Fiche n°6

INTERPRETER, REPRESENTER et TRAITER DES DONNEES STATISTIQUES

I. *Introduction : de l'importance des statistiques et des probabilités ?*

Les **statistiques** représentent le domaine des mathématiques qui permet d'**étudier des données réelles** que l'on a obtenues :

- soit par un **recensement** (recueil complet de toutes les données, sur toute la population étudiée) : ceci permet une image précise de ce que l'on désire observer mais pose des problèmes techniques évident s'il y a un grand nombre de données ;
- soit par l'étude d'**échantillons** (recueil de données que sur une partie seulement de la population à étudier : par exemple, sondages…) : ceci présente donc quelques possibilités d'erreur (incertitude) qu'il faut minimiser.

*A ne pas confondre avec les **probabilités** qui proposent des modèles **pour prévoir sans avoir expérimenté** le résultat d'expérience dans lequel intervient le **hasard**…*

Quelques repères dans le temps…

- Les premiers relevés d'hommes et de bien ont eu lieu **vers 3000 ans avant J.-C. en Mésopotamie** ;
- **L'Egypte des pharaons** organisait régulièrement des recensements notamment pour les impôts ;
- **Tycho Brahe** (1546-1601), astronome danois, utilise la moyenne arithmétique pour réduire les erreurs d'observations ;
- Premières analyses de situation de probabilité vers le XVIIe siècle : **Pierre de Fermat, Blaise Pascal**… ;
- Au XVIIIe et XIXe siècle se développe la **théorie des erreurs** ;
- Essor à partir XXe siècle :
  - En statistiques, les **ordinateurs** permettent de **nombreuses simulations**…
  - En probabilités, développement de la théorie actuelles des probabilités…

Quelle utilité dans le monde contemporain ?

- ***Pour trouver et décrire une relation*** : en médecine, on établit le risque cardio-vasculaire lié au tabac en étudiant le pourcentage de fumeurs chez les cardiaques et le pourcentage de cardiaques chez les fumeurs et les non-fumeurs ;
- ***Prendre une décision*** : l'amélioration annuelle des semences de céréales par croisements successifs, les contrôles de fabrication et de fiabilité dans l'industrie, d'efficacité d'un médicament, etc. sont très dépendants des tests statistiques ;
- ***Prévoir et planifier*** : de nombreuses statistiques économiques sont publiques (INSEE) et servent par exemple de base aux négociations syndicales ou intergouvernementales.

Exemples de questions d'actualités :

- **Climat** : assiste-t-on à un réchauffement de la planète ?
- **Santé** : faut-il encore vacciner les enfants contre la variole ? faut-il mettre en quarantaine une population lors d'une pandémie comme le « Coronavirus Covid-19 » ?
- **Paris sportifs** : une activité à risques ?
- **Qualité industrielle** : comment faire pour être « sûr » que dans un lot de 1000 piles électriques vendues, il y en a au moins 995 qui fonctionnent correctement ?
- **Météo** : fera-t-il beau dimanche ?
- **Population** : quel pays aura le plus d'habitant en 2020 ?

II. *Graphiques : comment choisir une bonne représentation ?*

On peut représenter des données avec plusieurs graphiques selon :

- ce que l'on souhaite observer : le **caractère** ;
- ce que l'on souhaite montrer grâce au graphique : pour comparer des **effectifs**, pour répartir des **fréquences** ou pour **regrouper des données** (histogramme).

Quel type de **caractère** dois-je étudier et observer ?

Caractères **numériques**

(nombres ordonnés > axe des abscisses)

Exemples : âge, temps ou durée, masse, notes d'un contrôle, prix ou salaires, etc.

Caractères **qualitatifs**

Exemples : couleur des yeux, boisson préférée, langue parlée, secteurs d'activités, jours de la semaine, etc.

