Fiche

Une étude statistique comprend en général les étapes suivantes :
1. on précise les questions auxquelles on veut répondre ;
2. on procède à une enquête, on collecte les données ;
3. on présente ces données dans un tableau ;
4. on représente cette série statistique à l'aide d'un diagramme ;
5. intervient enfin le mathématicien qui procède au calcul de paramètres permettant de caractériser toute la série statistique à l'aide de quelques nombres.

1. Comment établir le tableau d'une série statistique ?

• Rappelons le sens de quelques termes en statistique :

- l'ensemble étudié lors d'une enquête statistique est la **population** ;
- un élément de cette population est un **individu** ;
- le nombre total d'individus de la population est sa taille ;
- le caractère étudié sur cette population est la **variable statistique** ;
- les valeurs prises par cette variable peuvent être appelées **modalités**.

• La variable est :

- **qualitative,** quand elle prend des valeurs non numériques ;
- **quantitative,** quand elle prend des valeurs numériques.

Quand elle est quantitative, elle peut être :

- **discrète,** quand elle prend un nombre fini de valeurs ;
- **continue,** quand elle prend toute valeur comprise entre deux nombres donnés.

Remarque

Quand le nombre de valeurs prises par la variable statistique est trop grand, on traite la variable comme une variable continue.

• Quand la variable statistique X est discrète, on compte pour chaque valeur le nombre d'individus prenant cette valeur : c'est l'**effectif**. On aboutit à un tableau du type :

Valeur de X	x_1	x_2	...	x_p
Effectif	n_1	n_2	...	n_p

On calcule parfois, pour chaque valeur, les **fréquences relatives** : c'est le rapport $\frac{\text{effectif de la valeur}}{\text{taille de la population}}$.

• Quand le nombre de valeurs prises par la variable statistique est trop grand ou quand la variable est continue, on regroupe les valeurs en **classes**. Ce sont des intervalles semi-ouverts $[x_i, x_{i+1}[$. On appelle **amplitude** de la classe le nombre : $x_{i+1} - x_i$ et **centre** de la classe le nombre : $\frac{x_{i+1}+x_i}{2}$. Pour chaque classe, on compte le nombre d'individus qui prennent une valeur supérieure ou égale à x_i et inférieure à x_{i+1}. Ce sera l'effectif de la classe. On aboutit à un tableau du type :

Valeur de X	$[x_1, x_2[$	$[x_2, x_3[$	...	$[x_{p-1}, x_p[$
Effectif	n_1	n_2	...	n_p

On calcule parfois, pour chaque classe de valeurs, sa fréquence relative : c'est le rapport $\frac{\text{effectif de la classe de valeurs}}{\text{taille de la population}}$.

