

UAA 1 :

Statistique descriptive à une variable

Théorie



L'élève doit SAVOIR :

1. Expliquer et donner la notation qui y est liée (si elle existe) des mots suivants :
 - (1) "effectif total"
 - (2) "modalités"
 - (3) "caractère quantitatif"
 - (4) "caractère qualitatif"
 - (5) "fréquence d'une modalité"
 - (6) "moyenne arithmétique"
 - (7) "médiane"
 - (8) "premier quartile"
 - (9) "deuxième quartile"
 - (10) "troisième quartile"
 - (11) "mode" (ou "classe modale")
 - (12) "variance"
 - (13) "écart-type"
2. Expliquer la différence entre un caractère discret et un caractère continu.
3. Associer un mot à sa définition.

L'élève doit ÊTRE CAPABLE DE :

1. Déterminer la population, l'effectif total, l'effectif d'une modalité, le caractère, les modalités et dire s'il s'agit d'un caractère qualitatif ou quantitatif, d'un caractère discret ou continu pour une série statistique.
2. Utiliser le vocabulaire statistique.
3. Etablir le tableau recensé ou le tableau groupé d'une série statistique ; et en déduire des renseignements.
4. Tracer un diagramme en bâtonnets, un histogramme, un diagramme des effectifs cumulés et un diagramme des fréquences cumulées.
5. Lire un graphique, un diagramme, un tableau et l'interpréter.
6. Déterminer la moyenne arithmétique, la médiane, les quartiles et le mode (ou la classe modale) ; et en donner une interprétation.
7. Donner l'étendue d'une série statistique.
8. Calculer la variance et l'écart-type et en donner une interprétation.

9. Lire une boîte à moustaches pour en déduire des informations sur une série statistique.
10. Représenter une boîte à moustache.
11. Utiliser l'inégalité de Tchebychev.
12. Utiliser la mémoire de la calculatrice.

"Il est prouvé que fêter les anniversaires est bon pour la santé.

Les statistiques montrent que les personnes qui en fêtent le plus deviennent les plus vieilles."

(Den Hartog)

« Il y a trois sortes de mensonges :

les petits mensonges, les sacrés mensonges et les statistiques. »

Jadis, lorsque les voyages par mer étaient beaucoup plus périlleux que maintenant, lorsque les navires battus par les tempêtes jetaient une partie de leur cargaison par-dessus bord, il fut admis que ceux dont les marchandises avaient été sacrifiées avaient droit à une indemnité équitable, aux frais de ceux dont les marchandises étaient arrivées à bon port. La valeur des biens perdus était remboursée suivant un accord entre tous ceux dont les marchandises se trouvaient à bord du même navire. Ce dommage en cours de route était appelé *havaria* (avarie) et cette appellation fut tout naturellement reportée sur le dédommagement que chaque particulier était invité à payer. C'est de ce mot latin que dérive le terme anglais moderne *average* (moyenne dans le sens de valeur centrale). L'idée d'une moyenne a donc des racines dans des formes primitives d'assurance.

Très naturellement, par suite de l'extension de la navigation, les assurances furent fondées sur des bases plus fermes : elles répartirent les risques, non plus entre ceux dont les marchandises étaient exposées au cours d'une traversée, mais entre un grand nombre de négociants. Par la suite, le fait de garantir de tels risques devient une profession spécialisée et rémunératrice, entraînant le paiement à l'assureur d'une prime en rapport avec le risque couvert.

M. J. Moroney, *Comprendre la statistique*, Penguin Books, 1951.

Actuellement, la moyenne reste la valeur la plus utilisée en statistique, mais on s'en sert avec prudence car sa signification et sa portée sont très différentes d'une situation à l'autre. La moyenne ne donne en effet aucune information sur la dispersion des données. C'est pourquoi elle doit être complétée par une valeur rendant compte de cette dispersion et qui conditionne la fiabilité de la moyenne. Cette valeur est basée sur le calcul de la variance. De même pour la médiane : sa portée sera nuancée et complétée par l'écart interquartile.



Les statistiques forment donc un ensemble de méthodes scientifiques consistant à collecter, à traiter les données collectées et à analyser ces données pour en tirer des conclusions et prendre des décisions judicieuses.

Exemple d'application : l'accidentologie (études scientifiques des accidents)

Grâce aux statistiques, nous avons aujourd'hui une meilleure connaissance de l'accidentologie. Les premières statistiques des accidents de la circulation remontent pratiquement à la naissance de l'industrie automobile. Mais, au fil des années, elles n'ont cessé de s'affiner, gagnant en fiabilité, en précision et en rapidité, afin de disposer d'une connaissance détaillée de notre accidentologie.

Il a trois bonnes raisons à cela.

- Il s'agit pour les pouvoirs publics d'avoir la **vision** la plus claire, la plus précise possible sur les causes et les conditions des accidents qui surviennent sur les différents réseaux (les lieux, les conditions atmosphériques, l'éclairage, les individus et les véhicules impliqués,...). Le choix des « armes » dépend étroitement de l'ennemi que l'on a à combattre... et seules les statistiques permettent de bien le définir.
- Une fois les **actions** décidées et mises en œuvre, il s'agit d'évaluer leur efficacité sur le terrain. Là encore, les statistiques permettent de mesurer objectivement les effets des actions entreprises, de façon à pouvoir les généraliser si celles-ci sont positives, ou bien à les amender si elles ne donnent pas entière satisfaction.
- La diffusion des statistiques permet de faire **partager** à l'ensemble des usagers notre **connaissance** sur l'accidentologie. C'est indispensable si l'on veut obtenir leur adhésion à une lutte qui passe forcément par un consensus social. Savoir, par exemple, qu'une baisse sensible des accidents qui coïncide avec la mise en place de nouvelles mesures peut convaincre des usagers, jusque-là incrédules, de l'intérêt de ces mesures...

Des mesures de vitesses sont réalisées sur les différents réseaux routiers. Les appareils utilisés permettent également de repérer les interdistances entre les véhicules. L'évolution des comportements sur certains points importants (l'alcool, le port de la ceinture et du casque, etc.) vient compléter les données scientifiques. Et d'autres enquêtes doivent s'y ajouter : sur le respect des feux rouges, l'utilisation des téléphones portables au volant, etc...

A. Données statistiques

1. Vocabulaire



A. 1. VOCABULAIRE
<https://bit.ly/3wxwzJ7>

Pour découvrir les premières notions de statistiques, nous allons nous intéresser au **nombre de frères et sœurs des élèves d'une classe de 4^e**.

1	2	1	2	0
0	2	2	1	1
1	3	2	2	5
2	1	0	6	0
2	1	2	1	2

Cette liste de nombres constitue un **tableau statistique** (ou une **série statistique**). Comme les nombres n'y sont pas disposés dans un ordre mathématique particulier, mais placés tels qu'ils ont été recueillis, il est appelé **tableau brut**.

Nous allons étudier ces nombres statistiquement...

Population : Ensemble des éléments auxquels se rapporte une étude statistique.

Exemple : La population est la classe de 4^e.

Un **individu** est un élément de la population.

Remarque : Ce vocabulaire particulier de population, tient au fait que beaucoup de travaux statistiques portent réellement sur des ensembles de personnes. Mais il s'applique à tout type de données : on pourrait, par exemple, étudier la taille des arbres, les causes des accidents de la circulation ou le type de véhicules impliqués, le chiffre d'affaires d'une société au cours des années, etc.

Effectif total : Nombre d'individus de la population

Notation : n .

Exemple : L'effectif total vaut 25.

Caractère : Caractéristique présentée par les individus de la population.

Exemple : Le caractère étudié est le nombre de frères et sœurs.

En statistiques, on étudie un **caractère** présenté par des **individus** d'une **population**. On dispose rarement des résultats relatifs à cette population et on se restreint à l'étude d'un **échantillon** de cette population.

Les individus recensés d'après un caractère statistique forment une **série statistique**.

Modalités : Différentes formes prises par un caractère.

Notation : x_i

Exemple : Les modalités sont les nombres 0, 1, 2, 3, 5 et 6.

Caractère quantitatif : Caractère dont les modalités sont mesurables.

Caractère qualitatif : Caractère dont les modalités ne sont pas mesurables.

Exemple : Le caractère est quantitatif.

Remarques :

- Un caractère qualitatif peut néanmoins s'exprimer par des nombres, comme le code postal, mais ces nombres ne représentent pas des quantités, ce ne sont que des « codages ».
- Un caractère quantitatif peut être **discret** s'il ne prend que des valeurs isolées (nombre d'enfants de la famille par exemple) ou **continu** s'il peut prendre toute valeur d'un intervalle (superficie de logement). Nous y reviendrons un peu plus loin dans ce chapitre.

