Université Mohamed Seddik Ben Yahia - Jijel

Introduction aux probabilités et à la statistique descriptive

Niveau : Première année Licence Mathématiques et Informatique

OBJECTIF DE L'ENSEIGNEMENT

Cet enseignement a pour objectif d'introduire l'analyse statistique descriptive et le calcul des probabilités. Il s'agit d'apprendre aux étudiants en première année licence mathématiques et Informatique de décrire et de présenter une série statistique à une seule dimension sous forme de tableaux et de graphiques, de l'analyser en calculant les paramètres de tendance centrale et les paramètres de dispersion. Puis, de leur apprendre à réaliser une analyse combinatoire et les calculs des probabilités.

CHAPITRE 01 : DESCRIPTION D'UNE SÉRIE STATISTIQUE À UNE SEULE DIMENSION

INTRODUCTION

La statistique est une science indispensable dans tous les domaines (médecine, économie, finance, biologie, sciences sociales …). C'est un ensemble de méthodes mathématiques et de techniques qui visent à collecter les données, les traiter, les présenter et les organiser puis à les interpréter afin de les rendre compréhensibles et aide ainsi à prendre des décisions rationnelles.

La statistique est un domaine des mathématiques appliquées. Ses principales branches sont :

- La **statistique descriptive** qui propose de présenter les données sous formes des tableaux statistiques et des résumés graphiques pour une analyse préliminaire (tableaux de bord …), des synthèses via divers indicateurs tels que les paramètres de tendance centrale, de dispersion et de formes.
- L'**analyse factorielle** qui a pour but de réduire un nombre important de variables pour décrire un phénomène en produisant des indicateurs statistiques de synthèse, en formant des groupes homogènes de variables statistiques et de groupes homogènes d'individus statistiques, telles que l'Analyse des composantes principales ACP, Analyse des correspondances multiples ACM, …
- La **classification** qui offre des typologies ou des groupes homogènes dans une population disparate.
- La **modélisation** qui vise à expliquer un phénomène donné en fonction des variables explicatives, en permettant également une prédiction de valeurs sur des individus statistiques nouveaux ou inconnus.

…

SECTION 01 : VOCABULAIRE STATISTIQUE ET DÉFINITIONS

I. VOCABULAIRE STATISTIQUE

1. Population statistique

La population statistique (notée Ω) est l'ensemble sur lequel on effectue des observations. On appelle donc la population statistique, l'ensemble d'unités statistiques constituant les unités observées. Elle est bien spécifiée s'il n'y a pas d'ambiguïté sur la définition de l'ensemble.

On peut distinguer deux types de populations statistiques.

- ***Population statistique finie :*** La population statistique est dite finie lorsque le nombre d'unités statistiques observées est limité par exemple : le nombre d'étudiants dans une section, nombre d'enfants dans un ménage …
- ***Population statistique infinie*** : La population statistique est dite infinie lorsque le nombre d'unités statistiques observées est très élevé ou illimité. Par exemple : Nombre d'étoiles dans le ciel, nombre de graines de blé collectées, nombre de pièces de monnaie en libre circulation …

2. Unité statistique (ou individu statistique)

On appelle unité statistique (ou individu statistique), tout élément de la population statistique étudiée. On la note ω.

3. Échantillon statistique

C'est un sous ensemble de la population statistique considérée. Le nombre d'individus ou d'unités dans l'échantillon est **la taille** de l'échantillon.

En statistique, « prélever » un échantillon consiste à extraire un ou plusieurs individus d'une population. Les renseignements obtenus sur un échantillon permettent de mieux connaître la population statistique. Le recours à un échantillon répond en général à la nécessité pratique (temps limité, coût élevé, …). Un échantillon doit être représentatif.

4. Caractère statistique (variable statistique)

Le caractère statistique est la propriété singulière que l'on se propose d'observer dans la population statistique. Un caractère étudié porte aussi le nom de variable statistique. Ce caractère doit caractériser l'intégralité des individus appartenant à la population statistique étudiée.

Il existe deux types de caractères, à savoir :

Caractère quantitatif : Lorsque la variable statistique X peut être exprimée numériquement, donc lorsqu'on peut la mesurer, elle est dite quantitative (ou mesurable). Les valeurs d'une variable quantitative sont des nombres exprimant une quantité sur lesquels les opérations arithmétiques sont possibles et ont un sens (somme, division...). La variable statistique peut être discrète (discontinue) ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre :

- Elle est discrète si elle ne prend que des valeurs entières, des valeurs isolées les unes des autres (exemple : nombre d'enfants d'une famille).
- Elle est dite continue lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini de R (exemple : diamètre de pièces, poids, salaires…).

Caractère qualitatif : Lorsque la variable ne se prête pas à des valeurs numériques, elle est dite qualitative (exemple : opinions politiques, nationalité, couleurs des yeux…). Elle peut être ordonnée ou nominale, dichotomique ou non. Ses valeurs sont des modalités statistiques (ou catégories) exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens.
On distingue deux types de variables qualitatives :

- Variable qualitative ordinale : l'ordre des modalités est important (Niveau scientifique : faible, moyen , fort)
- Variable qualitative nominale : l'ordre des modalités n'est pas important (Nationalité des participants à un colloque international : Algérienne, tunisienne, française, …)

Une variable est dichotomique si elle n'a que deux modalités.