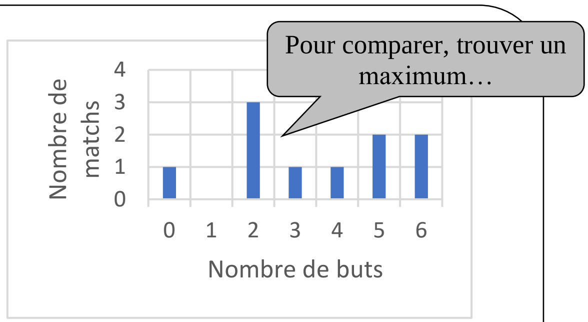


Diagramme en barres (bâtons)

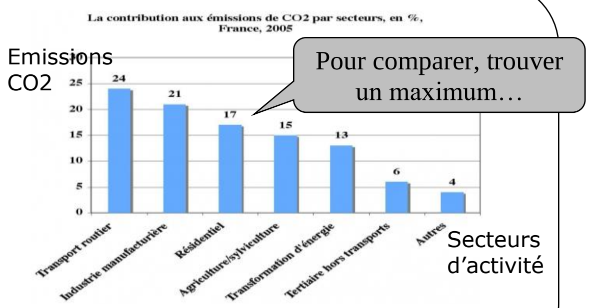
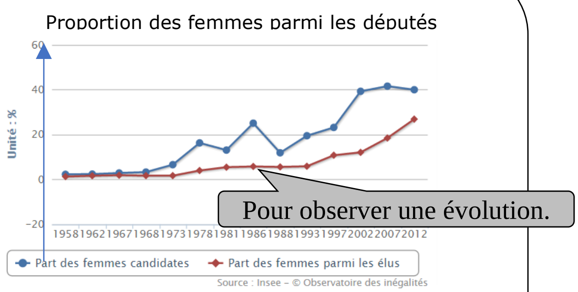


Diagramme en barres (bâtons)



Courbe

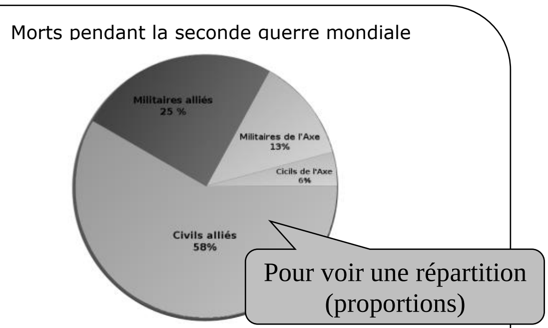
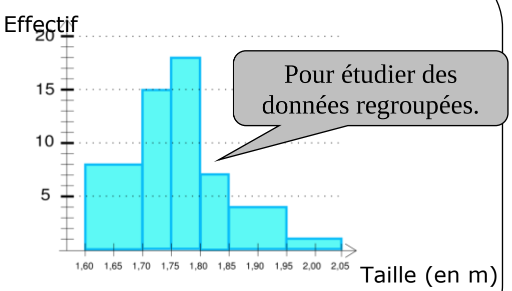
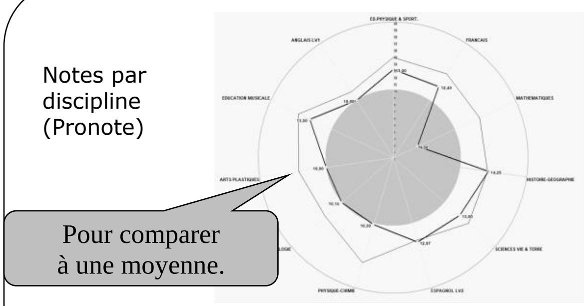


Diagramme circulaire

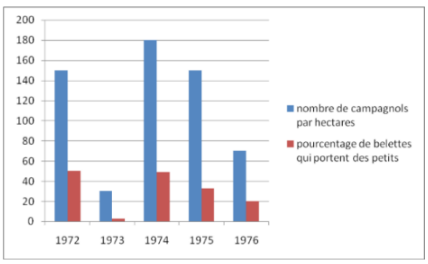


Histogramme (regroupement)



Et bien d'autres diagrammes…

Remarque On peut parfois comparer deux caractères sur un même graphique.
Par exemple, ce graphique ci-contre permet de comparer des populations entre proies (campagnols) et prédateurs (belettes).