Dans la suite de ce cours, nous n'étudierons que des variables statistiques quantitatives.

2. Tri de données



A. 2. TRI DE DONNÉES

Après avoir effectué une enquête à propos d'un caractère statistique, les résultats sont écrits dans un **tableau brut**. Comme ce tableau n'est pas très parlant (tel quel il ne donne pas une idée claire des résultats), on rassemble les résultats dans un **tableau recensé**.

Effectif d'une modalité : Nombre d'individus qui présentent cette modalité.

Notation : e_i

Fréquence d'une modalité : Quotient de l'effectif de la modalité par l'effectif total.

Notation et formule : $f_i = \frac{e_i}{n}$

Effectif cumulé d'une modalité : Somme des effectifs de la modalité et des modalités qui la précèdent dans le tableau.

Notation et formule: $\varepsilon_i = e_1 + e_2 + e_3 + \dots + e_i$

Fréquence cumulée d'une modalité : Somme des fréquences de la modalité et des modalités qui la précèdent dans le tableau.

Notation et formule : $F_i = f_1 + f_2 + f_3 + \dots + f_i$

On peut ainsi construire le **tableau recensé** en ordonnant les modalités dans l'ordre **croissant** et en calculant, pour chaque modalité, les diverses grandeurs que nous venons de définir.

Modalités	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
x_i	e_i	f_i (en %)	ε_i	F_i (en %)
0	4	16	4	16
1	8	32	12	48
2	10	40	22	88
3	1	4	23	92
5	1	4	24	96
6	1	4	25	100
Total :	25	100		

Dans cet exemple, nous avons donc : $x_1 = 0$, $e_2 = 8$, $f_5 = 4\%$ ou encore $\varepsilon_4 = 23$.

Cette présentation des informations est utile pour répondre à différentes questions :

1. Combien d'élèves ont exactement 2 frères et/ou sœurs ? **10 élèves**
2. Quel pourcentage des élèves n'a pas de frère et/ou sœur ? **16 %**
3. Quel pourcentage des élèves a au plus 3 frères et/ou sœurs ? **92 % des élèves**
4. Quel pourcentage des élèves a au moins 3 frères et/ou sœurs ? **12 % des élèves**



Un tableau recensé contiendra donc toujours cinq colonnes nommées "modalités" (rangées dans l'ordre croissant), "effectifs", "fréquences", "effectifs cumulés" et "fréquences cumulées".

Remarque importante :

Les fréquences en % sont généralement exprimées par un nombre comportant une seule décimale (donc trois décimales pour les fréquences exprimées par des nombres décimaux).

L'arrondi se fait selon la règle habituelle :

- si le premier chiffre qui suit la dernière décimale conservée est inférieur à cinq, la dernière décimale reste telle quelle,
- si le premier chiffre qui suit la dernière décimale conservée est supérieur ou égal à cinq, la dernière décimale est augmentée de 1 unité.

A cause de l'arrondi, la somme des fréquences peut être légèrement différente de 1 : il convient de toujours négliger cette différence et de toujours écrire 1.

Qu'est-ce qu'un caractère continu ?



A. 2. CARACTÈRE CONTINU

A priori, nous savons que certaines variables statistiques ne peuvent posséder qu'un nombre fini de modalités : ce sont des variables statistiques discrètes.

Par exemple, si on lance une paire de dés 100 fois de suite et qu'à chaque lancer, on note la somme des points figurant sur les faces supérieures des deux dés, on obtient 100 nombres mais, au plus, 11 modalités : 2, 3, 4, ..., 12. La somme des points dans le jet de deux dés est donc une variable statistique discrète.



Notons qu'en 6^{ème} année, nous étudierons la probabilité (fortement liée à la notion de fréquence) que la somme des deux dés soit 7, par exemple.

Mais il arrive que des variables statistiques puissent prendre n'importe quelle valeur dans un intervalle donné ou, tout simplement, n'importe quelle valeur parmi un grand nombre de valeurs possibles. On parle alors de **variables statistiques continues**.

Dans ce cas, les modalités sont trop nombreuses ! Il est alors préférable de les regrouper par **classes**, ce qui permet d'éviter des tableaux recensés très longs dans lesquels les fréquences, trop petites, n'auraient plus vraiment de signification.

Ces classes ont généralement la même largeur.

On parle de manière analogue de l'effectif et de la fréquence d'une classe. Le tableau dans lequel on rassemble les modalités en classes s'appelle un **tableau groupé**.

Par exemple, voici le tableau groupé donnant la superficie de logements en m² sur un échantillon de 1000 foyers.

Les 3 dernières colonnes sont construites comme pour un tableau recensé.

Modalités x_i	Effectifs e_i	Fréquences f_i (en %)	Effectifs cumulés $\mathcal{E}_i$	Fréquences cumulées F_i (en %)
[20;40[	50	5	50	5
[40;60[	130	13	180	18
[60;80[	190	19	370	37
[80;100[	160	16	530	53
[100;140[	320	32	850	85
[140;200[	150	15	1000	100
Total :	1000	100		

Remarque importante :

Il faut se rendre compte qu'en considérant un tableau groupé (en classes), on diminue l'information contenue dans ce tableau. Un choix judicieux des classes peut heureusement limiter la perte. En ce qui nous concerne, nous ne nous intéresserons pas aux normes de construction des classes.

B. Graphiques



A. 3. GRAPHIQUES

Après avoir collecté des données et les avoir ordonnées dans des tableaux recensés ou groupés, il est intéressant de pouvoir les **représenter graphiquement**. Il existe différents types de graphiques : à bâtonnets, en rectangles, circulaires, figuratifs, en pyramides, ... En fonction de ce que l'on souhaite montrer, on préférera l'un plutôt que l'autre.



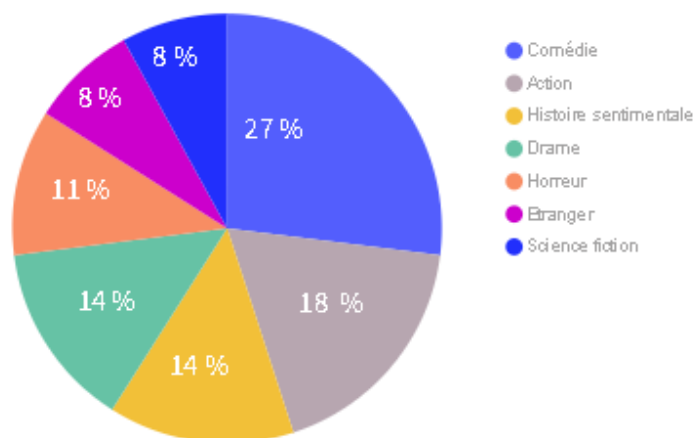
Si le caractère quantitatif est discret (tableau recensé) ou s'il est continu (tableau groupé), on n'utilisera pas le même type de représentations, même si elles peuvent porter le même nom.

1. Diagramme des effectifs pour un caractère qualitatif

Le **diagramme circulaire** (communément appelé *camembert*) est particulièrement adapté pour les séries à caractère qualitatif.

Chaque modalité est représentée par un secteur circulaire dont l'aire et l'angle sont proportionnels à l'effectif correspondant.

Exemple : Voici le diagramme circulaire concernant le type de film préféré dans la classe de cinéma de Mme Robert.



Ce type de graphique ne sera pas utilisé dans ce cours.

1. Graphique des effectifs pour un caractère quantitatif

(1) Caractère quantitatif discret

Le **diagramme en bâtonnets** est particulièrement adapté pour les séries à caractère quantitatif discret (tableau recensé).

Sur l'axe des abscisses, on indique les modalités du caractère discret étudié.

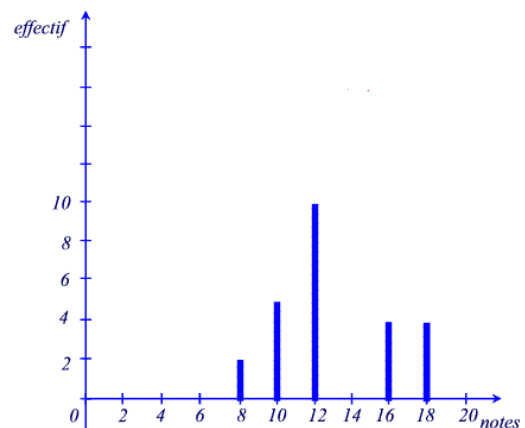
Sur l'axe des ordonnées, on indique les valeurs des effectifs.

