1 SERIES STATISTIQUES A UNE VARIABLE

«Ce qui est d'une hypocrisie insupportable
c'est d'accepter les privilèges d'une classe
sans en accepter les fonctions.»
André MAUROIS

MARCHE D'APPROCHE

1. INTRODUCTION

Le mot "statistique" vient du latin "status", qui signifie "Etat". A l'origine, la statistique rassemblait exclusivement des renseignements concernant l'Etat : recensement d'une population, évaluation des ressources, état des stocks en denrées alimentaires etc..., ce qui explique le vocabulaire employé actuellement.
Les premières statistiques connues remontent à plus de 4000 ans en Chine (tables de statistiques agricoles). La Bible évoque plusieurs opérations de recensement, dont le dénombrement des Israélites aptes à porter les armes (livre des Nombres, ch.1).
De nos jours, les méthodes statistiques sont employées en médecine (évaluation de l'efficacité d'un traitement, lien entre maladie et mode de vie ...) aussi bien qu'en agronomie (sélection des variétés, études de descendance ...) ou dans l'industrie (contrôle de qualité, organisation du travail...) sans oublier la sociologie dont les enquêtes et sondages de toutes sortes fleurissent dans les médias.

2. LE VOCABULAIRE

2. 1. Population – Individus – Caractères

On étudie un (ou plusieurs) **caractère** présenté par les **individus** d'une **population.** On dispose rarement des résultats relatifs à cette population et on se restreint à l'étude d'un (ou plusieurs) **échantillon** de cette population.

■ **On souhaite étudier les conditions de vie des familles résidant dans un quartier d'une petite ville de province. Cette étude porte, par exemple, sur les caractères suivants :**

- ♦ nombre d'enfants par famille
- ♦ catégorie socioprofessionnelle (CSP) du chef de famille
- ♦ superficie, en mètres carrés, du logement de la famille.

Le caractère étudié peut être **qualitatif** (catégorie socioprofessionnelle) ou **quantitatif.**
Dans ce dernier cas, il peut être **discret** s'il ne prend que des valeurs isolées (nombre d'enfants), ou **continu** s'il peut prendre toute valeur d'un intervalle (superficie du logement).

2. 2. Modalités – Effectifs – Fréquences

Les valeurs que peut prendre le caractère sont appelées **modalités**. S'il existe p modalités, on les ordonne et on les note : x_1, x_2,..., x_i,... x_p . Le nombre d'individus correspondant à la modalité x_i est son **effectif** : on le note n_i.
L'effectif total de la population étudiée est souvent appelé sa **taille**. On la note traditionnellement N et on a la relation :

$N = n_1 + n_2 + ... + n_p$, que l'on écrit encore : $N = \sum_{i=1}^{p} n_i$.

La **fréquence** de la modalité x_i est le quotient de son effectif n_i par l'effectif total. On la note f_i et, par suite : $f_i = \frac{n_i}{N}$. Il est clair que, quel que soit i, f_i appartient à [0 ; 1].

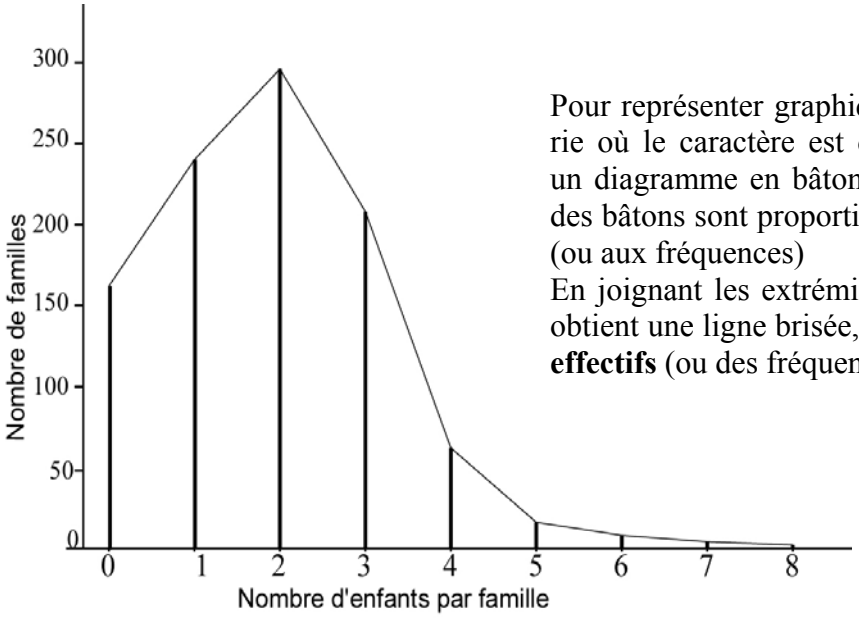
De plus, $\sum_{i=1}^{p} f_i = 1$.