III. Effectifs et fréquences

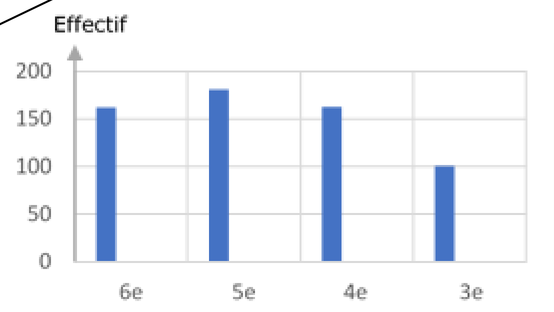
1. Effectifs

Le tableau ci-contre donne la répartition des ***effectifs*** des élèves dans un collège dont l'***effectif total*** est de 607 élèves.

	6ème	5ème	4ème	3ème	Total
Effectifs	162	181	163	101	**607**

Exemple L'effectif des 5èmes est de 181 élèves.

Graphique On peut représenter ces effectifs par un diagramme en barres où la *hauteur de chaque barre* est proportionnelle à l'effectif qu'elle représente.



2. Fréquences (et pourcentages)

Définition ***Fréquence*** = $\dfrac{\textbf{\textit{effectif}}}{\textbf{\textit{effectif total}}}$

	6ème	5ème	4ème	3ème	Total
Effectifs	162	181	163	101	**607**
Fréquences	0,267	0,298	0,268	0,167	**1**

Exemple

La fréquence des élèves de 3ème dans ce collège est 0,167 environ car $\mathbf{\dfrac{101}{607} \approx 0{,}167}$

A savoir ¤ Une fréquence est un nombre compris entre 0 et 1.

¤ ***Fréquence en %*** = $\dfrac{\textbf{\textit{effectif}}}{\textbf{\textit{effectif total}}} \times 100$

On obtient alors :

	6ème	5ème	4ème	3ème	Total
Effectifs	162	181	163	101	**607**
Fréquences (en %)	26,7	29,8	26,8	16,7	**100**

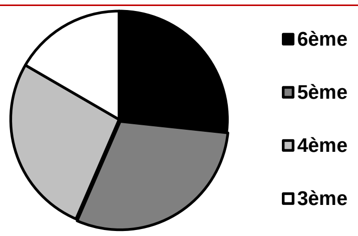
Exemple

Le pourcentage de 4ème dans ce collège est 26,8 % environ car $\mathbf{\dfrac{163}{607} \times 100 \approx 26{,}8}$

A savoir Dans un **diagramme circulaire**, la *mesure de chaque angle* est proportionnelle à l'effectif ou la fréquence qu'il représente.

Mesure d'un angle (en degrés) = $\dfrac{\textbf{\textit{effectif}}}{\textbf{\textit{effectif total}}} \times 360$

	6ème	5ème	4ème	3ème	Total
Effectifs	162	181	163	101	**607**
Mesures (en degrés)	96	107	97	60	**360**



Exemple L'angle représentant les 5èmes mesure 107° environ car $\mathbf{\dfrac{181}{607} \times 360 \approx 107}$

IV. Moyenne pondérée d'une série statistique

EXERCICE TYPE 1

On a interrogé plusieurs personnes sur leur taille : les résultats ont été reportés sur un tableur dans la feuille de calcul suivante.

	A	B	C	D	E	F
1	Taille (en m)	1,70	1,75	1,80	1,85	*Total*
2	Effectif	3	6	4	2	15
3	Fréquence (en %)					100

1. Quelle formule permet de calculer le nombre total de personnes interrogées ?
2. a. Quel est le pourcentage de personnes mesurant au moins 1m80 ?
   b. Dans la cellule B3, quelle formule permet de calculer le pourcentage de personne mesurant 1m70 ?
3. a. La formule « =MOYENNE(B1:E1) » permet-elle de calculer la taille moyenne pour ce groupe de personnes ?
   b. Calculer la taille moyenne pour ce groupe de personnes.

Solution

1. La formule permettant de calculer le nombre total de personnes interrogées est :
   =SOMME(B2:E2)

Sur tableur, dans une formule, le symbole « **:** » (deux points) signifie « **jusqu'à** ».