Pour chaque modalité, on trace un bâton dont la hauteur indique l'effectif.

Tous les bâtons ont la même largeur, celle-ci n'a aucune signification.

Exemple : Voici le diagramme en bâtonnets des notes obtenues (sur 20) par des élèves d'une classe de 3^e lors d'un contrôle de français.

Notes	Effectifs
8	2
10	5
12	10
16	4
18	4



(2) Caractère quantitatif continu

L'**histogramme** est privilégié pour représenter les séries à caractère quantitatif continu (tableau groupé).

Un histogramme est constitué d'une succession de rectangles accolés dont la largeur de la base correspond à l'amplitude de la classe considérée et dont la surface est proportionnelle à l'effectif associé à cette classe.

a. Si les classes n'ont pas la même largeur :

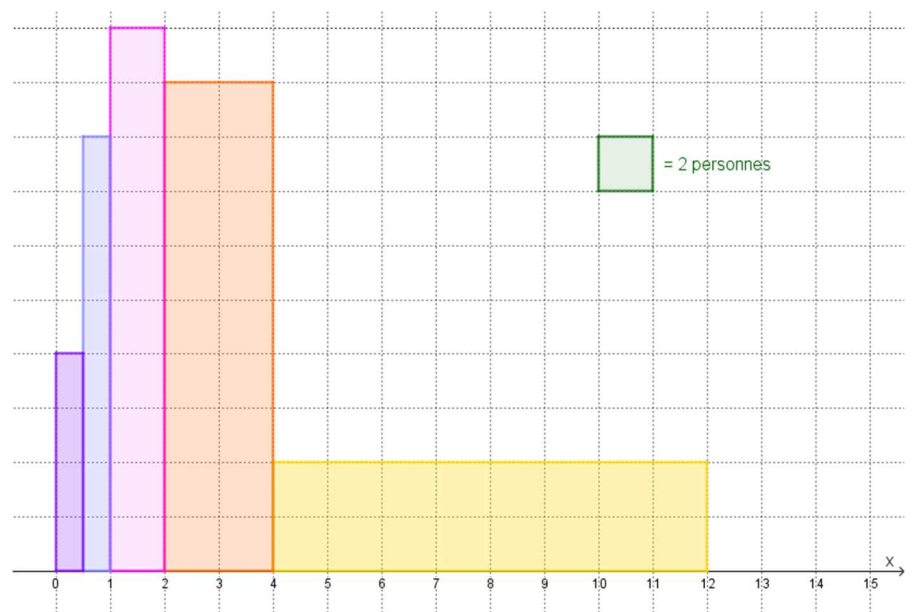
On trace **un seul axe** : celui des abscisses, sur lequel on indique les modalités.

Pour chaque classe, on construit un rectangle dont la base est la largeur de la classe et dont la hauteur est proportionnelle à l'effectif, en fonction de l'échelle choisie.

Il s'agit donc de choisir une échelle adéquate par rapport aux données. On mentionne, sur le graphique, ce que représente un carreau de l'histogramme.

Exemple : On étudie le temps passé devant la télévision en heures par jour en France en 2010.

Classes	Effectifs
[0;0,5[	4
[0,5;1[	8
[1;2[	20
[2;4[	36
[4;12[	32



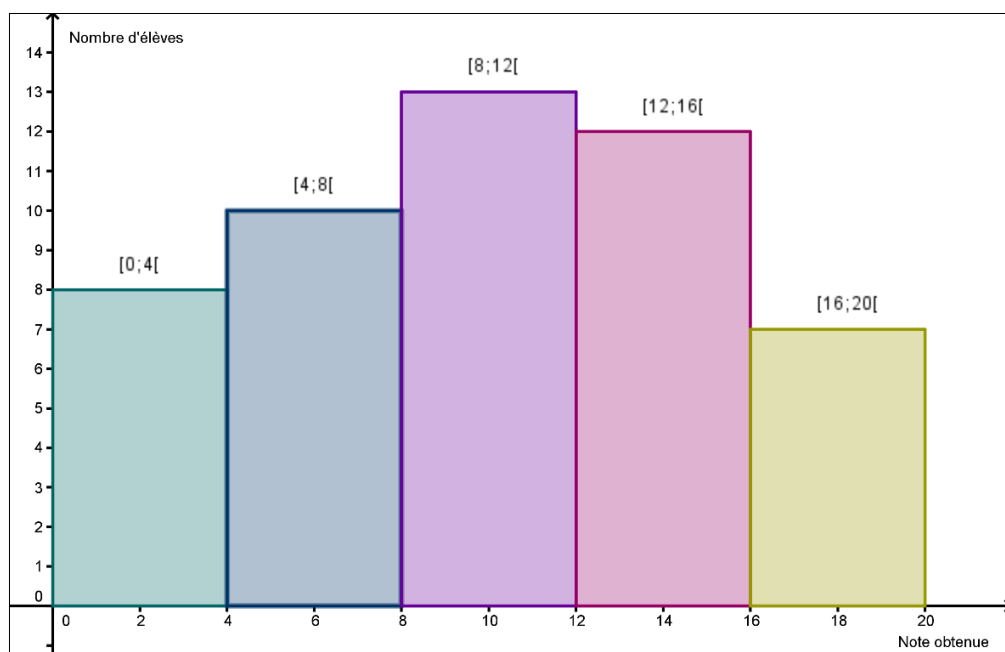
Temps d'utilisation de la télévision en heures par jour

b. Si les classes ont la même largeur :

Si toutes les classes ont la même largeur, les rectangles auront tous la même largeur à la base et leur hauteur sera proportionnelle à l'effectif de la classe. C'est uniquement dans ce cas qu'il sera utile d'inclure un axe des ordonnées gradué et d'y indiquer les effectifs.

Exemple : Voici l'histogramme présentant les notes attribuées aux cinquante élèves ayant présenté l'examen de chimie organique.

Notes	Nombre d'élèves
[0;4[	8
[4;8[	10
[8;12[	13
[12;16[	12
[16;20[	7



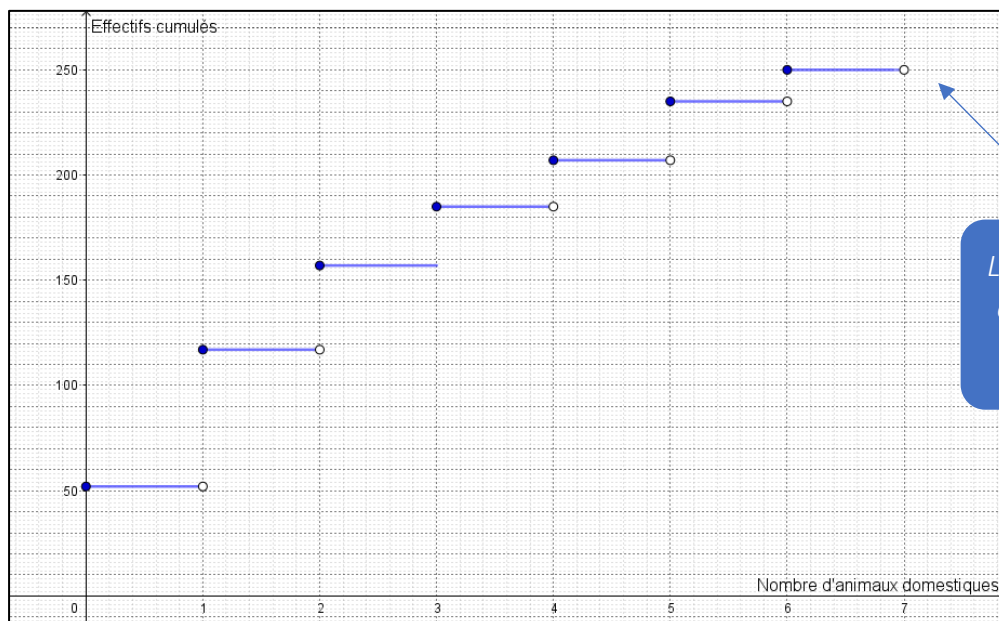
2. Graphique des effectifs cumulés et fréquences cumulées

(1) Caractère quantitatif discret

Le **diagramme des effectifs cumulés** est un graphique en escalier. On construit celui-ci en traçant des paliers qui débutent par un point plein dont l'abscisse est la modalité et l'ordonnée est l'effectif cumulé correspondant. Les paliers se terminent par un point vide dont l'abscisse est la modalité qui suit.

Exemple : Voici le diagramme des effectifs cumulés du nombre d'animaux domestiques de 250 personnes interrogées au hasard dans la rue.