Une série statistique est l'ensemble des couples (x_i, n_i) ou (x_i, f_i) avec $i \in [\![1 ; p]\!]$

■ La notation $[\![1 ; p]\!]$ remplace $[1 ; p] \cap \mathbb{N}$.

Exemple : Nombre d'enfants par famille sur un échantillon de 1000 foyers.

x_i	0	1	2	3	4	5	6	7	8
n_i	162	240	297	208	62	16	8	5	2



Pour représenter graphiquement une telle série où le caractère est discontinu, on utilise un diagramme en bâtons où les « hauteurs » des bâtons sont proportionnelles aux effectifs (ou aux fréquences)
En joignant les extrémités de ces bâtons, on obtient une ligne brisée, appelé **polygone des effectifs** (ou des fréquences).

2. 3. Regroupement en classes

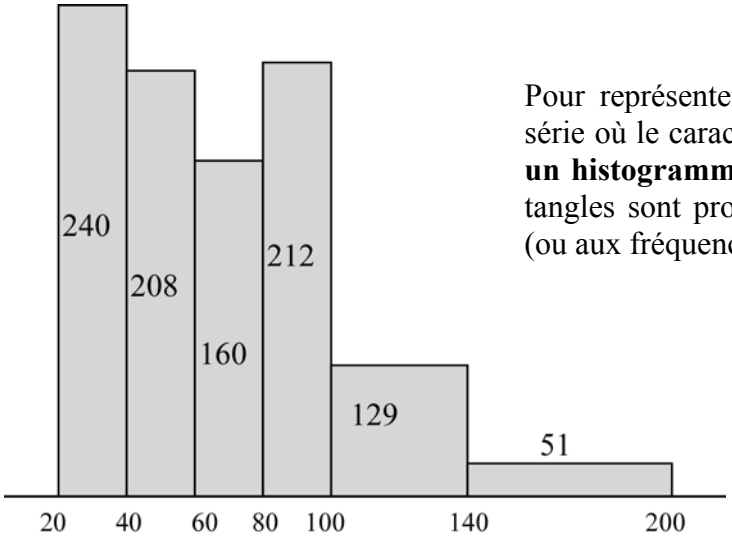
Dans le cas d'une série statistique où le caractère étudié est continu, on peut regrouper les valeurs du caractère en ***classes***.

Exemple : Superficie du logement en m² sur un échantillon de 1000 foyers.

x_i	[20, 40[	[40, 60[	[60, 80[	[80,100[	[100, 140[	[140, 200]
n_i	240	208	160	212	129	51

Pour chaque classe $[a_i, a_{i+1}[$ on désigne par c_i le nombre $\frac{a_i + a_{i+1}}{2}$. Ce nombre est le ***centre*** de la classe concernée. La différence $a_{i+1} - a_i$ est l'***amplitude*** de la classe.

Une telle série statistique est notée : $([a_i, a_{i+1}[, n_i)$ avec $i \in [\![1, p]\!]$.



Superficie des logements

Pour représenter graphiquement une telle série où le caractère est **continu**, on utilise **un histogramme** où les « aires » des rectangles sont proportionnelles aux effectifs (ou aux fréquences).