Par exemple, **B2:E2** signifie de **B2 jusqu'à E2…**

2. a. Il y a 4+2 = 6 personnes mesurant au moins 1m80 (c'est-à-dire 1m80 ou plus).
   $\frac{\text{effectif}}{\text{effectif total}} \times 100 = \frac{6}{15} \times 100 = 40.$
   40 % des personnes interrogées mesurent au moins 1m80.
   b. La formule permettant de calculer le pourcentage de personne mesurant 1m70 est :
   =B2/15*100
3. a. Non, la formule « =MOYENNE(B1:E1) » ne permet pas de calculer la taille moyenne pour ce groupe de personnes car elle ne prend pas en compte les effectifs.
   b. $1,70\times3 + 1,75\times6 + 1,80\times4 + 1,85\times2 = 26,5$ et $26,5 \div 15 \approx 1,766$
   La taille moyenne de ce groupe de personnes est **environ 1,77 m**.

EXERCICE TYPE 2

Tailles entre 1,80 m compris et 2 m non compris

Déterminer la taille moyenne des élèves de cette classe :

Taille (en m)	[1,50 ; 1,60[	[1,60 ; 1,70[	[1,70 ; 1,80[	[1,80 ; 2[	*Total*
Centre	(1,50 + 1,60) ÷ 2 = 1,55	(1,60 + 1,70) ÷ 2 = 1,65	(1,70 + 1,80) ÷ 2 = 1,75	(1,80 + 2) ÷ 2 = 1,90	
Effectif	3	13	8	2	26

Solution Pour des données regroupées, on utilise le **centre** de l'intervalle pour calculer la moyenne (voir les calculs dans le tableau)

$1,55\times3 + 1,65\times13 + 1,75\times8 + 1,90\times2 = 43,9$ et $43,9 \div 26 \approx 1,69$

La taille moyenne des élèves de la classe est ***environ 1,69 m***.

V. *Médiane d'une série statistique*

Définition La **médiane** d'une série ordonnée est **une valeur** telle qu'il y ait **autant de valeurs inférieures** ou égales **que de valeurs supérieures** ou égales.

EXERCICE TYPE 3 Déterminer les médianes des séries de notes suivantes.
- série A : 13, 13, 20, 19, 18, 15, 15
- série B : 8, 8, 9, 12, 15, 17, 12, 11, 14, 14
- série C : 17, 14, 3, 16, 5, 17

Solution Pour déterminer une médiane, ***il faut d'abord ordonner la série***.

- série A : $\underbrace{13 \leq 13 \leq 15}_{\text{3 notes}} \leq \boxed{15} \leq \underbrace{18 \leq 19 \leq 20}_{\text{3 notes}}$. La médiane de cette série est **15**.

- série B : $\underbrace{8 \leq 8 \leq 9 \leq 11 \leq 12}_{\text{5 notes}}$ ... $8 \leq 8 \leq 9 \leq 11 \leq \boxed{12 \leq 12} \leq 14 \leq 14 \leq 15 \leq 17$. La médiane est **12**.

- série C : $3 \leq 5 \leq \boxed{14 \leq 16} \leq 17 \leq 17$.

La médiane de cette série est entre 14 et 16. Par habitude, on prendra alors la valeur centrale : la médiane de cette série C est donc **15**.

Remarques

¤ Deux séries peuvent avoir la même moyenne mais pas la même médiane (séries B et C).

¤ Deux séries peuvent avoir la même médiane mais pas la même moyenne (séries A et C).

	Série A	Série B	Série C
Médiane	**15**	**12**	**15**
Moyenne	**≈ 16,1**	**12**	**12**

VI. *L'étendue*

Définition L'***étendue*** d'une série est la différence entre les deux valeurs extrêmes.

EXERCICE TYPE 4 Déterminer l'étendue des séries A, B et C de l'exercice type 3.

Solution
- série A : 20 − 13 = 7. L'étendue de cette série est 7.
- série B : 17 − 8 = 9. L'étendue de cette série est 9.
- série C : 17 − 3 = 14. L'étendue de cette série est 14.

Remarque

¤ La moyenne et la médiane donnent des informations sur la **position** et la répartition des valeurs.