Nombre d'animaux domestiques	Nombre de personnes	Fréquences (en %)	Effectifs cumulés	Fréquences cumulées (en %)
0	52	20,8	52	20,8
1	65	26	117	46,8
2	40	16	157	62,8
3	28	11,2	185	74
4	22	8,8	207	82,8
5	28	11,2	235	94
6	15	6	250	100

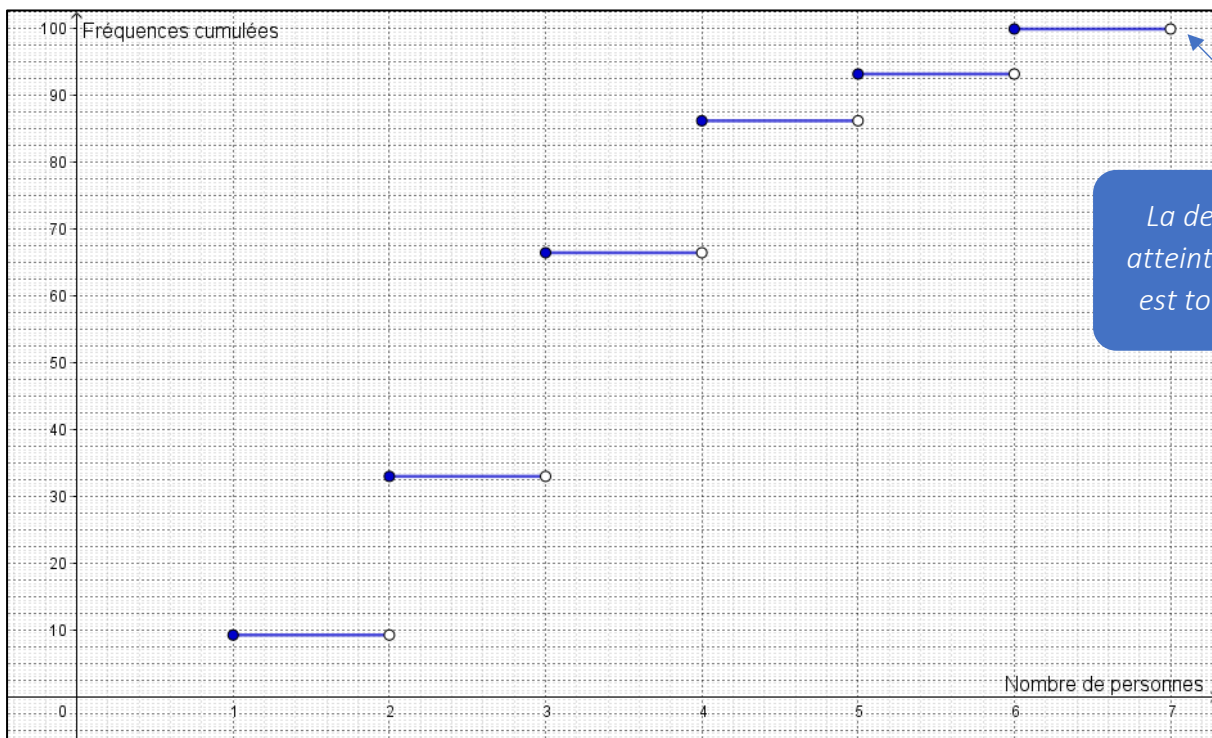


La dernière valeur atteinte en ordonnée est toujours égale à l'effectif total.

Le **diagramme des fréquences cumulées** est également en escalier et se construit de la même manière que celui des effectifs cumulés.

Exemple : On étudie le nombre de personnes à bord d'une voiture.

Nombre de personnes	Nombre de voitures	Fréquences (en %)	Effectifs cumulés	Fréquences cumulées (en %)
1	37	9,25	37	9,25
2	95	23,75	132	33
3	134	33,5	266	66,5
4	79	19,75	345	86,25
5	28	7	373	93,25
6	27	6,75	400	100



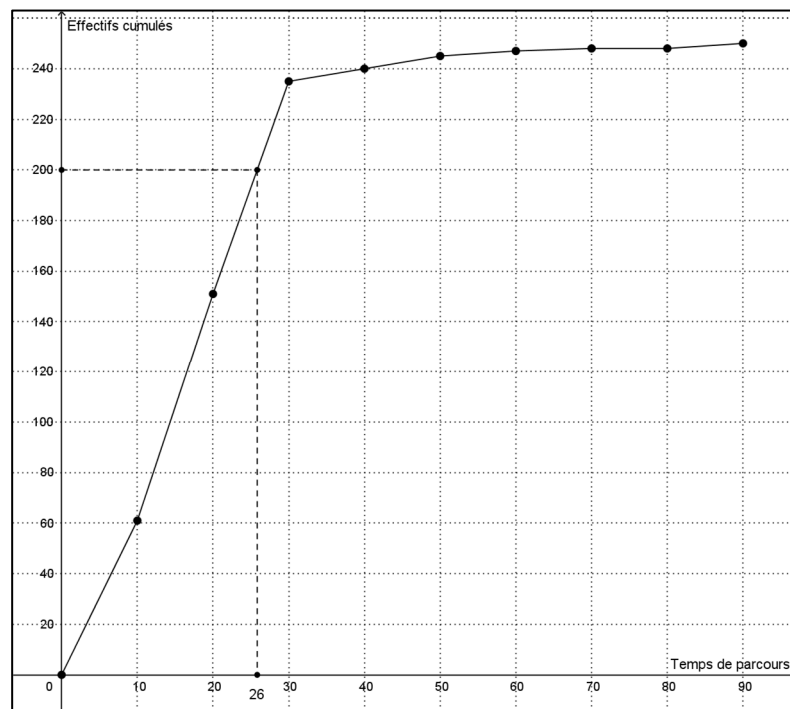
(2) Caractère quantitatif continu

Le **graphique des fréquences cumulées** est une ligne polygonale qui relie les points dont l'abscisse est la plus grande valeur de l'intervalle ouvert décrivant la classe considérée et l'ordonnée l'effectif cumulé de cette classe.

Exemple : Voici le tableau groupé du temps de parcours (en minutes) du domicile à leur collège pour 250 élèves.

Temps (en minutes)	Effectifs	Fréquences (en %)	Effectifs cumulés	Fréquences cumulées (en %)
[0;10[	61	24,4	61	24,4
[10;20[	90	36	151	60,4
[20;30[	84	33,6	235	94
[30;40[	5	2	240	96
[40;50[	5	2	245	98
[50;60[	2	0,8	247	98,8
[60;70[	1	0,4	248	99,2
[70;80[	0	0	248	99,2
[80;90[	2	0,8	250	100

Et le diagramme des effectifs cumulés correspondant :



Remarque : Ce choix de relier les points par des segments de droite revient à considérer que les valeurs du caractère sont régulièrement distribuées à l'intérieur de chaque classe.

Par exemple, on a représenté les points de coordonnées $(30; 235)$ et $(40; 240)$, ce qui signifie qu'il y a 5 $(= 240 - 235)$ temps de parcours dans la classe $[30; 40[$.

Relier ces deux points par un segment de droite revient à faire comme si ces 5 temps étaient répartis régulièrement dans l'intervalle $[30; 40[$: par exemple comme si on avait 30, 32, 34, 36, 38 minutes.

Ce n'est évidemment pas forcément vrai !