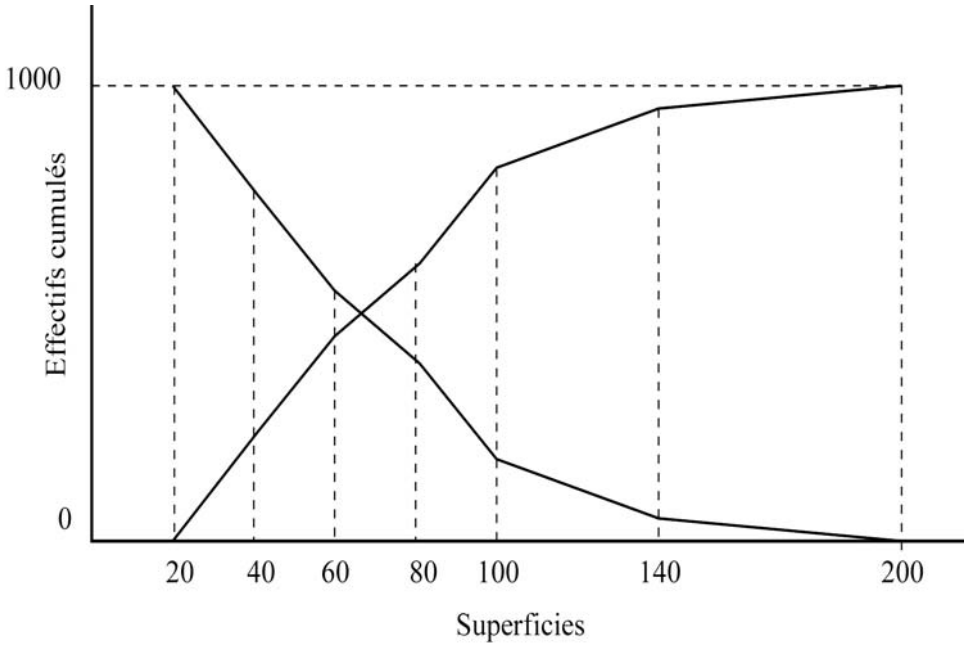
2. 4. Effectifs cumulés – Fréquences cumulées

Chacune des séries statistiques précédentes peut être représentée sous forme d'effectifs (ou de fréquences cumulées). Pour cela on peut construire un tableau d'effectifs (ou de fréquences) de la façon suivante :

- On écrit, dans la première ligne, **toutes les bornes**.
- On réalise ensuite une **partition** de l'effectif total, **relativement à chaque borne**, à l'aide de la relation «strictement inférieur à...» (les classes étant fermées à gauche et ouvertes à droite).

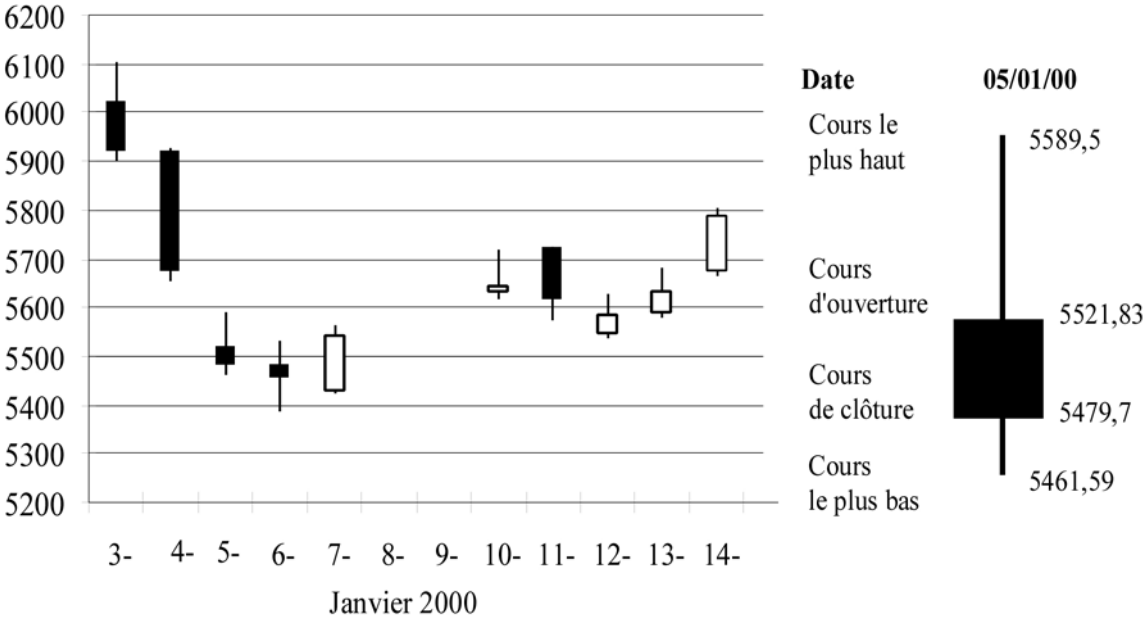
	Bornes	20	40	60	80	100	140	200
Effectifs cumulés croissants	< à...	0	240	448	608	820	949	1000
Effectifs cumulés décroissants	≥ à....	1000	760	552	392	180	51	0

On représente alors la série statistique par deux polygones dits **polygones des effectifs (ou fréquences) cumulés.**



2. 5. Un graphique plus élaboré.

