ou non).

Situation familiale A_i	Nombre de personnes dans cette situation n_i	Fréquences $f_i = n_i/n$	f_i en % $100\ f_i$
célibataire	150	150/360	
marié	120	.	
divorcé	10	.	
veuf	80	.	
Total	360 = n	1	100%

Présentation en tableau d'un caractère qualitatif

Image 1 Figure 4 : Ex: Situation familiale dans un groupe de 360 individus



Représentation graphique du caractère « Situation familiale »

Image 2 Figure 5 :Représentation en secteurs ou en camemberts

GROUPE	ni
A	35
B	9
O	40
AB	16

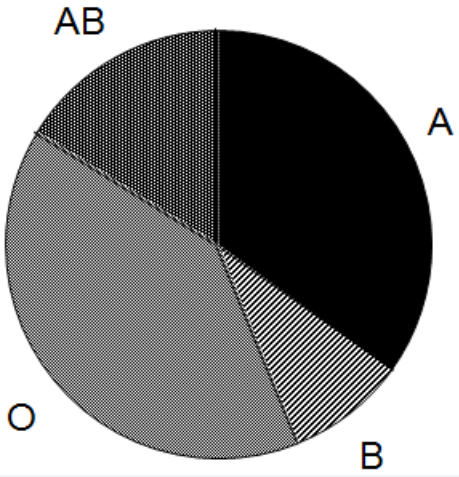


Figure 6 : groupage de 100 personnes

2. Paramètres d'une distribution

a) Les paramètres de position

Définition : Moyenne arithmétique (mean)

$E(x)$ = moyenne de la distribution théorique des éléments x.

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$ désigne **la moyenne arithmétique d'une population** finie comportant N éléments (N = effectif). Mêmes unités physiques que x.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{K} n_i \cdot x_i = \sum_{i=1}^{K} f_i \cdot x_i$ désigne **la moyenne arithmétique** de n éléments (n = effectif) tirés d'une population finie ou infinie. Mêmes unités que x, où n_i et f_i désigne respectivement l'effectif et la fréquence de la modalité x_i (ou le centre de la classe dans le cas d'une **variable classée**).

Exemple

- Soit la série de chiffres $\{8, 5, 9, 13, 25\}$. La moyenne arithmétique de cette série de chiffres se calcule ainsi :

$$\bar{x} = \frac{8 + 5 + 9 + 13 + 25}{5} = \frac{60}{5} = 12.$$

- Calcul de la moyenne arithmétique quand les valeurs sont groupées par classes par exemple :

Calculons la taille moyenne des 20 individus :

La taille moyenne est donc : $\bar{x} = \frac{3410}{20} = 170.5cm$

Remarque

- Soient plusieurs populations d'effectifs $n_1, n_2, ..., n_k$, de moyennes respectives $\bar{x_1}, \bar{x_2}, ..., \bar{x_k}$. La moyenne globale est la moyenne des moyennes i.e. $\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{K} n_i \cdot \bar{x}_i$.
- **La moyenne arithmétique conserve les changements d'échelle et d'origine** i.e. pour $a, b \in \mathbb{R}; \overline{ax+b} = a\bar{x} + b$.
- La somme algébrique des écarts de tous les termes de la série à la moyenne est nulle: i.e. $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$.
- La somme des carrés des différences de tous les termes d'une série statistique à un nombre quelconque est **minimal** lorsque ce nombre es la moyenne arithmétique de cette série: i.e. $\sum_{i=1}^{K} n_i \cdot (x_i - x)^2$ est minimale si et seulement si $x = \bar{x}$.

Définition : Médiane (median)

Symbole: Me_x; Mêmes unités physiques que x.

La médiane est la valeur de la variable qui se situe au centre de la série statistique, classée en ordre croissant. La médiane sépare la série en deux groupes d'égale importance.

- Si le nombre d'observations n est **impair** alors la médiane est la valeur de la série ordonnée située à la position $\frac{n+1}{2}$ .i.e. $Me = x_{(n+1)/2}$.
- Si n est **pair** : Me est la **moyenne arithmétique** des deux observations centrales.i.e. $Me = \frac{x_{n/2} + x_{(n/2)+1}}{2}$.

Remarque

Si **la variable est continue**, pour **calculer la médiane**, on repère d'abord la classe dans laquelle elle se trouve : c'est celle dont l'effectif cumulé est immédiatement supérieur à $\frac{n}{2}$, notons la $]x_1; x_2]$;

notons n_2 l'effectif cumulé de cette classe et n_1 l'effectif cumulé de la classe qui précède . En faisant l'hypothèse que les valeurs sont **uniformément réparties** à l'intérieur des classes, on a d'après le **théorème de Thalès** l'expression suivante :

$Me \approx x_1 + (x_2 - x_1) \cdot \frac{\frac{n}{2} - n_1}{n_2 - n_1}$.

Ce calcul fournit **une valeur approchée de la médiane**.

Exemple

1. **pour la série: [1, 32, 128, 129, 1000235], Me = 128.**
2. **pour la série [1, 32, 128, 129, 532, 1000235], Me = 128,5.**

Définition : Mode (mode)

Symbole: Mo_x Mêmes unités physiques que x.