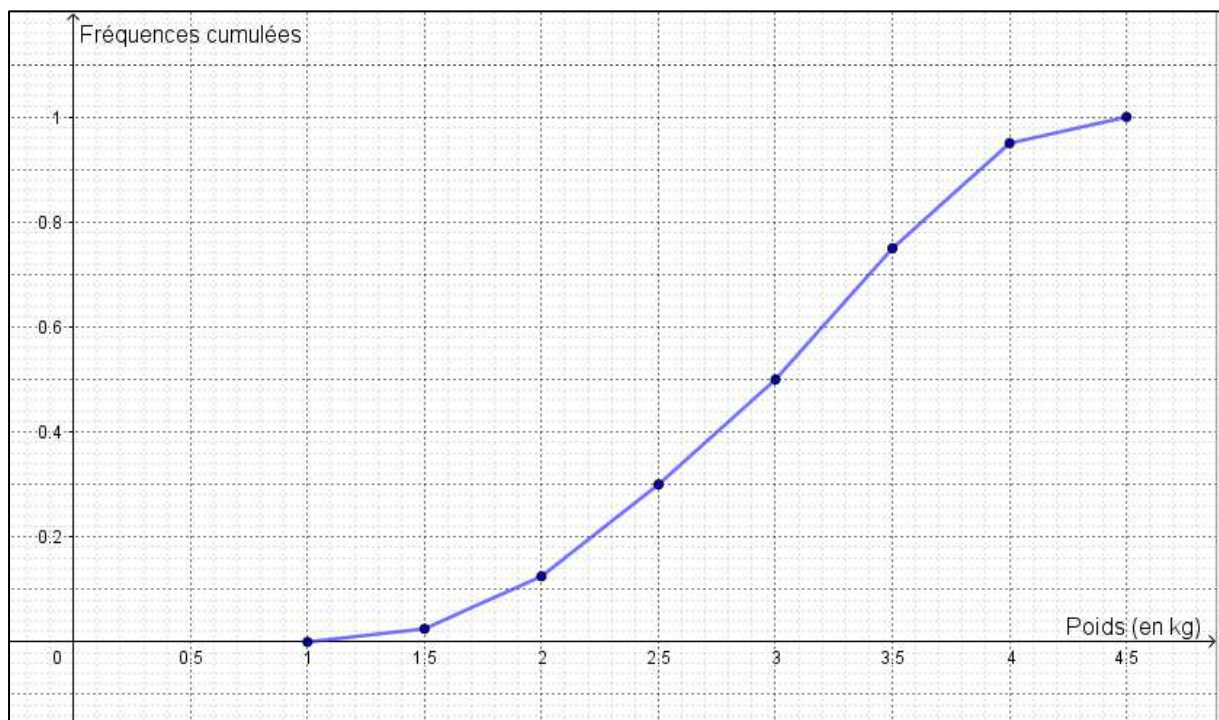
Cette idée de régularité entre chaque point permet de faire des **extrapolations** (c'est-à-dire des déductions à partir de données partielles) sur les valeurs du caractère, bien sûr si l'on n'a pas le tableau détaillé de ces valeurs.

Par exemple, on a **extrapolé**, sur le graphique, le 200^{ème} temps de parcours. On a trouvé qu'il était d'environ 26 minutes.

Le **graphique des fréquences cumulées** se construit de la même manière que celui des effectifs cumulés.

Exemple : Voici le diagramme des fréquences cumulées pour le poids de 240 enfants, à leur naissance.

Poids (en kg)	Nombre d'enfants	Fréquences	Effectifs cumulés	Fréquences cumulées
[1;1,5[	6	0,025	6	0,025
[1,5;2[	24	0,1	30	0,125
[2;2,5[	42	0,175	72	0,3
[2,5;3[	48	0,2	120	0,5
[3;3,5[	60	0,25	180	0,75
[3,5;4[	48	0,2	228	0,95
[4;4,5[	12	0,05	240	1



C. Paramètres de position

Un tableau recensé associé à quelques graphiques permet de visualiser globalement les résultats d'une enquête statistique. Mais on va tirer davantage d'informations à propos de la population observée. Une idée est de regarder la tendance centrale qui est décrite par des paramètres de position.

1. Moyenne arithmétique



B.1. MOYENNE ARITHMÉTIQUE

On a demandé à 34 élèves de 4^{ème} le temps qu'ils passaient chaque jour à surfer sur internet, pendant leurs vacances d'été. Voici le tableau recensé de leurs réponses :



Modalités	Effectifs	Fréquences	Effectifs cumulés	Fréquences cumulées
x_i	e_i	f_i	ε_i	F_i
0	6	0,18	6	0,18
1	12	0,35	18	0,53
2	5	0,15	23	0,68
3	3	0,09	26	0,77
4	2	0,06	28	0,83
5	5	0,15	33	0,98
12	1	0,03	34	1
Total :	34	1		

On voudrait connaître le temps moyen passé sur internet par élève.

Moyenne Somme de toutes les modalités du caractère observé, divisée par l'effectif

arithmétique: total : $\bar{x} = \frac{x_1 \cdot e_1 + x_2 \cdot e_2 + x_3 \cdot e_3 + \dots + x_p \cdot e_p}{n}$

où p désigne le nombre de modalités.

Exemple : $\bar{x} = \frac{0 \times 6 + 1 \times 12 + 2 \times 5 + 3 \times 3 + 4 \times 2 + 5 \times 5 + 12 \times 1}{34} = 2,24$

Cela signifie que, en moyenne, les élèves surfaient 2,24h par jour pendant leurs vacances.

Remarques :

1. Pour calculer la moyenne arithmétique d'un caractère quantitatif continu, on remplace chaque modalité par le centre c_i de chaque classe.

Par exemple, le centre de la classe $[4;8[$ est $\frac{4+8}{2} = 6$.

La moyenne ainsi obtenue n'a pas la même valeur que si on effectue le calcul à partir du tableau brut (sans classe), mais elle constitue une bonne approximation de celle-ci.

Exemple : Dans une ferme, à une date déterminée, on a pesé les œufs qui ont été produits (les masses des œufs sont exprimées en grammes).



Masse de l'œuf	$[28;37[$	$[37;47[$	$[47;52[$	$[52;57[$	$[57;62[$	$[62;72[$	$[72;82[$
Nombre d'œufs	3	51	74	112	92	62	6

On calcule la moyenne :

$$\bar{x} = \frac{32,5 \times 3 + 42 \times 51 + 49,5 \times 74 + 54,5 \times 112 + 59,5 \times 92 + 67 \times 62 + 77 \times 6}{400} = 55,24 \text{ g}$$

2. Calculer la moyenne arithmétique d'un caractère qualitatif n'a aucun sens.
3. La moyenne n'est pas nécessairement une valeur observée du caractère étudié. Par exemple, il n'existerait évidemment pas de famille comprenant 2,67 enfants...
4. La moyenne est influencée par toutes les valeurs et est malheureusement très sensible aux valeurs extrêmes, au point d'en perdre parfois une bonne partie de sa représentativité, surtout dans des échantillons de petite taille. Ainsi la moyenne des six salaires mensuels suivants 3500 ; 4200 ; 4600 ; 5000 ; 6200 ; 36500 est égale à 10 000, alors qu'un seul salaire dépasse cette moyenne.

2. Médiane



B.2. MÉDIANE

Définition : La **médiane** d'une série statistique à caractère quantitatif, notée M , est la modalité qui divise la population en deux groupes de même effectif.

Les modalités du premier groupe sont inférieures à la médiane, les modalités de l'autre sont supérieures à la médiane.

(1) Caractère quantitatif discret

Pour déterminer la médiane d'un caractère quantitatif discret, il suffit de rechercher la modalité dont la fréquence cumulée atteint ou dépasse 50%.

Exemple : L'espérance de vie dans les pays européens en 2015 est donnée dans le tableau ci-dessous.

Espérance de vie (en années)	71	72	73	74	75	76	77	78	79	81	82	83
Effectifs	1	1	1	1	5	2	4	4	1	9	9	5
Fréquences (en %)	2,3	2,3	2,3	2,3	11,6	4,7	9,3	9,3	2,3	20,9	20,9	11,6
Effectifs cumulés	1	2	3	4	9	11	15	19	20	29	38	43
Fréquences cumulées (en %)	2,3	4,6	6,9	9,2	20,8	25,5	34,8	44,1	46,4	67,3	88,2	100

81 est la modalité pour laquelle la fréquence cumulée atteint ou dépasse 50 %, il s'agit donc de la médiane : $M = 81$.

Cela signifie que la moitié des européens a une espérance de vie inférieure à 81 ans.

(2) Caractère quantitatif continu

Pour un caractère quantitatif continu, on utilise le diagramme des fréquences cumulées : une valeur *approchée* de la médiane s'obtient en recherchant l'abscisse du point dont l'ordonnée est 0,5 (50 %).

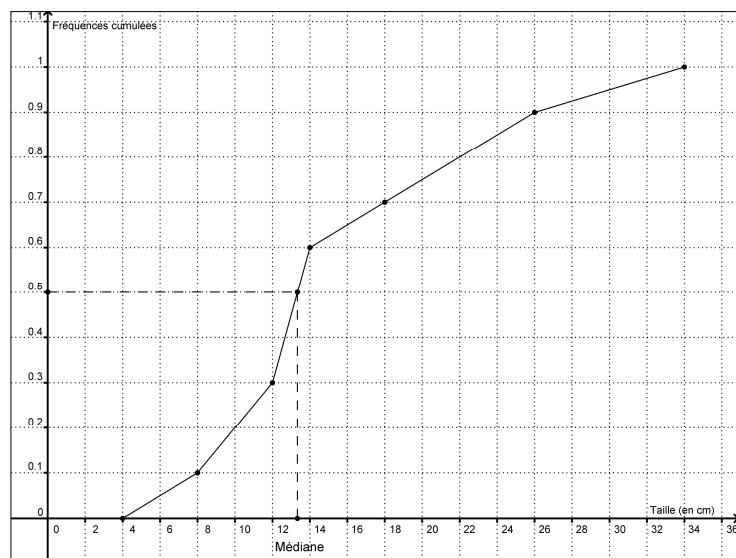
On détermine une valeur *précise* de la médiane en utilisant l'**interpolation linéaire** (méthode consistant à calculer la valeur prise par une fonction passant par deux points) :

- 1 On repère la classe à laquelle la fréquence cumulée 0,5 appartient. Sur le diagramme des fréquences cumulées, ce point appartient au segment reliant les points $A(x_A; F_A)$ et $B(x_B; F_B)$.
- 2 On calcule la valeur de la modalité correspondante avec la formule
$$M = \frac{0,5 - F_A}{F_B - F_A} \cdot L + x_A$$
 où L est l'amplitude de la classe considérée : $L = x_B - x_A$.