Remarque

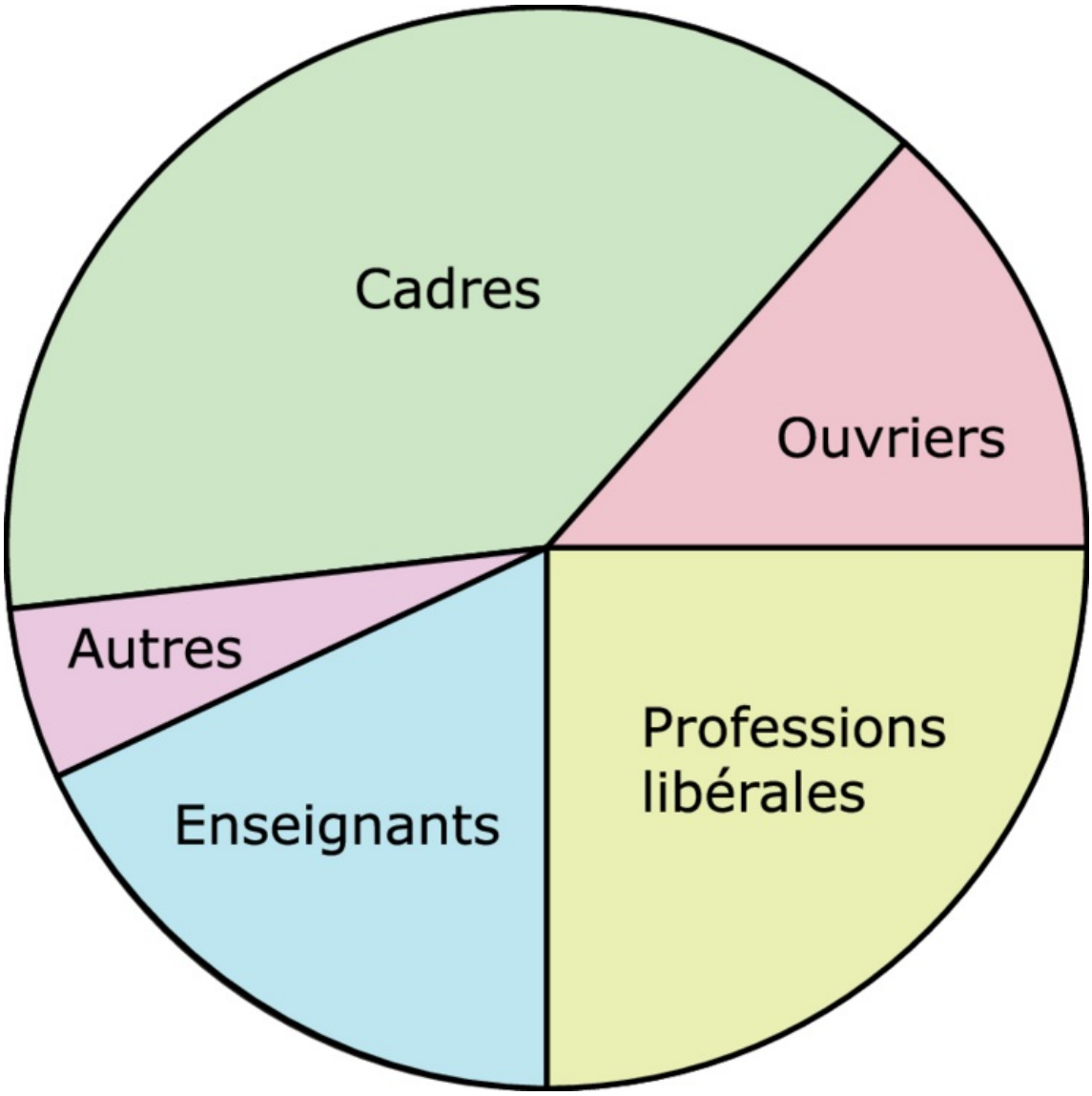
Lors du regroupement des valeurs par classes, on s'efforce d'avoir des classes de même amplitude et qui ne soient pas trop nombreuses. Souvent cependant, les valeurs extrêmes posent problème, d'où des premières ou dernières classes qui sont soit ouvertes soit d'amplitudes différentes.

Exercice n°1

Exercice n°2

2. Comment représenter une série statistique ?

• Pour représenter une variable statistique discrète, on utilise un **diagramme en bâtons** (chaque bâton a une hauteur proportionnelle à l'effectif et/ou à la fréquence) ou un **diagramme circulaire** (chaque secteur est proportionnel à l'effectif et/ou à la fréquence).
Par exemple, la répartition sociologique de 60 étudiants est la suivante : 8 ouvriers ; 23 cadres ; 15 professions libérales ; 11 enseignants et 3 autres.
Pour représenter cette série par un diagramme circulaire, on calcule pour chaque secteur l'angle au centre. Pour le secteur « ouvriers », l'angle au centre est de $(360 \div 60) \times 8 = 48$, soit 48°.
On procède de même pour les autres secteurs et on obtient le diagramme suivant :



• Pour représenter une variable statistique continue, on trace un **histogramme**. L'histogramme est constitué de rectangles juxtaposés dont la surface est proportionnelle à l'effectif de la classe correspondante.
Si les classes ont des amplitudes égales, la hauteur des rectangles est proportionnelle à l'effectif. Si les classes ont des amplitudes inégales, on représente la classe ayant la plus petite amplitude ; puis on compense une amplitude k fois plus grande par une hauteur k fois plus petite.