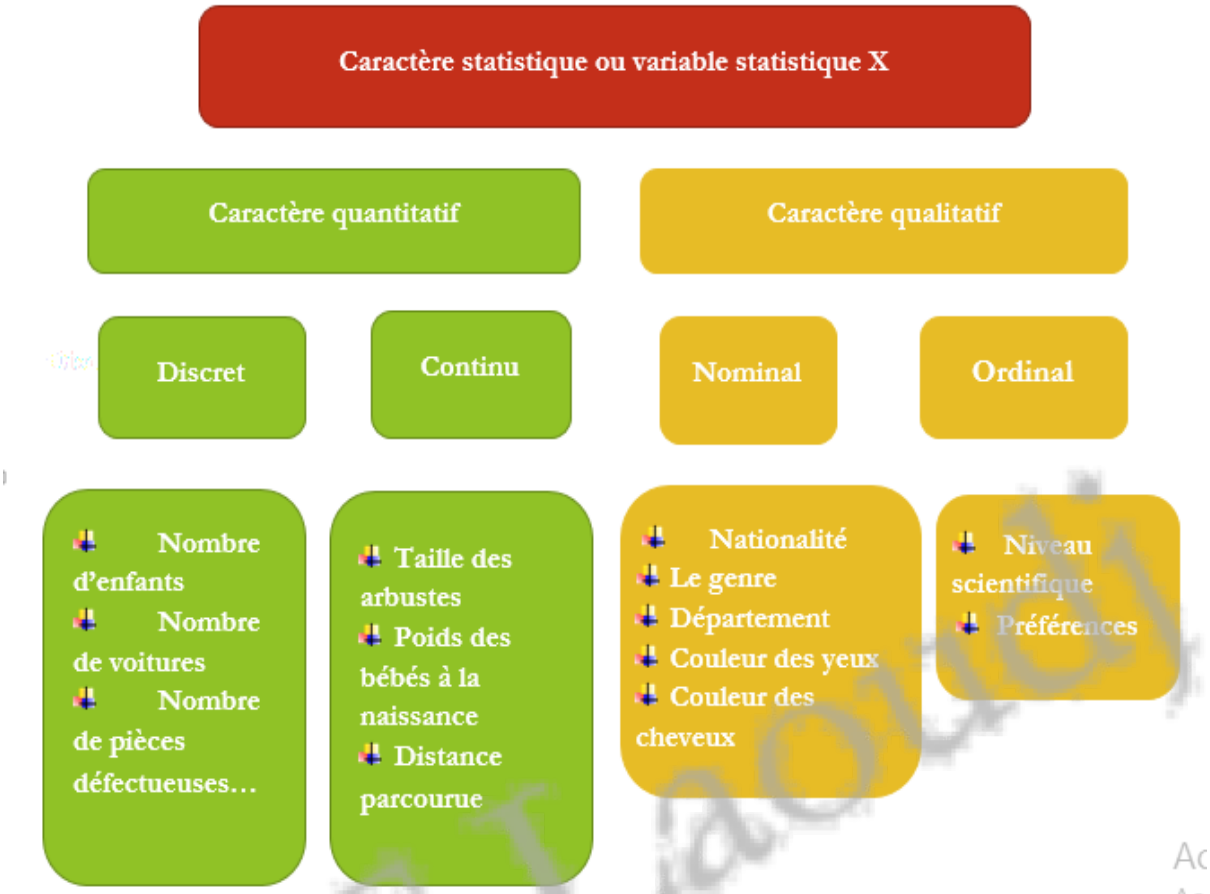
5. Les modalités statistiques

Chaque unité statistique occupe une position ou une place dans la population statistique par rapport au caractère ou variable étudiée X. Les unités se répartissent dans les modalités de la variable.

Remarques importantes :

Avant chaque étude statistique, vous devez :

- Identifier avec précision la population statistique et l'unité statistique.
- Cibler le caractère étudié pour confirmer l'appartenance de chaque individu à la population étudiée.
- Chaque individu doit appartenir à une modalité du caractère et une seule seulement, pour que la somme des individus statistiques des différentes modalités soit égale à la somme des individus statistiques de la population statistique étudiée.
- En revanche, une modalité peut contenir plusieurs unités.



Quelques exemples

Exemple 1 :

On sélectionne un échantillon de 20 habitations et on s'intéresse à leur répartition selon leur nombre de pièces :

4 1 3 1 4 5 5 5 6 6 2 1 4 4 4 5 2 1 1 2

Population statistique = 20 habitations

Caractère statistique = nombre de pièces

Type du caractère statistique (ou variable) = quantitatif .discret (ou discontinu)

Modalités statistiques = 1, 2, 3, 4, 5, 6.

Exemple 2 :

On s'intéresse au poids des bébés à la naissance. La répartition de 10 bébés nés à la maternité de Jijel en Juin 2022 selon leur poids en **kg** est la suivante :

1,85 1,96 3,33 1,99 4,05 4,60 3,20 3,33 1,90 2,75

Population statistique = 10 bébés nés à la maternité de Jijel en Juin 2022

Caractère statistique = poids des bébés en kg

Type du caractère = Quantitatif continu

Modalités statistiques = Formation des classes

[1,50 kg – 2,50 kg [[2,50 kg– 3,50 kg[[3,50 kg – 5 kg [

Exemple 03 :

On réalise une étude sur le niveau des étudiants de première année universitaire en informatique (niveau faible, moyen, fort).

Population statistique = Étudiants en première année universitaire.

Caractère statistique = Niveau scientifique en Informatique.

Type du caractère = Qualitatif ordinal.

Modalités statistiques = faible, moyen, fort.

II. DÉFINITIONS

1. Variable statistique

Une variable statistique (ou caractère statistique) est ce qui est observé ou mesuré sur les individus d'une population statistique. On peut la définir mathématiquement comme une application X de Ω (population statistique) dans $\mathbb{R}$ (si X est continue) et de Ω dans $\mathbb{N}$ (si X est discrète).