Exemple : Un arboriculteur a mesuré la taille des fleurs dans l'une de ses serres.

Voici le tableau groupé et le diagramme des fréquences cumulées :

Classes	Fréquences cumulées (en %)
[4;8[	6
[8;12[	10
[12;14[	60
[14;18[	70
[18;26[	90
[26;34[	100



L'abscisse du point dont l'ordonnée vaut 0,5 (50%) vaut approximativement 13.

On a donc $M = 13$.

Précisons la valeur de la médiane, en utilisant l'interpolation linéaire :

1

La médiane appartient à la classe $[12;14[$, elle se situe donc sur le segment déterminé par les points $A(12;0,3)$ et $B(14;0,6)$.

2

On utilise la formule : $M = \frac{0,5 - F_A}{F_B - F_A} \cdot L + x_A$

$$= \frac{0,5 - 0,3}{0,6 - 0,3} \cdot 2 + 12$$

$$= 13,3$$

Cela signifie que la moitié des fleurs mesurent moins de 13,3 cm.



En 2018 en Belgique, le **salaire mensuel moyen** était de 3489 euros brut alors que le salaire mensuel médian était de 3053 euros brut.

Source : statbel.fgov.be

3. Quartiles



B.3. QUARTILES

Alors que la médiane divise la population en deux groupes de même effectif, les **quartiles** partagent la population en quatre groupes de même effectif.

Le **premier quartile**, noté Q_1 , est la modalité qui partage la population en un premier groupe comprenant un quart des effectifs et un second groupe comprenant les trois autres quarts.

Le **deuxième quartile**, noté Q_2 , est égal à la médiane.

Le **troisième quartile**, noté Q_3 , est la modalité qui partage la population en un premier groupe comprenant trois quarts des effectifs et un second groupe comprenant le quart restant.

(1) Caractère quantitatif discret

Pour déterminer le premier quartile d'un caractère quantitatif discret, il suffit de rechercher la modalité dont la fréquence cumulée atteint ou dépasse 25%.

Pour déterminer le troisième quartile d'un caractère quantitatif discret, il suffit de rechercher la modalité dont la fréquence cumulée atteint ou dépasse 75%.

Exemple : L'espérance de vie dans les pays européens en 2015 est donnée dans le tableau ci-dessous.

Espérance de vie (en années)	71	72	73	74	75	76	77	78	79	81	82	83
Effectifs	1	1	1	1	5	2	4	4	1	9	9	5
Fréquences (en %)	2,3	2,3	2,3	2,3	11,6	4,7	9,3	9,3	2,3	20,9	20,9	11,6
Effectifs cumulés	1	2	3	4	9	11	15	19	20	29	38	43
Fréquences cumulées (en %)	2,3	4,6	6,9	9,2	20,8	25,5	34,8	44,1	46,4	67,3	88,2	100

76 est la modalité pour laquelle la fréquence cumulée atteint ou dépasse 25 %, il s'agit donc du premier quartile : $Q_1 = 76$.

Cela signifie qu'un quart des européens a une espérance de vie inférieure à 76 ans.

82 est la modalité pour laquelle la fréquence cumulée atteint ou dépasse 75 %, il s'agit donc du troisième quartile : $Q_3 = 82$.

Cela signifie que trois quarts des européens ont une espérance de vie inférieure à 82 ans.

(2) Caractère quantitatif continu

Sur le diagramme des fréquences cumulées, une valeur *approchée* du premier quartile s'obtient en recherchant l'abscisse du point dont l'ordonnée est 0,25 (25 %).

On détermine une valeur *précise* du premier quartile en utilisant l'**interpolation linéaire** :

- 1 On repère la classe à laquelle la fréquence cumulée 0,25 appartient. Sur le diagramme des fréquences cumulées, ce point appartient au segment reliant les points $A(x_A; F_A)$ et $B(x_B; F_B)$.
- 2 On calcule la valeur de la modalité correspondante avec la formule
$$Q_1 = \frac{0,25 - F_A}{F_B - F_A} \cdot L + x_A$$
 où L est l'amplitude de la classe considérée.

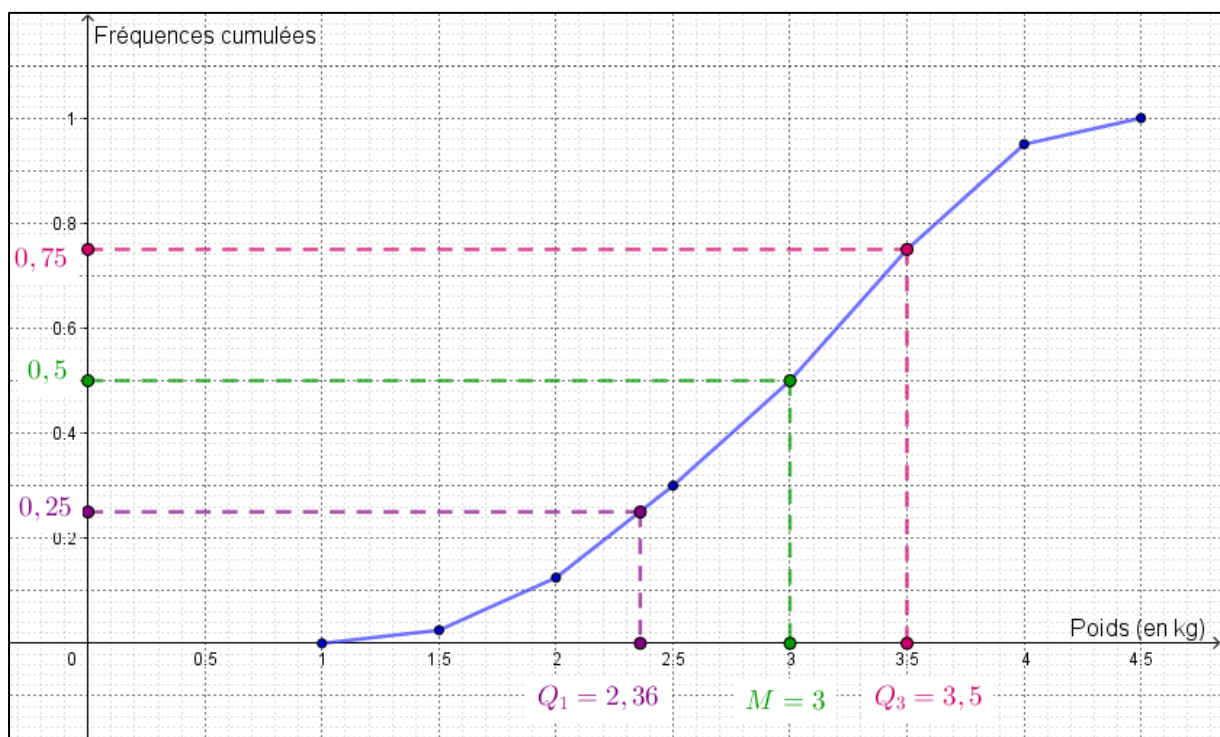
Sur le diagramme des fréquences cumulées, une valeur *approchée* du troisième quartile s'obtient en recherchant l'abscisse du point dont l'ordonnée est 0,75 (75 %).

On détermine une valeur *précise* du premier quartile en utilisant l'**interpolation linéaire** :

- 1 On repère la classe à laquelle la fréquence cumulée 0,75 appartient. Sur le diagramme des fréquences cumulées, ce point appartient au segment reliant les points $A(x_A; F_A)$ et $B(x_B; F_B)$.
- 2 On calcule la valeur de la modalité correspondante avec la formule
$$Q_3 = \frac{0,75 - F_A}{F_B - F_A} \cdot L + x_A$$
 où L est l'amplitude de la classe considérée.