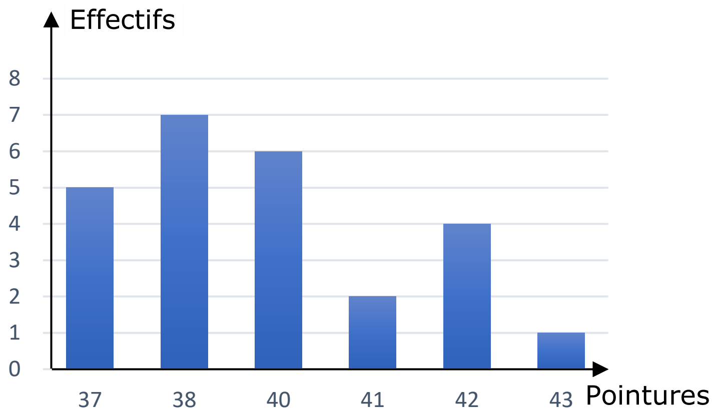
¤ L'étendue donne uniquement des informations sur la **dispersion**, mais pas sur la répartition entre la valeur minimale et la valeur maximale.

VII. SITUATIONS STATISTIQUES (EXERCICES BILAN CYCLE 4)

EXERCICE TYPE 5

Une enquête a été réalisée dans une classe de 4e : ses résultats sont indiqués dans le diagramme ci-contre.

1. Déterminer le nombre de personnes :
   - **a.** chaussant moins de 40.
   - **b.** chaussant au moins du 41.
2. Quel est le pourcentage de personnes chaussant du 38 ?
3. Déterminer la pointure moyenne de ce groupe de personnes.
4. Thibaut affirme que, dans cette étude, la médiane est 39. A-t-il raison ? Justifier.



Solution

1. **a.** Les personnes chaussant moins de 40 sont celles dont la pointure est 37, 38 ou 39 : il y a donc 5 + 7 = **12 personnes chaussant moins de 40**.
   b. Les personnes chaussant au moins du 41 sont celles dont la pointure est 41, 42 ou 43 : il y a donc 2 + 4 +1 = **7 personnes chaussant au moins du 41**.

2. Il y a 7 de personnes chaussant du 38.

   Au total, il y a 5 + 7 + 6 + 2 + 4 + 1 = 25 collégiens interrogés.

   $$\frac{\text{effectif}}{\text{effectif total}} \times 100 = \frac{7}{25} \times 100 = 28.$$

   28 % des personnes interrogées chaussent du 38.

3. Pour calculer la pointure moyenne de ce groupe de personnes, il faut absolument tenir compte des effectifs dans chaque pointure…

   $37\times5 + 38\times7 + 40\times6 + 41\times2 + 42\times4 + 43\times1 = 984$

   $984 \div 25 = 39{,}36$

   La pointure moyenne de ce groupe de personnes est 39, 36.

4. Thibaut affirme que, dans cette étude, la médiane est 39. D'après la définition de la médiane, cela signifierait qu'il y a autant de personnes ayant une pointure inférieure ou égale à 39 que de personnes ayant une poiture supérieure ou égale à 39…

   Comme il y a 12 personnes ayant une pointure inférieure ou égal à 39 mais 13 ayant une pointure supérieure ou égal à 39, **la pointure 39 ne peut pas être une médiane de cette série statistique**.

Remarque : dans cet exemple, la médiane de la série est la pointure 40. En effet, comme il y a 25 personnes interrogées et que 25 ÷ 2 = 12,5, la médiane devra être la 13ème valeur de la série ordonnée pour qu'il y ait 12 valeurs inférieures ou égal et 12 valeurs supérieures ou égales… La médiane de cette série est donc 40.

EXERCICE TYPE 6

Une enquête a été réalisée auprès de 2 500 personnes à partir de la question suivante :

« A quel âge avez-vous trouvé un emploi correspondant à votre qualification ? ».

Les résultats de l'enquête ont été reportés sur un tableur dans la feuille de calcul suivante.

