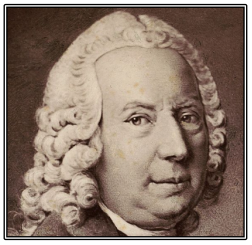
Statistiques

Capacités attendues en fin de chapitre :

- Indicateur de tendance centrale d'une série statistique.
- Linéarité de la moyenne.
- Indicateur de dispersion : écart interquartile, écart-type
- Décrire verbalement les différences entre deux séries statistiques en s'appuyant sur des indicateurs ou des représentations graphiques données.

Le mathématicien du chapitre :



Daniel Bernoulli (1700-1782) est un mathématicien, médecin et physicien suisse. Il met en place le premier modèle statistique qui permet de faire avancer la prévention en épidémiologie. Il montre qu'il est plus avantageux d'inoculer le virus de la variole en prévention afin que les personnes ne meurent plus de cette maladie.

I. *Préambule*

La statistique est une discipline scientifique qui étudie les *caractères* des *individus* d'une *population* donnée.

II. *Vocabulaire des statistiques*

Voici de nombreuses définitions, indispensables pour la compréhension du chapitre :

- ***Population :*** ensemble des individus sur lesquels porte l'étude statistique.

Remarque : les individus ne sont pas nécessairement des êtres humains. Il peut s'agir d'oranges, d'oiseaux…

- ***Caractère :*** propriété étudiée sur chaque individu.
  On l'appelle aussi parfois variable statistique.

✓ Un caractère est ***quantitatif*** lorsqu'il prend des valeurs numériques mesurables (poids, taille, etc…)
✓ Un caractère est ***qualitatif*** lorsqu'il ne prend pas de valeurs numériques (couleur, langue vivante, filière en première, etc…)
✓ Un caractère est ***discret*** lorsqu'il prend des valeurs isolées les unes des autres.
✓ Un caractère est ***continu*** lorsqu'il prend ses valeurs dans un intervalle. On utilise alors le vocabulaire de ***classe*** pour nommer cet intervalle.

Remarque : on a donc plusieurs combinaisons possibles (en utilisant qualitatif, quantitatif, discret et continu) mais certaines n'existent pas…

- ***Modalités :*** valeurs prises par le caractère.
- ***Effectifs :*** L'effectif est le nombre n d'éléments correspondant à une même valeur du caractère.
- ***Effectif total :*** C'est le nombre d'éléments sur lequel porte l'étude.
- ***Fréquence :*** donnée par la formule : $f_i = \frac{n_i}{N}$ elle est souvent exprimée en pourcentage.
- ***Classes :*** une classe est un intervalle noté $[a; b[$ avec $a < b$

✓ L'***amplitude*** de la classe est la longueur de l'intervalle :$b - a$. Dans un exercice de seconde, les classes n'ont pas nécessairement toutes la même amplitude.
✓ Le ***centre de la classe*** est le nombre $\frac{a+b}{2}$

- Les ***séries cumulées*** sont obtenues en ajoutant toutes les valeurs inférieures à une modalité.

- ✓ Les ***Effectifs Cumulées Croissants (ECC)*** répondent à la question : Combien d'éléments ont une valeur inférieure ou égale à une valeur donnée du caractère.
- ✓ Les ***Fréquences Cumulées Croissantes (FCC)*** répondent à la question : Quel est le pourcentage d'éléments qui ont une valeur inférieure ou égale à une valeur donnée du caractère.

Remarque :

Les séries cumulées ne peuvent être employées que ***si le caractère est quantitatif***.

Exemple :

On a mesuré la durée de vie en heures de 300 ampoules. Les résultats sont consignés ci-dessous.

Durée de vie	[900 ; 950[	[950 ; 1000[	[1000 ; 1050[	[1050 ; 1100[
Effectif	75	165	36	24
ECC	75	240	276	300

La population étudiée est les 300 ampoules. Le caractère étudié est la durée de vie d'une ampoule. Ce caractère est quantitatif et continu. C'est pour cela que les données sont regroupées par classe d'amplitude égale à 50 heures. Il y a 240 ampoules qui fonctionnent moins de 1000 heures.