Exemple : Reprenons l'exemple du poids de 240 enfants, à leur naissance.

Poids (en kg)	Nombre d'enfants	Fréquences	Effectifs cumulés	Fréquences cumulées
[1;1,5[	6	0,025	6	0,025
[1,5;2[	24	0,1	30	0,125
[2;2,5[	42	0,175	72	0,3
[2,5;3[	48	0,2	120	0,5
[3;3,5[	60	0,25	180	0,75
[3,5;4[	48	0,2	228	0,95
[4;4,5[	12	0,05	240	1



On peut y lire très facilement la valeur de la médiane et celle du troisième quartile.

Calculons valeur exacte du premier quartile, en utilisant l'interpolation linéaire :

- 1 Grâce au tableau groupé, on sait que le premier quartile appartient à la classe $[2; 2,5[$, il se situe donc sur le segment déterminé par les points $A(2; 0,125)$ et $B(2,5; 0,3)$.

- 2 On utilise la formule :
$$Q_1 = \frac{0,25 - F_A}{F_B - F_A} \cdot L + x_A$$

$$= \frac{0,25 - 0,125}{0,3 - 0,125} \cdot 0,5 + 2$$

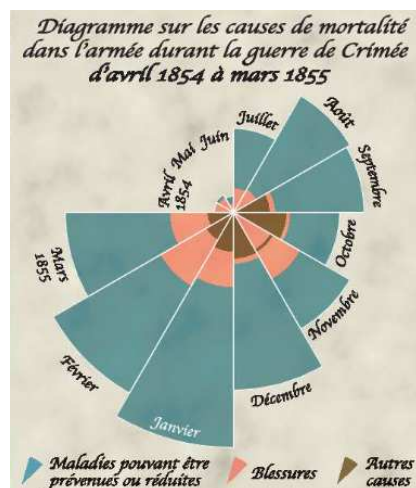
$$= 2,36$$

Cela signifie qu'un quart des nouveau-nés pèse moins de 2,36 kg.



Florence Nightingale (1820-1910), infirmière britannique, organisa une intervention humanitaire sur le front de la guerre de Crimée (1853-1856) où la mortalité était importante. A son retour, elle étudia les données rassemblées sur les causes de la mortalité : cette dernière étant en grande partie due aux conditions sanitaires.

Nightingale milita pour que le gouvernement britannique investisse davantage dans les soins aux soldats. Pour exposer des faits précis tout en rendant le contenu plus clair, elle inventa les diagrammes en crête de coq. En 1859, elle devint la première femme à être admise à la Société Royale de Statistique.



4. Mode

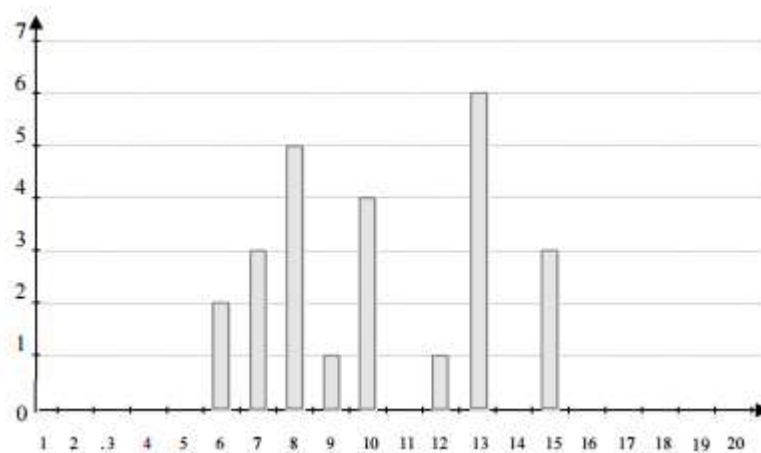
Le **mode** est la modalité pour laquelle l'effectif est le plus élevé.

On parle de **classe modale** dans le cas d'un caractère quantitatif continu.

Contrairement à la moyenne et à la médiane, il peut y avoir plusieurs modes.

Exemples :

- Voici le diagramme représentant la répartition des notes obtenues (sur 20) par les élèves d'une classe de 3^{ème} lors d'un contrôle de français.



Pour cette série statistique, le mode est 13.

- Une station de ski réalise une enquête auprès de 300 skieurs qui la fréquentent. Les résultats de l'enquête sont présentés dans le tableau ci-dessous et indiquent la répartition en classe des skieurs en fonction de leur âge (en années).

Age	[0;10[	[10;20[	[20;30[	[30;40[	[40;50[	[50;60[	[60;70[	[70;80[	[80;90[
Effectifs	27	45	48	39	42	36	33	24	36

Pour cette série statistique, la classe modale est $[20;30[$.

5. Choix du paramètre de position

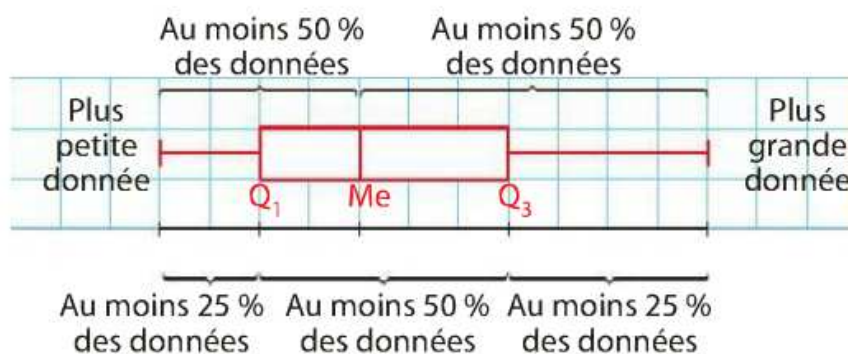
Il n'y a pas de réponse fixe à cette question, chaque paramètre de position présente des avantages et des inconvénients :

Valeurs centrales	Avantages	Inconvénients
Moyenne	• Très largement utilisée. • Utilise toutes les données.	• Influencée par des valeurs extrêmes. • Peut ne correspondre à aucune valeur de la liste.
Médiane	• Est souvent une valeur donnée. • Utile lorsque des valeurs extrêmes peuvent peser sur la moyenne.	• Cible une seule valeur.
Mode	• Très facile à déterminer. • Représente une valeur typique. • Est toujours une valeur de la liste.	• N'inclut pas toutes les valeurs.

6. Boîte à moustaches ou diagramme en boîte

Il est possible de résumer, sous la forme d'un graphique, l'information fournie par l'étendue et les trois quartiles.

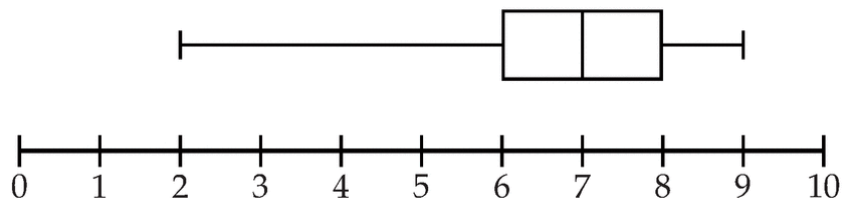
Ce diagramme consiste en une boîte rectangulaire symbolisant l'intervalle interquartile. Celle-ci est prolongée par des segments (les "moustaches") dont les extrémités représentent les valeurs extrêmes de la série statistique. La médiane est marquée par un trait vertical dans la boîte.



Exemple : On considère une série statistique dont le diagramme en boîte est donné ci-dessous.

(1) Lisons :

- le minimum : 2
- Q_1 : 6
- la médiane : 7
- Q_3 : 8
- le maximum : 9



(2) Déduisons-en l'intervalle interquartile : $[6;8]$

D. Paramètres de dispersion

Actuellement, la moyenne reste la valeur la plus utilisée en statistique, mais on s'en sert avec prudence car sa signification et sa portée sont très différentes d'une situation à l'autre. La moyenne ne donne en effet aucune information sur la dispersion des données. C'est pourquoi elle doit être complétée par une valeur rendant compte de cette dispersion et qui conditionne la fiabilité de la moyenne. Cette valeur est basée sur le calcul de l'écart-type.

Il existe d'autres paramètres de dispersion que l'écart-type que nous allons découvrir dans cette section.

1. Etendue

Une première indication intéressante est de connaître la longueur de l'intervalle entre les modalités extrêmes, c'est-à-dire l'étendue.

L'étendue est la différence entre la plus grande et la plus petite modalité.