	A	B	C
1	Âge	Centre	Effectif
2	[18 ; 22[		100
3	[22 ; 26[		200
4	[26 ; 30[		400
5	[30 ; 34[		1 100
6	[34 ; 38[		700
7	Total		2 500

1. Représenter les résultats de cette enquête par un histogramme.

2. Dans la cellule C7, quelle formule permet de calculer le nombre total de personnes interrogées ?

3. **a.** Compléter la colonne B de cette feuille de calcul.

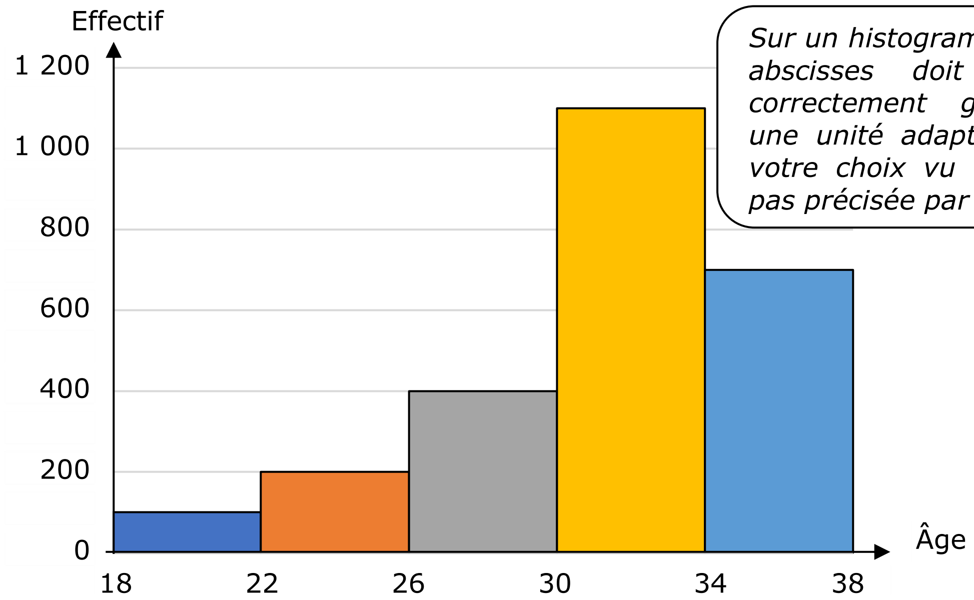
b. La formule « =MOYENNE(B2:B6) » permet-elle de calculer l'âge moyen de ce groupe de personnes ?

c. Calculer cet âge moyen.

4. Est-il vrai que la médiane de cette série statistique est inférieure à 30 ans ? Justifier.

<u>*Solution*</u>

1.



Sur un histogramme, l'axe des abscisses doit aussi être correctement gradué (avec une unité adaptée… mais de votre choix vu qu'elle n'était pas précisée par l'énoncé.)

2. Dans la cellule C7, on calcule le nombre total de personnes interrogées grâce à la formule :

=SOMME(C2:C6)

3. **a.** On obtient la feuille de calcul ci-contre.

	A	B	C
1	Âge	Centre	Effectif
2	[18 ; 22[	**20**	100
3	[22 ; 26[	**24**	200
4	[26 ; 30[	**28**	400
5	[30 ; 34[	**32**	1 100
6	[34 ; 38[	**36**	700
7	Total		2 500

b. Non, la formule « =MOYENNE(B2:B6) » ne permet pas de calculer l'âge moyen de ce groupe de personnes car elle ne prend pas en compte les effectifs pour chaque classe d'âge.

c. $20\times100 + 24\times200 + 28\times400 + 32\times1100 + 36\times700 = 78\,400$

$78\,400 \div 2\,500 = 31{,}36$

L'âge moyen de ce groupe de personnes est 31,36 ans.

4. Il y a 2 500 personnes interrogées : la moitié des personnes interrogées représente donc 1 250 personnes.

Le nombre de personnes dont l'âge est inférieur à 30 ans est 100+200+400 = 700.

La médiane de cette série statistique n'est donc pas inférieure à 30 ans puisque 700 ne représente pas la moitié des effectifs (voir définition de la médiane si besoin…).