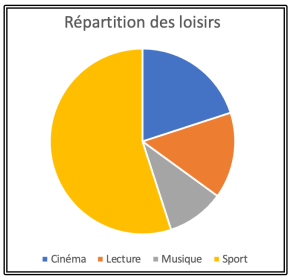
III. Quelques représentations graphiques

a) Le diagramme circulaire

On rappelle qu'un angle plein a pour mesure 360 degrés.
On étudie le budget pour les loisirs consacrés par une famille de 5 personnes (2 parents et 3 enfants).

1. Compléter le tableau ci-dessous.

Loisir	Somme (en euros)	Fréquence (en %)	Angle (en degrés)
Cinéma	40	20	72
Lecture	30	15	54
Musique	20	10	36
Sport	110	55	198
Total	200 €	100	360



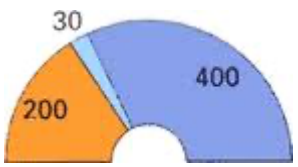
2. Donner le caractère de cette série statistique. Est-il quantitatif discret, quantitatif continu ou qualitatif ?

Le caractère étudié ici est qualitatif car on s'intéresse aux loisirs de la famille.

3. On choisit de construire un diagramme circulaire pour représenter cette série. Après avoir compléter le tableau, vous continuerez la construction du diagramme ci-dessus.

Remarque : On utilise un diagramme semi-circulaire quand une notion d'ordre intervient dans les modalités. C'est le cas, par exemple, lors d'élections.

Exemples ;



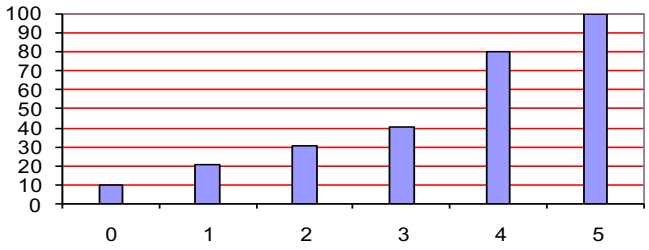
Voici le diagramme semi-circulaire représentant le résultat des élections législatives en France en 1871.Le secteur bleu représente les Bonapartistes et le bleu ciel les royalistes. Quel groupe représente la couleur orange ? Les Républicains

b) *Le diagramme en bâtons*

Exemple : Retrouver les données à partir d'un graphique.

On considère le graphique suivant des fréquences cumulées croissantes (FCC) données en pourcentage.

Compléter le tableau des fréquences en pourcentage cumulées croissantes suivant à l'aide du diagramme ci-contre :



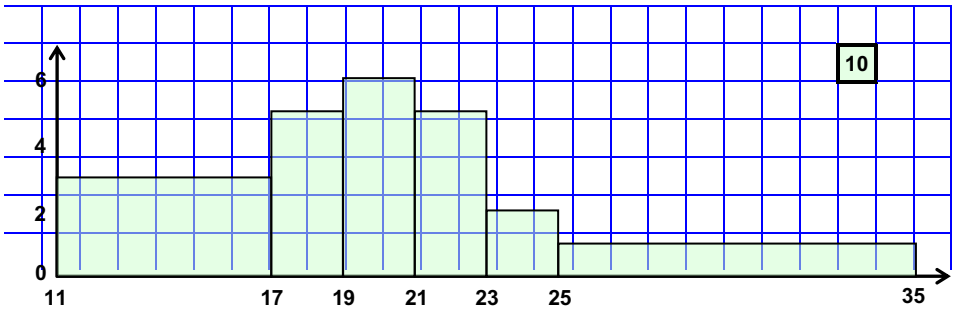
Valeurs	0	1	2	3	4	5
FCC (%)	10	20	30	40	80	100
F (%)	10	10	10	10	40	20

c) *L'histogramme*

Attention :

Ce sont les surfaces des rectangles qui sont proportionnelles aux effectifs et non pas les hauteurs. Il faut être vigilant car les classes n'ont pas la même amplitude.