Lorsque le caractère quantitatif étudié est continu, l'étendue est la différence entre la plus grande valeur de la dernière classe et la plus petite de la première classe.

Exemples :

1. Voici les résultats obtenus à un test :

Modalités	1	2	3	4	5	6	7	8	9	10
Effectifs	4	6	11	25	27	30	23	18	5	1

L'étendue de cette série statistique vaut $10 - 1 = 9$.

2. Voici le tableau de la répartition des salaires mensuels des employés d'une entreprise :

Classes	[600;1000[	[1000 ;1400[	[1400 ;1800[	[1800 ;2200[	[2200 ;2600[
Effectifs	3	21	112	124	52

L'étendue de cette série statistique vaut $2600 - 600 = 2000$.

Comme l'étendue ne tient compte que des 2 valeurs extrêmes, elle risque d'être influencée par une valeur exceptionnelle.

2. Ecart interquartile

Une autre indication intéressante est de connaître la longueur de l'intervalle qui comprend 50 % des observations centrales de la distribution, c'est-à-dire l'écart interquartile.

L'**écart interquartile** est la différence $Q_3 - Q_1$.

L'écart interquartile indique la dispersion des valeurs de la série autour de la médiane : un grand écart interquartile correspond à une grande dispersion ; il a également l'avantage de ne pas être sensible aux valeurs extrêmes.

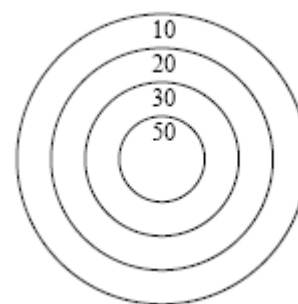
3. Ecart moyen

L'écart moyen est la moyenne des écarts à la moyenne :

$$\frac{(x_1 - \bar{x}) \cdot e_1 + (x_2 - \bar{x}) \cdot e_2 + (x_3 - \bar{x}) \cdot e_3 + \dots + (x_p - \bar{x}) \cdot e_p}{n} \text{ où } p \text{ est le nombre de modalités.}$$

Exemple : Eric joue aux fléchettes en vue d'un championnat. Il tire vingt fléchettes sur la cible.

Les résultats obtenus sont les suivants :



Nombre de points x_i	Nombre de tirs pour Eric
0	1
10	4
20	5
30	6
50	4

En moyenne, Eric obtient $\bar{x} = \frac{0 \times 1 + 10 \times 4 + 20 \times 5 + 30 \times 6 + 50 \times 4}{20} = 26$ points lorsqu'il lance une fléchette.

L'écart moyen vaut : $\frac{(0 - 26) \times 1 + (10 - 26) \times 4 + (20 - 26) \times 5 + (30 - 26) \times 6 + (50 - 26) \times 4}{20} = 0$.

Conclusion : Quelle que soit la situation, l'écart moyen vaut 0 !

4. Ecart moyen absolu

Puisque l'écart moyen ne fournit aucun renseignement utile, on remplace les écarts à la moyenne par leur valeur absolue, c'est-à-dire $|x_i - \bar{x}|$.

L'**écart moyen absolu** est la moyenne des valeurs absolues des écarts à la moyenne :

$$\frac{|x_1 - \bar{x}| \cdot e_1 + |x_2 - \bar{x}| \cdot e_2 + |x_3 - \bar{x}| \cdot e_3 + \dots + |x_p - \bar{x}| \cdot e_p}{n} \quad \text{où } p \text{ est le nombre de modalités.}$$

Exemple : Reprenons l'exemple d'Eric :

L'écart moyen absolu vaut :

$$\frac{|0 - 26| \times 1 + |10 - 26| \times 4 + |20 - 26| \times 5 + |30 - 26| \times 6 + |50 - 26| \times 4}{20} = 12.$$

Cela signifie donc qu'à chaque lancer, Eric obtient un score qui s'écarte de 12 points au-dessus, ou en-dessous de la moyenne.

5. Variance et écart-type

Pour étudier davantage la **dispersion** d'une série, il sera plus avantageux d'accentuer les valeurs extrêmes par rapport à la moyenne. Ainsi, au lieu de calculer les valeurs absolues des écarts, on calcule leurs carrés : $(x_i - \bar{x})^2$.

La **variance** est la moyenne des carrés des écarts à la moyenne de toutes les modalités de la série statistique : $V = \frac{(x_1 - \bar{x})^2 \cdot e_1 + (x_2 - \bar{x})^2 \cdot e_2 + (x_3 - \bar{x})^2 \cdot e_3 + \dots + (x_p - \bar{x})^2 \cdot e_p}{n}$ où p est le nombre de modalités.

Exemple : Reprenons l'exemple d'Eric :

La variance vaut :

$$V = \frac{(0 - 26)^2 \times 1 + (10 - 26)^2 \times 4 + (20 - 26)^2 \times 5 + (30 - 26)^2 \times 6 + (50 - 26)^2 \times 4}{20} = 214$$

Remarque : Dans le cas d'un caractère quantitatif continu, les x_i doivent être remplacés par les centres des classes, comme pour le calcul de la moyenne.

Le principal défaut de la variance est qu'elle a la même dimension que le carré des valeurs des modalités. Pour pallier ce problème, il suffit de prendre la racine carrée positive de la variance.

L'écart-type est la racine carrée de la variance : $\sigma = \sqrt{V}$

Concrètement, l'écart-type donne une certaine mesure de l'écart entre les modalités de la série statistique et la moyenne de celle-ci :

- plus l'écart-type d'une série est petit, plus les modalités sont concentrées autour de la moyenne, donc plus la série est homogène ;
- plus l'écart-type d'une série est grand, plus les modalités de la série sont éloignées de la moyenne, donc moins la série est homogène.

L'écart-type est donc d'un grand intérêt pour comparer deux séries statistiques de même nature : celle dont l'écart-type est le plus élevé est la plus dispersée.

Les paramètres de dispersion utilisés sont donc l'étendue, l'écart interquartile, la variance et l'écart-type.

6. Inégalité de Tchebychev

Il est parfois utile de déterminer quel est l'intervalle de l'échantillon qui contient un pourcentage donné des observations. Ou encore de calculer le pourcentage des observations qui figurent dans un intervalle donné centré sur la moyenne.

Les mathématiciens français Irénée-Jules Bienaymé et russe Pafnouti Tchebychev ont établi l'inégalité suivante :

Inégalité de Tchebychev : Pour une série statistique de moyenne $\bar{x}$ et d'écart-type σ et pour tout réel k strictement positif, la proportion des données qui appartiennent à l'intervalle $[\bar{x} - k\sigma; \bar{x} + k\sigma]$ est supérieur ou égale à $1 - \frac{1}{k^2}$.

Exemple : On a pesé cent castors :

Masse arrondie au kg	Nombre de castors mâles adultes
$[18; 20[$	4
$[20; 22[$	17
$[22; 24[$	32
$[24; 26[$	33
$[26; 28[$	11
$[28; 30[$	3

(1) Calculons la moyenne et l'écart-type.

$$\bar{x} = \frac{19 \times 4 + 21 \times 17 + 23 \times 32 + 25 \times 33 + 27 \times 11 + 29 \times 3}{4 + 17 + 32 + 33 + 11 + 3} = 23,78$$

$$V = \frac{(19 - 23,78)^2 \times 4 + (21 - 23,78)^2 \times 17 + (23 - 23,78)^2 \times 32 + (25 - 23,78)^2 \times 33 + (27 - 23,78)^2 \times 11 + (29 - 23,78)^2 \times 3}{4 + 17 + 32 + 33 + 11 + 3} = 4,87$$

$$\sigma = \sqrt{4,87} = 2,21$$

- (2) Utilisons l'inégalité de Tchebychev pour déterminer l'intervalle qui contient plus de 50 % des effectifs.

On cherche l'intervalle $[\bar{x} - k\sigma; \bar{x} + k\sigma]$ avec k qui vérifie $1 - \frac{1}{k^2} = 0,5$.

On résout cette dernière équation :

$$1 - \frac{1}{k^2} = 0,5 \Leftrightarrow 1 - 0,5 = \frac{1}{k^2}$$

$$\Leftrightarrow 0,5 = \frac{1}{k^2}$$

$$\Leftrightarrow k^2 = \frac{1}{0,5}$$

$$\Leftrightarrow k^2 = 2$$

$$\Leftrightarrow k = \pm\sqrt{2}$$

k étant un réel strictement positif, on écarte la solution $k = -\sqrt{2}$.

Ainsi, l'intervalle est $[23,78 - \sqrt{2} \times 2,21; 23,78 + \sqrt{2} \times 2,21] = [20,7; 26,9]$.