X : $\Omega \to \mathbb{R}\ (ou\ \mathbb{N})$

$\omega \to X(\omega)$

$X(\omega)$ Représente les modalités de la variable statistique X étudiée, qui caractérisent les individus statistiques ω .

Les valeurs des k modalités sont : $x_1, x_2, \dots x_k$

- Si la variable statistique est quantitative continue, les valeurs de $X(\omega) \in \mathbb{R}$.
- Si la variable statistique est quantitative discrète, les valeurs de $X(\omega) \in \mathbb{N}$.
- Si la variable statistique est qualitative, on peut coder $X(\omega)$ en 0, 1, 2 … mais les opérations mathématiques sur ces valeurs n'ont aucun sens.

Exemple : On sélectionne un échantillon de 15 habitations (logements) et on compte le nombre de pièces dans chaque habitation, on trouve les résultats suivants :

4 1 3 1 4 5 5 5 6 6 2 1 4 4 4

Population statistique = 20 habitations.

$$\Omega = \{ w_1, w_2, \dots, w_{15}\}$$

La première habitation est composée de quatre pièces, on note : $X(w_1) = 4$ pièces

La deuxième habitation est composée d'une seule pièce, on note : $X(w_2) = 1$ pièce.

…

2. Effectif (ou fréquence absolue)

On appelle **effectif** d'une modalité statistique ou d'une classe $[e_i\ ,\ e_{i+1}]$ (si X est continue),le nombre de fois où les valeurs x de cette modalité (ou de cette classe) apparait dans la population statistique étudiée. L'effectif est noté "n_i". L'effectif est parfois appelé fréquence absolue.

On appelle **l'effectif total** de la population, noté N, la somme des k effectifs (k modalités) particuliers n_i correspondant à chacune des valeurs, soit :

$$n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i = N$$

Par exemple : [20 ans, 22 ans [et l'effectif correspondant est égal à 15, cela signifie que 15 étudiants sont âgés entre 20 et 21 ans révolus (22 n'appartient pas à la classe).

3. Fréquence (ou fréquence relative)

On appelle **fréquence relative** d'une modalité statistique ou de la classe $[e_i, e_{i+1}[$, le rapport de l'effectif n_i associé à cette modalité (ou classe) et de l'effectif total N de la population observée.

$$f_i = \frac{n_i}{N} \text{ pour tout } i = 1, \dots, K$$

On peut dire aussi que la fréquence f_i qui correspond à la classe $[e_i, e_{i+1}[$ est égale à la proportion d'unités statistiques qui se retrouvent dans cette classe.

Ce rapport est égal au pourcentage d'individus présentant classe $[e_i, e_{i+1}[$, par raport à l'ensemble de la population observée f_i est toujours comprise entre 0 et 1.

On peut l'exprimer en pourcentage %, noté p_i est toujours comprise entre 0 et 100%.

Pour une série statistique à **k** modalités statistiques, on a :

$$\sum_{i=1}^{K} f_i = f_1 + f_2 + \cdots + f_k = 1$$

4. Effectifs cumulés et fréquences cumulées

Effectifs cumulés

Dans une série statistique où les valeurs de la variable étudiée X peuvent être rangées selon un ordre croissant (ou décroissant) ; on peut également définir des effectifs cumulés. Un effectif cumulé correspond à l'effectif d'une modalité auquel on ajoute les effectifs de toutes les modalités précédentes. On cumule les n_i. Et il peut être **croissant** $\overrightarrow{N}$ comme il peut être **décroissant** $\overleftarrow{N}$**.**

Fréquences cumulées

Dans une série statistique où les valeurs de la variable étudiée X peuvent être rangées selon un ordre croissant on peut également définir des fréquences cumulées. Une fréquence cumulée correspond à la fréquence d'une modalité à laquelle on ajoute les fréquences de toutes les modalités précédentes.

Elle peut être croissante F comme elle peut être décroissante G**.** On cumule les fréquences f_i , en ajoutant chaque fréquence f_i à la somme de celle qui précèdent. F et G sont toujours comprises entre 0 et 1. On peut aussi les exprimer en pourcentage % et sont toujours comprises entre 0 et 100%.

la fonction F est définie comme étant la somme des fréquences f_i des valeurs inférieures à x : $F_X(x) = P(X < x)$

Et la fonction G est définie comme étant la somme des valeurs supérieures ou égales à x :

$G_X(x) = P(X \geq x) = 1 - F_X(x)$

5. Définition d'un Tableau statistique

Un tableau statistique est un outil qui permet de décrire et d'organiser les données d'une série statistique de façon claire. On en trouve les modalités du caractère statistique, l'effectif et la fréquence. Il est possible d'y ajouter les effectifs cumulés ou les fréquences cumulées.

Un tableau doit toujours avoir un titre, un numéro et une source.

Caractère discret

Exemple : Répartition de 20 logements selon leur nombre de pièces

Nombre de pièces **xi**	nombre de logements **ni**	Fréquence **fi**	Effectif cumulés croissant $\overrightarrow{N}$	Effectif cumulés décroissants $\overleftarrow{N}$	Fréquence cumulées croissantes **F**	Fréquences cumulées décroissantes **G**
1	5	0,25	5	20	0,25	1,00
2	3	0,15	8	15	0,40	0,75
3	1	0,05	9	12	0,45	0,60
4	5	0,25	14	11	0,70	0,55
5	4	0,20	18	6	0,90	0,30
6	2	0,10	20	2	1,00	0,10
Total	20	1,00				

Source : Réalisé par nos soins

- **Interprétation des fréquences cumulées croissantes F :**

$F_X(1) = P(X < 1) = 0$ soit 0%.