Exercice n°2

Exercice n°3

4. Comment tracer la courbe des effectifs cumulés ?

• Une courbe des effectifs cumulés (ou des fréquences cumulées) est croissante ou décroissante.

• Pour tracer la courbe des **effectifs cumulés croissants,** on détermine d'abord pour chaque classe $[x_i, x_{i+1}[$ l'effectif cumulé croissant N_i , c'est-à-dire le nombre d'individus qui prennent une valeur inférieure à x_{i+1}. On place ensuite dans un repère les points (x_{i+1}, N_i) , on obtient ainsi la courbe des effectifs cumulés croissants.

• Pour tracer la courbe des **fréquences relatives cumulées croissantes,** on procède de même. N_i est alors remplacé par F_i qui désigne le pourcentage des individus qui prennent une valeur inférieure à x_{i+1}.

• En remplaçant « inférieur » par « supérieur », on obtient de même les courbes des effectifs cumulés **décroissants** ou celle des fréquences relatives cumulées décroissantes.

Exercice n°7

4. Comment calculer une moyenne ?

• Quand la **série statistique** est **discrète**, de taille n, on peut la représenter sous forme d'un tableau du type :

Valeur de X	x_1	x_2	...	x_p	
Effectif	n_1	n_2	...	n_p	n

où $n_1 + n_2 + \ldots + n_p = n$.
On appelle moyenne de X le nombre :
$\overline{X} = \frac{1}{n}(n_1x_1 + n_2x_2 + \ldots + n_px_p)$.

• Quand la série statistique est **continue**, de taille n, on a un tableau du type :

Valeur de X	$[x_1,\ x_2[$	$[x_2,\ x_3[$	...	$[x_p,\ x_{p+1}[$	
Effectif	n_1	n_2	...	n_p	n

Pour calculer la moyenne d'une telle série, on utilise la formule précédente en remplaçant
x_i
par le centre
c_i
de l'intervalle $[x_i,\ x_{i+1}[$.
La moyenne de X est alors le nombre :
$\overline{X} = \frac{1}{n}(n_1c_1 + n_2c_2 + \ldots + n_pc_p)$, où $c_i = \frac{x_{i+1}+x_i}{2}$.

Exercice n°4

Exercice n°5

5. Comment utiliser les propriétés de la moyenne ?

Lorsque l'on **modifie les valeurs d'une série statistique par des opérations simples**, il n'est pas toujours nécessaire de recommencer le calcul de la moyenne.
On utilise les propriétés suivantes :
- si $\overline{X}$ est la moyenne des nombres $x_1,\ x_2,\ \ldots,\ x_n$ et $\overline{Y}$ celle des nombres $y_1,\ y_2,\ \ldots,\ y_n$, alors la moyenne des nombres $x_1 + y_1,\ x_2 + y_2,\ \ldots,\ x_n + y_n$ est $\overline{X+Y}$;
- si k est un réel quelconque et $\overline{X}$ la moyenne des nombres $x_1,\ x_2,\ \ldots,\ x_n$, alors la moyenne des nombres $k + x_1,\ k + x_2,\ \ldots,\ k + x_n$ est $k + \overline{X}$;
- si λ est un réel quelconque et $\overline{X}$ la moyenne des nombres $x_1,\ x_2,\ \ldots,\ x_n$, alors la moyenne des nombres $\lambda x_1,\ \lambda x_2,\ \ldots,\ \lambda x_n$ est $\lambda\overline{X}$.