Le mode est la valeur de la variable statistique la **plus fréquente**.

Dans le cas d'une variable statistique continue, on parle plutôt de **classe modale**.

Remarque

1. **Le mode** peut être calculé pour tous les types de variable, **quantitative et qualitative**.
2. Pour les **variables qualitatives**, le mode correspond à la **classe** ayant **la plus forte fréquence**.
3. **Le mode ou la classe modale n'est pas obligatoirement unique**.

b) Les paramètres de dispersion

Les paramètres de dispersion renseignent sur l'étalement de la distribution de fréquence autour de la moyenne.

Définition : Étendue de variation (range)

Synonyme: **plage** de variation; Mêmes unités physiques que x.

L'étendue notée e, est **la différence** entre **la plus grande** (maximum)et **la plus petite** (minimum) des valeurs observées.i.e. $e = x_{Max} - x_{Min}$.

Définition : Variance (variance)

C'est la caractéristique de dispersion la plus utilisée avec l'écart quadratique moyen.

Symboles: $s_x{}^2$ pour un échantillon, et σ^2("$sigma^2$")ou $Var(x)$ pour une population ou distribution théorique.

Pour une population statistique d'effectif N dont la moyenne vraie μ est connue par théorie ou par hypothèse, on utilise la formule suivante:

1er cas : série non classée $Var(x) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\mu})^2$.

2ème cas : série classée $Var(x) = \sigma^2 = \frac{1}{N}\sum_{i=1}^{K} n_i \cdot (x_i - \bar{\mu})^2 = \sum_{i=1}^{K} f_i \cdot (x_i - \bar{\mu})^2$.

Dans **le cas d'une variable statistique continue**,x_i représente **le centre de la** i **ème classe.**

Méthode : Autre expression de la variance : Théorème de KOENIG

1er cas : série non classée $Var(x) = (\frac{1}{N}\sum_{i=1}^{N} x_i^2) - \bar{\mu}^2$.

2ème cas : série classée $Var(x) = (\frac{1}{N}\sum_{i=1}^{K} n_i \cdot x_i^2) - \bar{\mu}^2 = (\sum_{i=1}^{K} f_i \cdot x_i^2) - \bar{\mu}^2$.

Définition : L'écart type (standard deviation)

l'écart type est la racine carrée de la variance.

Symboles: σ pour une population ou une distribution théorique s_x pour un échantillon, $s_x = \sigma(x) = \sqrt{Var(x)}$.

Remarque

1. **La variance (ou écart-type)** est toujours **positive ou nulle**.
2. **L'écart-type** a l'avantage d'être un **nombre de même dimension** que les données (contrairement à la variance qui en est le carré).
3. **l'écart type** (tout comme **la variance**) est sensible aux **valeurs extrêmes**.

Définition : Coefficient de variation (coefficient of variation)

Le coefficient de variation C_v d'un caractère statistique X est égal à l'écart type divisé par la moyenne.i.e. $C_v(X) = \frac{\sigma(X)}{\bar{X}} \cdot 100$.

Définition : L'écart moyen absolu

L'écart moyen absolu est la somme des valeurs absolues des écarts à la moyenne divisée par le nombre d'observations : $e_{moy} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|$.

Définition : L'écart médian absolu

L'écart médian absolu est la somme des valeurs absolues des écarts à la médiane divisée par le nombre d'observations : $e_{med} = \frac{1}{N}\sum_{i=1}^{N}|x_i - Me|$

Définition : L'écart interquartile

- **Le premier quartile** Q_1 de la série est la valeur x_i dont l'indice i est le plus petit entier supérieur ou égal à $\frac{n}{4}$.
- **Le troisième quartile** Q_3 de la série est la valeur x_j dont l'indice j est le plus petit entier supérieur ou égal à $\frac{3n}{4}$.
- La médiane est parfois appelée **second quartile**, c'est-à-dire $Q_2 = Me$.
- L'intervalle $]Q_1; Q_3[$ est appelé **intervalle interquartile**
- Le nombre $Q_3 - Q_1$ est appelé **écart interquartile.**

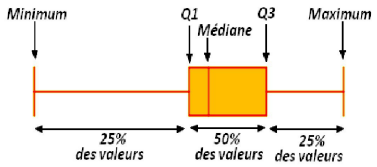


Image 3 Diagramme en boîte

Un **diagramme en boîte**, ou pour utiliser le terme anglais, un **boxplot**, est une **alternative à l'histogramme** pour **résumer graphiquement** la distribution en terme de paramètres de position et de dispersion

Remarque

Dans l'intervalle interquartile on trouve **50%** de la population.

Les moments d'une distribution (moments)

- On appelle **moment à l'origine d'ordre** $r \in \mathbb{N}$ le paramètre: $m'_r = \frac{1}{N}\sum_{i=1}^{N} x_i^r$.

- On appelle **moment centré d'ordre** $r \in \mathbb{N}$ le paramètre :
  $m_r = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^r$.