La proportion des logements à moins d'une pièce est 0%.

$F_X(2) = P(X < 2) = 0{,}25$ soit 25%

La proportion des logements à moins de deux pièces (à une pièce seulement) s'élève à 25%.

…

$F_X(6) = P(X < 6) = 0{,}90$ soit 90%

La proportion des logements à moins de six pièces s'élève à 90%.

$F_X(7) = P(X < 7) = 1$ soit 100%

La proportion des logements à moins de sept pièces s'élève à 100%.

Ici le nombre de pièces est inférieur strictement à 7. La modalité « 7 » ne figure pas dans le tableau.

Les effectifs croissants s'interprètent de la même façon (voir TD).

- **Interprétation des fréquences cumulées décroissantes G :**

$G_X(x) = P(X \geq x)$

$G_X(1) = P(X \geq 1) = 1$ soit 100%

La proportion des logements ayant au moins une pièce est 100%.

$G_X(2) = P(X \geq 2) = 0{,}75$ soit 75%

La proportion des logements ayant au moins deux pièces s'élève à 75%.

…

$G_X(6) = P(X \geq 6) = 0{,}10$ soit 10%

La proportion des logements ayant au moins six pièces s'élève à 10%.

$G_X(7) = P(X \geq 7) = 0$ soit 0%

La proportion des logements ayant au moins sept pièces s'élève à 0%.

Les effectifs décroissants s'interprètent de la même façon.

Remarque importante : ***Pour éviter les erreurs lors de l'interprétation et de la représentation du diagramme cumulatif de F et G, on peut décaler les valeurs de G (la dernière colonne du tableau précédent). Ceci est possible puisque*** $G_X(x) = 1- F_X(x)$

Même remarque pour les effectifs décroissants $\overleftarrow{N}$.

Nombre de pièces **xi**	nombre de logements **ni**	Fréquence **fi**	Effectif cumulés croissants $\overrightarrow{N}$ **00**	Effectif cumulés décroissants $\overleftarrow{N}$ **20**	Fréquence cumulées croissantes **F** **0,00**	Fréquences cumulées décroissantes **G** **1,00**
1	5	0,25	5	15	0,25	0,75
2	3	0,15	8	12	0,40	0,60
3	1	0,05	9	11	0,45	0,55
4	5	0,25	14	06	0,70	0,30
5	4	0,20	18	02	0,90	0,10
6	2	0,10	20	00	1,00	0,00
Total	20	1,00				

Source : Réalisé par nos soins

Remarque : Lorsque nous disposons d'un nombre important de modalités statistiques pour un caractère X discret, ou lorsque le caractère de la population est de nature continue, on regroupe les valeurs en classe $[e_i, e_{i+1}[$.

- **Caractère Continu**

Exemple : Répartition de 425 salariés, d'une entreprise Z, selon leurs salaires nets (en dollars américains) en 2010.

Montant du salaire (en dollars Américains) **xi**	nombre de salariés **ni**	Fréquence **fi**	Effectif cumulés croissants $\overrightarrow{N}$	Effectif cumulés décroissants $\overleftarrow{N}$	Fréquence cumulées croissantes **F**	Fréquences cumulées décroissantes **G**
[750 - 1000 [	150	0,35	150	425	0,35	1,00
[1000 - 1500 [	100	0,24	250	275	0,59	0,65
[1500- 2500 [	99	0,23	349	175	0,82	0,41
[2500 - 3000 [	50	0,12	399	76	0,94	0,18
[3000 - 5000 [	26	0,06	425	26	1,00	0,06
Total	425	1,00				

Source : Réalisé par nos soins

- **Interprétation des fréquences cumulées croissantes F :**

$F_X(750) = P(X < 750) = 0$ soit 0%.

La proportion des salariés qui perçoivent un salaire inférieur à 750 dollars américains est 0%.

$F_X(1000) = P(X < 1000) = 0{,}35$ soit 35%

La proportion des salariés qui perçoivent un salaire moins de 1000 dollars s'élève à 35%.

…

$F_X(3000) = P(X < 3000) = 0{,}94$ soit 94%

La proportion des salariés qui perçoivent un salaire moins de 3000 dollars s'élève à 94%.

$F_X(5000) = P(X < 5000) = 1$ soit 100%

La proportion des salariés qui perçoivent un salaire moins de 5000 dollars s'élève à 100%.

- **Interprétation des fréquences cumulées décroissantes G :**

$G_X(x) = P(X \geq x)$

$G_X(750) = P(X \geq 750) = 1$ soit 100%

La proportion des salariés qui perçoivent un salaire d'au moins 750 dollars s'élève à 100%

$G_X(1000) = P(X \geq 1000) = 0{,}65$ soit 65%

La proportion des salariés qui perçoivent un salaire d'au moins 1000 dollars s'élève à 65%

…

$G_X(3000) = P(X \geq 3000) = 0{,}06$ soit 6 %

La proportion des salariés qui perçoivent un salaire d'au moins 3000 dollars s'élève à 6%

$G_X(5000) = P(X \geq 5000) = 0$ soit 0%

La proportion des salariés qui perçoivent un salaire d'au moins 5000 dollars s'élève à 0%

Les effectifs décroissants s'interprètent de la même façon.