Exercice n°6

6. Comment calculer une médiane ?

• La médiane est le nombre qui sépare la série ordonnée en valeurs croissantes **en deux groupes de même effectif**.
Pour la trouver, on écrit la liste de toutes les valeurs de la série par ordre croissant, chacune d'elles étant répétée autant de fois que son effectif.
On distingue ensuite deux cas :
- si l'effectif total n est un nombre impair, la médiane est le terme de rang $\frac{n+1}{2}$
- si l'effectif total n est un nombre pair, la médiane est le centre de l'intervalle formé par les termes de rang $\frac{n}{2}$ et $\frac{n}{2} + 1$

• Quand la série est regroupée par classes, on **détermine la médiane graphiquement** à partir du polygone des effectifs ou des fréquences cumulés.

On calcule pour chaque classe $[x_i,\ x_{i+1}[$ l'effectif cumulé croissant N_i, c'est-à-dire le nombre d'individus qui prennent une valeur inférieure à x_{i+1}. On place ensuite dans un repère les points $(x_{i+1},\ N_i)$, on obtient ainsi le polygone des effectifs cumulés croissants.
La médiane est l'abscisse du point dont l'ordonnée est $\frac{n}{2}$.

Exercice n°7

7. Quels autres paramètres peut-on calculer ?

Les mathématiciens disent parfois qu'il existe autant de paramètres statistiques que de statisticiens. Sans aller jusque-là, on peut donner ou calculer, outre la moyenne et la médiane, les paramètres suivants :
- les **valeurs extrêmes**, c'est-à-dire la plus grande valeur X_{max} et la plus petite valeur X_{min} atteintes par la série ;
- l'**étendue**, c'est-à-dire la différence entre la plus grande et la plus petite valeur prises par la variable, soit $X_{max} - X_{min}$;
- le **mode** (ou la classe modale), c'est-à-dire la valeur (ou la classe) ayant le plus grand effectif ;
- le **premier quartile Q_1**, qui est la valeur de la variable au-dessous de laquelle on trouve le quart de l'effectif. Si la série est **discrète**, c'est la valeur de la variable dont le rang est égal ou immédiatement supérieur au quart de l'effectif. Si la série est **continue**, on lit la valeur correspondant à 25 % de l'effectif sur le polygone des fréquences ou des effectifs cumulés. On peut calculer une valeur plus précise par interpolation linéaire.
- le **troisième quartile Q_3**, qui est la valeur de la variable au-dessous de laquelle on trouve les trois-quarts de l'effectif. Si la variable est discrète, c'est la valeur de la variable dont le rang est égal ou immédiatement supérieur aux trois-quarts de l'effectif. Si la série est continue, on lit la valeur correspondant à 75 % de l'effectif sur le polygone des fréquences ou des effectifs cumulés.
- l'écart interquartile, qui est égal à la différence $Q_3 - Q_1$.

Exemple :
Le tableau indique la répartition des logements d'une ville en fonction du nombre de pièces.

Nombre de pièces x_i	1	2	3	4	5	6	7
Pourcentages n_i	10	15	20	30	12	8	5
Pourcentages cumulés croissants	10	10 + 15 = 25	25 + 20 = 45	45 + 30 = 75	87	95	100
Pourcentages cumulés décroissants	100	100 − 10 = 90	90 − 15 = 75	75 − 20 = 55	25	13	5

Mode = 4.
Médiane = 3,5.
Q_1 = 2 car 25 % des logements ont deux pièces ou moins.
Q_3 = 4 car 75 % des logements ont quatre pièces ou moins, c'est à dire que 25 % ont cinq pièces ou plus.

Remarque

Un paramètre quel qu'il soit n'a guère de sens en lui-même. Les enseignements que l'on peut tirer d'une série statistique proviennent plus souvent de la comparaison des paramètres entre eux.