Pour représenter l'évolution d'un caractère en fonction du temps on peut utiliser des graphiques plus complexes donnant simultanément plusieurs informations. Par exemple le graphique boursier ci-dessous donne l'évolution du CAC 40 du 3 au 14 janvier 2000. La partie droite du graphique en explique la conception.
Lorsque la boîte est noire, la valeur de clôture est en baisse. Lorsqu'elle est blanche ce cours de clôture est en hausse.



3. PARAMETRES D'UNE SERIE STATISTIQUE

3. 1. Principe

Le but de cette étude est de substituer à l'ensemble des valeurs de la série statistique à étudier, quelques paramètres dont les valeurs numériques résumeront aussi fidèlement que possible les données initiales. Il faut cependant être conscient que vouloir représenter une série statistique par quelques paramètres, même judicieusement choisis, conduit à une perte non négligeable d'information.

3. 2. Les types de paramètres

Nous allons simplement indiquer ici quelques avantages ou inconvénients liés à l'utilisation de tel ou tel de ces paramètres. On les classe en deux types :

Les paramètres de position

♦ **La moyenne arithmétique** $\bar{x}$ est le paramètre le plus utilisé en raison de ses propriétés algébriques (linéarité). Ce paramètre a l'inconvénient d'être sensible à des valeurs aberrantes. Si on veut calculer la moyenne des revenus d'une population le choix de la moyenne arithmétique n'est pas pertinent car les très hauts revenus tirent le résultat vers le haut.

♦ **La médiane** a justement pour avantage d'être peu sensible aux valeurs extrêmes qui peuvent ne pas être fiables. Dans l'exemple ci-dessus elle est beaucoup plus pertinente que la moyenne arithmétique. Cependant, elle a le très gros inconvénient de mal se prêter aux calculs algébriques.

♦ **Les quantiles** servent à mesurer la symétrie d'une série statistique mais aussi la concentration du caractère étudié.

■ **Les paramètres de dispersion**

♦ **L'étendue** est le plus simple d'utilisation mais ne fournit qu'un renseignement bien grossier sur la série concernée. Il faut cependant savoir qu'il intervient pour l'estimation d'un écart-type théorique dans des échantillons de faible effectif (moins de dix données).

♦ **L'écart-type** σ est fondamental pour tous les phénomènes de type gaussien. Le fait que dans ces distributions la quasi-totalité (95%) des effectifs soit comprise entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ permet de résumer la série statistique à l'aide des deux seuls paramètres $\bar{x}$ et σ.

♦ **L'interquartile** est une caractéristique de dispersion très simple qui permet d'éliminer l'influence des valeurs extrêmes. Il est de très loin préférable à l'étendue mais ne prend en compte que 50% de l'effectif total. Si on souhaite prendre en compte un pourcentage plus important de l'effectif (80%) on pourra, par exemple, utiliser l'interdécile.

CAMP DE BASE

1. PARAMETRES DE POSITION

Nature de la variable / Paramètre	Variable **DISCRETE** Série (x_i, n_i) où $i \in [\![1 ; p]\!]$ avec $N = \sum_{i=1}^{p} n_i$	Variable **CONTINUE** Série $([a_i, a_{i+1}[, n_i)$ où $i \in [\![1 ; p]\!]$, avec $c_i = \frac{a_i + a_{i+1}}{2}$ et $N = \sum_{i=1}^{p} n_i$
MODE (noté M_0)	On appelle ***mode*** toute valeur x_i dont l'effectif (ou la fréquence) est maximal.	On appelle ***classe modale*** toute classe pour laquelle $\frac{n_i}{a_{i+1}-a_i}$ (ou $\frac{f_i}{a_{i+1}-a_i}$) est maximal (sur l'histogramme, la hauteur du rectangle correspondant est maximale).
MEDIANE (notée m_e)	On appelle ***médiane un*** nombre réel m_e tel qu'il y ait autant de valeurs x_i inférieures ou égales à m_e que de valeurs x_i supérieures ou égales à m_e. Si $N = 2k+1$, $m_e = x_{k+1}$ Si $N = 2k$, on prend par convention: $m_e = \frac{x_k + x_{k+1}}{2}$	On appelle ***médiane le*** nombre réel m_e abscisse : ♦ du point d'ordonnée $\frac{N}{2}$ des polygones des effectifs cumulés ♦ ou du point d'ordonnée $\frac{1}{2}$ des polygones des fréquences cumulées.
MOYENNE ARITHMETIQUE (notée $\bar{x}$)	On appelle ***moyenne arithmétique*** le nombre réel $\bar{x}$ tel que : $\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i$ ou $\bar{x} = \sum_{i=1}^{p} f_i x_i$	$\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i$