SECTION 02 : REPRÉSENTATION GRAPHIQUE D'UNE SÉRIE STATISTIQUE A UNE SEULE DIMENSION

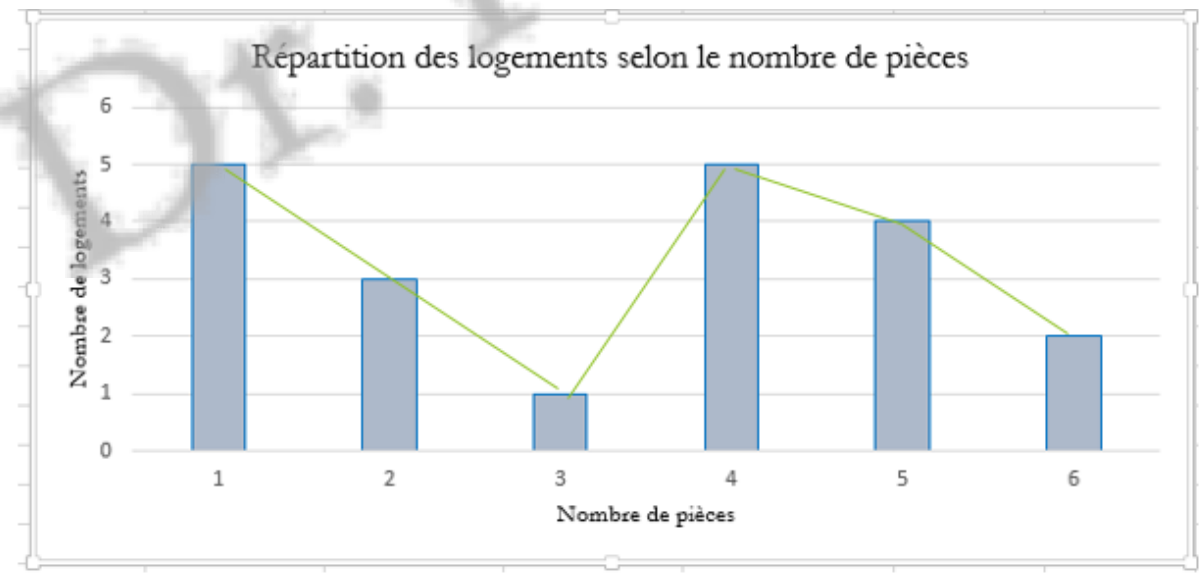
INTRODUCTION

Une série statistique associée à une variable statistique donnée X est organisée et représentée par un tableau statistique qui établit la répartition des individus statistiques selon le caractère étudié. Cependant, avec l'augmentation du nombre de modalités statistiques, sa lecture devient difficile. Il est donc recommandé **d'accompagner** ce tableau par des représentations graphiques adéquates. En effet, un graphique offre une vue d'ensemble meilleure, claire et rapide de la série statistique que le tableau statistique.

I. CAS D'UN CARACTÈRE QUANTITATIF DISCRET

1. DIAGRAMME EN BÂTONS ET POLYGONE DES EFFECTIFS

Le graphique adéquat pour représenter les effectifs **« ni »** ou les fréquences **« fi »** est « le diagramme en bâtons » pour une série statistique à un seul caractère quantitatif discret. Il est aussi appelé « diagramme en barres ». En abscisses, on trouve les valeurs du caractère X étudié à **k** modalités statistiques : x_1 , x_2 , ….x_k et en ordonnées, les effectifs ni correspondantes : n_1 , n_2 , ….n_k .



Source : Réalisé par nos soins

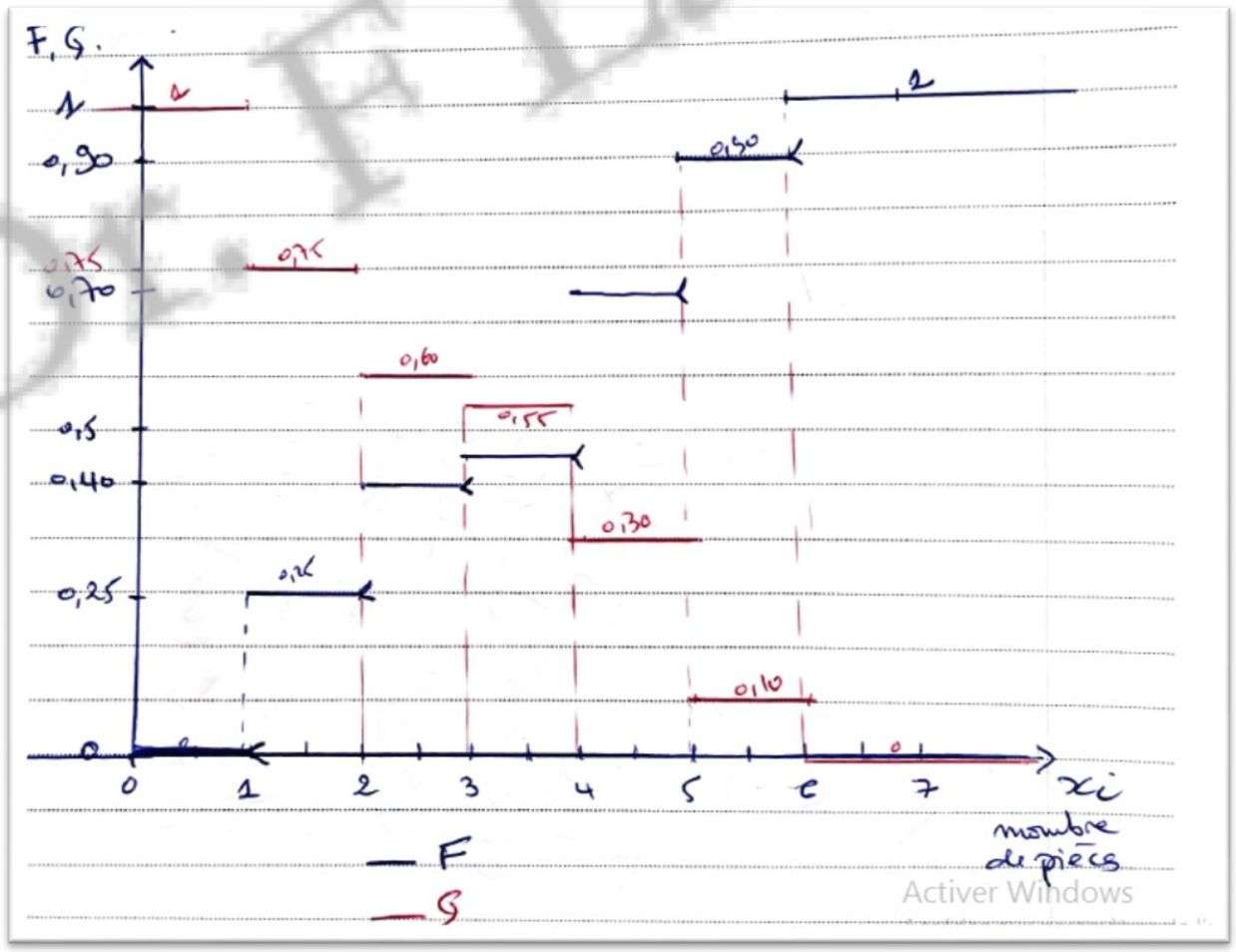
L'aire des bâtons n'a aucun sens. C'est la hauteur de chaque bâton qui renseigne l'effectif (ou la fréquence) de la modalité statistique.