Exercice n°8

Exercice n°9

Exercice n°10

8. Comment calculer une variance et un écart type ?

Soit la série statistique de taille n suivante :

X	x_1	x_2	...	x_p	
Effectif	n_1	n_2	...	n_p	n

On rappelle que la **moyenne** de X est le nombre : $\overline{X} = \frac{1}{n}\left(n_1x_1 + n_2x_2 + \ldots + n_px_p\right)$.
On appelle **variance** de la série statistique X, le nombre :
$V(X) = \frac{1}{n}\left(n_1\left(x_1 - \overline{X}\right)^2 + n_2\left(x_2 - \overline{X}\right)^2 + \ldots + n_p\left(x_p - \overline{X}\right)^2\right)$, qu'on réécrit ainsi :
$V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i\left(x_i - \overline{X}\right)^2$.
L'**écart type** de X est le nombre : $s(X) = \sqrt{V(X)}$.

Exemple

On étudie X l'âge des employés d'une entreprise. On obtient :

Âge	[20 ; 25[	[25 ; 30[	[30 ; 35[	[35 ; 40[	[40 ; 45[	[45 ; 50[	[50 ; 55[	
Effectif	150	300	600	750	450	600	150	3 000

La moyenne de X est :

$$\overline{X} = \frac{1}{3\,000}\left(\begin{array}{l} 150 \times 22,5 + 300 \times 27,5 + 600 \times 32,5 + \\ 750 \times 37,5 + \cdots + 150 \times 52,5 \end{array}\right)$$

$\overline{X} = 38,25$.

La variance de X est :

$$V(X) = \frac{1}{3\,000}\left(\begin{array}{l} 150(22,5 - 38,25)^2 + 300(27,5 - 38,25)^2 + \\ 600(32,5 - 38,25)^2 + \\ 750(37,5 - 38,25)^2 + \cdots + 150(52,5 - 38,25)^2 \end{array}\right)$$

$V(X) = 60,6875$.

Et l'écart type de X est : $s(X) = \sqrt{60,6875} \approx 7,79$.

Remarques

• La variance, l'écart type mesurent la façon dont les valeurs de X se dispersent autour de la moyenne. Ce sont des **paramètres de dispersion** (alors que la moyenne et la médiane sont des paramètres de position, ils précisent vers quelles valeurs se situe la série).

• On peut aussi calculer la variance à l'aide de la formule suivante :

$$V(X) = \frac{1}{n}\left(n_1 x_1^2 + n_2 x_2^2 + \ldots + n_p x_p^2\right) - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2.$$

• Dans le cas où, au lieu d'avoir une valeur x_i, on a un intervalle, les formules sont les mêmes en remplaçant x_i par le **centre de l'intervalle**.

Exercice n°11

Exercice n°12

À retenir

• La moyenne de X est le nombre : $\overline{X} = \frac{1}{n}\left(n_1 x_1 + n_2 x_2 + \ldots + n_p x_p\right)$.

• La médiane est le nombre qui sépare la série en deux groupes de même effectif.

• Au-dessous du premier quartile on trouve le quart de l'effectif, au-dessous du troisième quartile on trouve les trois-quarts de l'effectif.

Soit X une série statistique.

• La variance de X est le nombre : $V(X) = \frac{1}{n}\sum_{i=1}^{p} n_i\left(x_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{p} n_i x_i^2 - \overline{X}^2$.

L'écart type de X est la racine carrée de la variance : $s(X) = \sqrt{V(X)}$.

• Le premier quartile de X, noté Q_1, est la plus petite valeur de la série telle qu'au moins 25 % des données soient inférieures ou égales à Q_1.

Le troisième quartile de X, noté Q_3, est la plus petite valeur de la série telle qu'au moins 75 % des données soient inférieures ou égales à Q_3.

L'intervalle interquartile est l'intervalle $[Q_1 ; Q_3]$.