2. PARAMETRES DE DISPERSION

<table>
<tr><th>Nature de la variable
Paramètre</th><th>Variable DISCRETE
Série (x_i, n_i) où $i \in [\![1 ; p]\!]$
avec $N = \sum_{i=1}^{p} n_i$</th><th>Variable CONTINUE
Série ($[a_i, a_{i+1}[, n_i)$ où
$i \in [\![1 ; p]\!]$, avec $c_i = \frac{a_i + a_{i+1}}{2}$
et $N = \sum_{i=1}^{p} n_i$</th></tr>
<tr><td>ETENDUE
(notée E)</td><td colspan="2">On appelle étendue la différence entre les valeurs extrêmes prises par la variable statistique.
$E = x_p - x_1$ | $E = a_{p+1} - a_1$</td></tr>
<tr><td>ECART MOYEN ABSOLU
(noté e_m)</td><td colspan="2">On appelle écart moyen absolu la moyenne arithmétique des écarts à la moyenne.
$e_m = \frac{1}{N}\sum_{i=1}^{p} n_i \lvert x_i - \bar{x}\rvert$ ou $e_m = \sum_{i=1}^{p} f_i \lvert x_i - \bar{x}\rvert$ | $e_m = \frac{1}{N}\sum_{i=1}^{p} n_i \lvert c_i - \bar{x}\rvert$ ou $e_m = \sum_{i=1}^{p} f_i \lvert c_i - \bar{x}\rvert$</td></tr>
<tr><td>VARIANCE
(notée V ou σ^2)</td><td colspan="2">$V = \frac{1}{N}\sum_{i=1}^{p} n_i (x_i - \bar{x})^2$ ou $V = \sum_{i=1}^{p} f_i (x_i - \bar{x})^2$ | $V = \frac{1}{N}\sum_{i=1}^{p} n_i (c_i - \bar{x})^2$ ou $V = \sum_{i=1}^{p} f_i (c_i - \bar{x})^2$
On a aussi (théorème de *Huyghens-König*).
$V = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - \bar{x}^2$ | $V = \frac{1}{N}\sum_{i=1}^{p} n_i c_i^2 - \bar{x}^2$</td></tr>
<tr><td>ECART-TYPE
(noté σ)</td><td colspan="2">On appelle écart-type σ (sigma) la racine carrée de la variance V</td></tr>
</table>

EN CORDEE

1. MOYENNE ARITHMETIQUE

1. 1. Changement de variable affine

Le professeur SABREDUR corrige les copies du devoir qu'il a donné à ses élèves. En fonction du barème qu'il s'est fixé, les notes obtenues par les vingt premières copies sont les suivantes :

6	4	7	9	11	8	5	12	3	7
1	13	5	7	9	2	8	10	10	5

1°) Calculer la moyenne arithmétique, notée $\bar{x}$, de cette série statistique.

2°) Conscient de la faiblesse des résultats, le professeur décide de relever les notes mais hésite entre deux méthodes.

a) Il envisage de multiplier chaque note par 1,5. Quelles sont alors les vingt notes obtenues ? Calculer la moyenne, $\bar{y}$, de cette nouvelle série statistique. Comparer $\bar{x}$ et $\bar{y}$.

b) Il envisage d'ajouter 2 points à chaque note. Quelle influence cette opération a-t-elle sur la moyenne des notes ?

☞ ☞ Suivez le guide

1°) La moyenne arithmétique de cette série statistique est $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i$ d'où $\bar{x} = 7{,}1$.

2°) a) On multiplie chaque note par 1,5 et on obtient la série statistique (y_i) suivante :

9	6	10,5	13,5	16,5	12	7,5	18	4,5	10,5
1,5	19,5	7,5	10,5	13,5	3	12	15	15	7,5

dont la moyenne est $\bar{y} = 10{,}65$.
Alors $\bar{y} = 1{,}5\ \bar{x}$.

Lorsqu'on multiplie chaque valeur de la série par une constante, la moyenne est multipliée par cette constante.

b) On ajoute deux points à chaque note et on obtient la série statistique (z_i) suivante :

8	6	9	11	13	10	7	14	5	9
3	15	7	9	11	4	10	12	12	7

dont la moyenne est $\bar{z} = 9{,}1$.
Alors $\bar{z} = \bar{x} + 2$.

Lorsqu'on ajoute à chaque valeur de la série une constante, cette constante s'ajoute à la moyenne de la série.

Le cas général est traité dans le paragraphe : "utilisation du symbole Σ".