Si l'on joint les sommets du diagramme en bâtons, on obtient le polygone des effectifs (ou des fréquences).

2. DIAGRAMME CUMULATIF

Les effectifs cumulés et les fréquences cumulées peuvent être représentés graphiquement. Le graphique adéquat est le « Diagramme cumulatif » appelé aussi (diagramme intégral). Sur l'axe des abscisses, on trouve les valeurs xi et sur l'axe des ordonnées, on trouve les fréquences cumulées.
Il est sous forme d'**escalier** tel qu'entre une valeur x_i et une valeur successive x_{i+1}, l'effectif est constant (ou la fréquence est constante).

Diagramme cumulatif des logements



Source : Réalisé par nos soins

II. CAS D'UN CARACTÈRE QUANTITATIF CONTINU

1. HISTOGRAMME ET POLYGONE DES EFFECTIFS

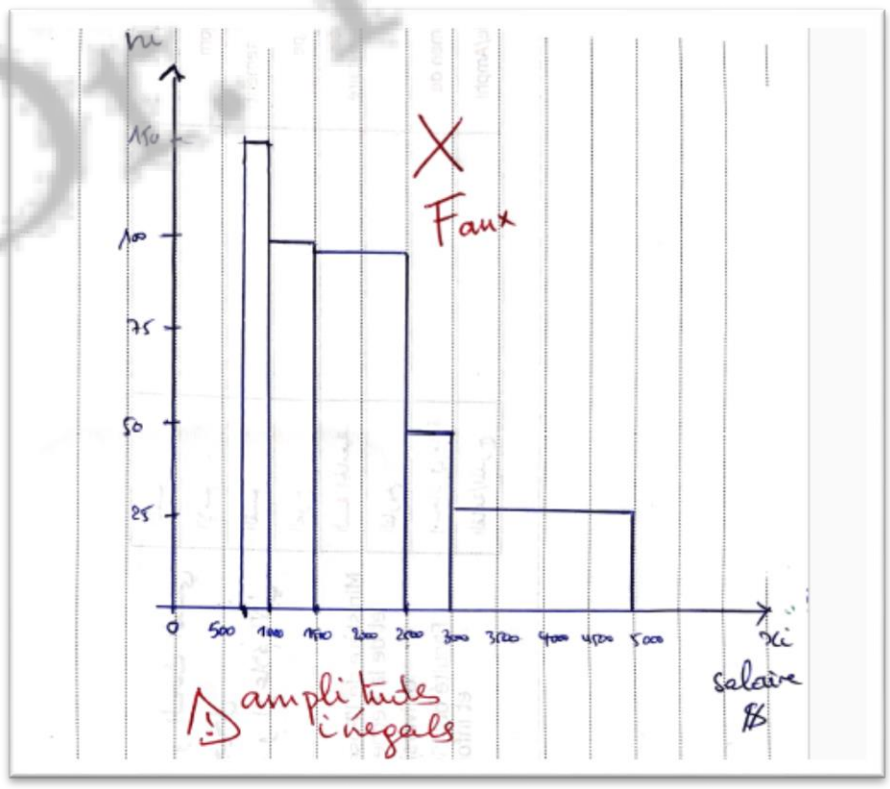
Le graphique adéquat pour représenter les effectifs (ou les fréquences) pour une série statistique associée au caractère statistique quantitatif continu est « l'HISTOGRAMME ». C'est une suite de rectangles accolés l'un à l'autre. Contrairement au diagramme en bâtons, la surface de chaque rectangle doit être proportionnelle à l'effectif de la classe.

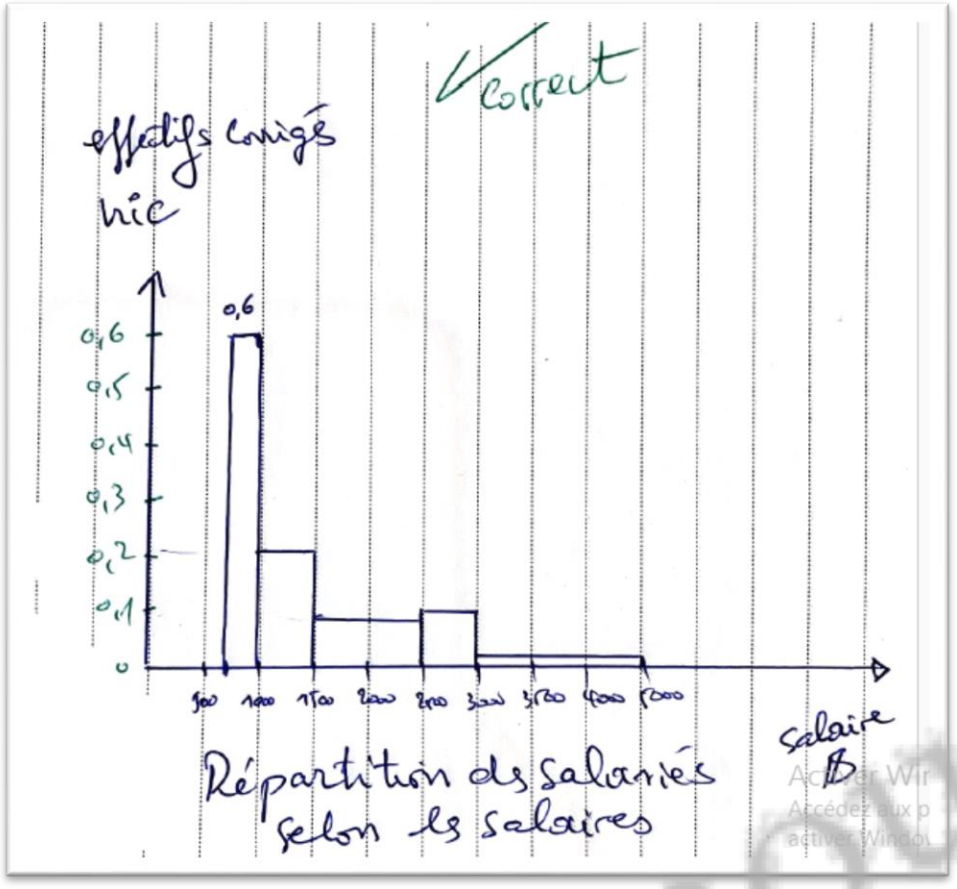
Attention :

Lorsque les amplitudes des différentes classes sont inégales, on représente sur l'histogramme les effectifs corrigés $n_i^c = \frac{ni}{ai}$

Avec ni est le nombre d'unités statistiques appartenant à la classe $[e_i, e_{i+1}[$ et ai est son amplitude : $a_i = e_{i+1} - e_i$

Répartition des 425 salariés selon leur salaire (en $).

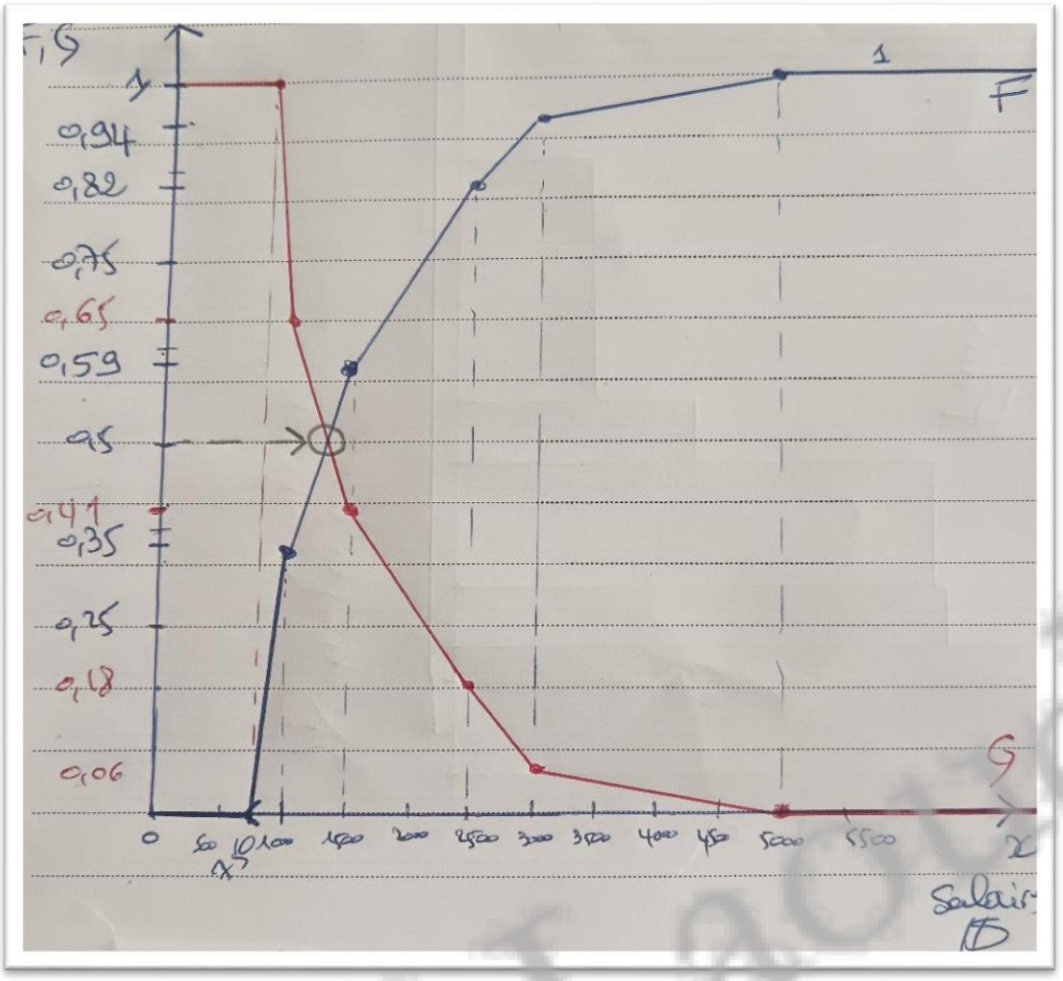




2. POLYGONE CUMULATIF

Le graphique adéquat pour représenter graphiquement les effectifs cumulés ou les fréquences cumulées pour une série statistique à un caractère quantitatif continu est le « Polygone cumulatif ». Sur l'axe des abscisses, on trouve les valeurs xi et sur l'axe des ordonnées, les fréquences cumulées.

Il n'est pas sous forme d'escalier tel qu'entre une valeur x_i et une valeur successive x_{i+1} , l'effectif (ou fréquence) est croissant pour la fonction F et décroissant pour la fonction G. On constate que lorsque la valeur **xi** coïncide avec la borne **ei** de la classe $[e_i, e_{i+1}[$; la fréquence cumulée **Fi** est une valeur exacte. Cependant, lorsque la valeur **xi** est comprise entre e_i et e_{i+1} ; la valeur de la fréquence cumulée **Fi** est une valeur approximative obtenue par une interpolation linéaire à partir des points de construction du polygone des fréquences cumulées de coordonnées (e_i , **F(ei))** et (e_{i+1} , **F(**e_{i+1}**)).**



III. CAS D'UN CARACTÈRE QUALITATIF

1. DIAGRAMME EN BÂTONS (TUYAUX D'ORGUE)

2. DIAGRAMME CIRCULAIRE (CAMEMBERT)

Appelé également digramme en secteurs.

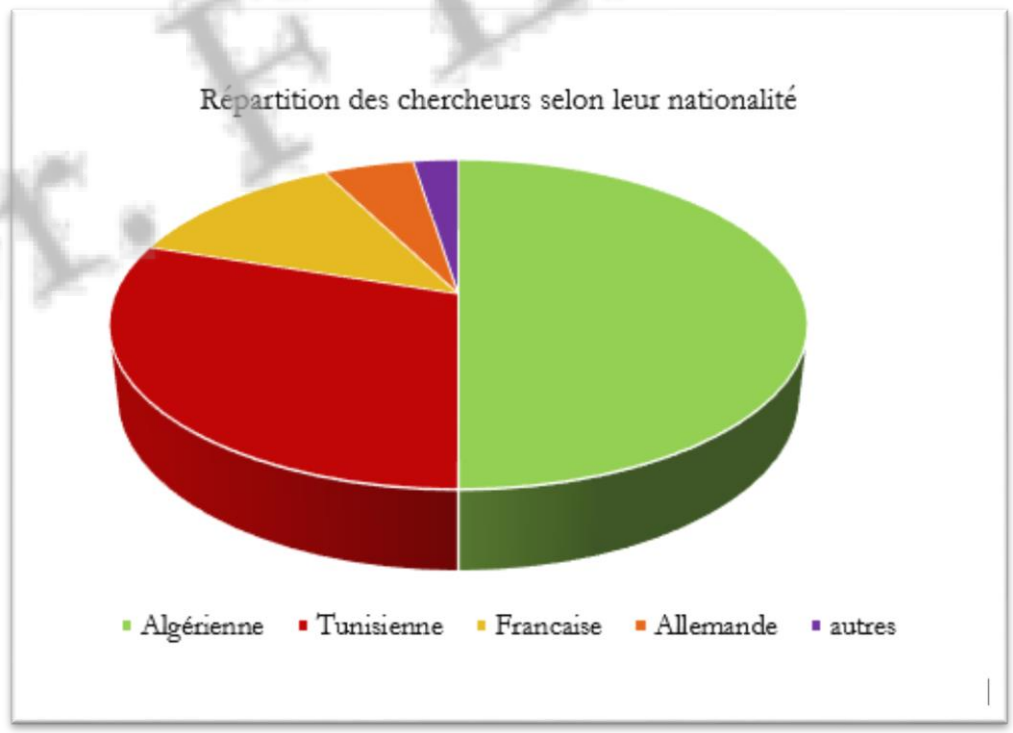
C'est un cercle (ou demi-cercle) divisé en secteurs des différentes modalités statistiques. Les mesures des angles des secteurs sont proportionnelles à l'effectif de chaque modalité statistique.

Les mesures des différents angles $\boldsymbol{\theta}$ sont obtenues par :

$\boldsymbol{\theta} = 360 * ni / N = 360 .fi$

Exemple : Répartition de 200 chercheurs participants à un colloque international selon leur nationalité.

Nationalité **xi**	nombre des chercheurs **ni**	fréquences **fi**	Pourcentage **%**	mesure de l'angle θ
Algérienne	100	0,5	50	180
Tunisienne	60	0,3	30	108
Française	25	0,125	12,5	45
Allemande	10	0,05	5	18
autres	5	0,025	2,5	9
total	200	1	100	360



On peut présenter qu'un demi-cercle et les mesures des angles θ sont obtenues par :

$\theta = 180 * ni / N = 180 .fi$

Remarque : le diagramme en bâtons peut servir à représenter une sérié statistique à caractère qualitatif.