Dans l'histogramme ci-dessous, l'effectif dans l'intervalle [23 ; 25[est égal à 40.



Compléter le tableau ci-dessous.

Classes	[11 ; 17[	[17 ; 19[	[19 ; 21[	[21 ; 23[	[23 ; 25[	[25 ; 35[	Total
Effectifs	180	100	120	100	40	100	640
Fréquences	0.28	0.16	0.19	0.16	0.06	0.16	1

Remarque : Le caractère est ici quantitatif continu. Nous verrons plus loin comment calculer la moyenne dans ces conditions.

d) *La courbe des FCC*

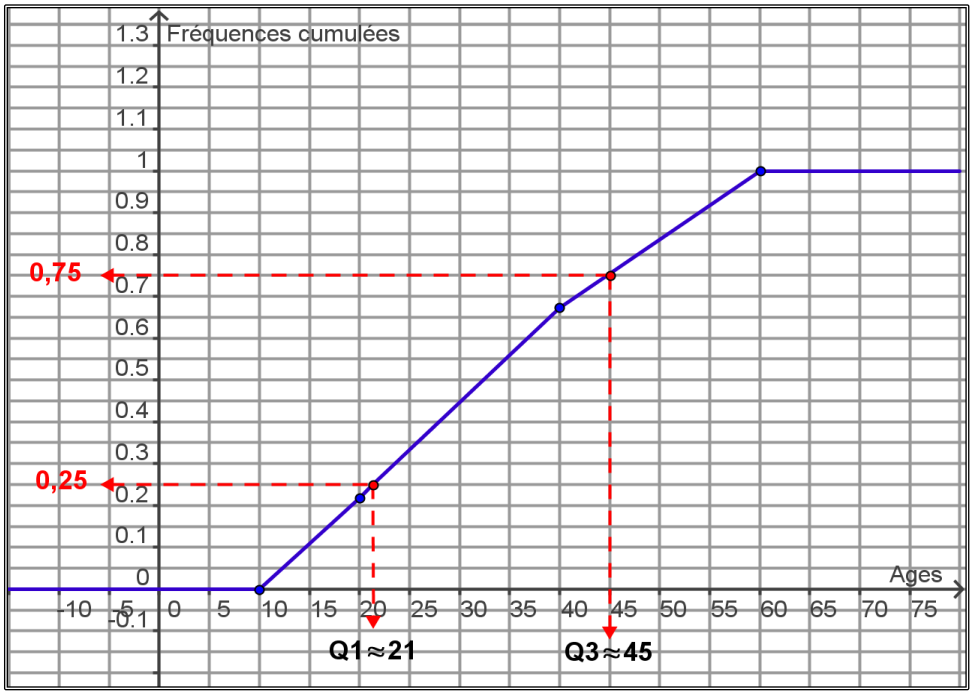
Très fréquemment, en statistiques, on étudie la courbe des fréquences Cumulées Croissantes. Elle est très utile pour retrouver graphiquement des éléments tels que la médiane ou les quartiles.

Exemple :

Voici la fréquentation d'un cinéma en fonction de l'âge. Compléter alors la ligne des fréquences puis celle des ***FCC*** puis tracer la courbe des fréquences cumulées croissantes.

âge	**[10 ; 20[**	**[20 ; 30[**	**[30 ; 40[**	**[40 ; 60[**	Total
Effectif	87	95	87	131	400
Fréquence	0.2175	0.2375	0.2175	0.3275	1
FCC	0.2175	0.455	0.6725	1	

Voici alors le tracé des ***FCC***.



On peut retrouver sur le graphique la médiane et les quartiles.

4. ***Les caractères statistiques***

a) ***La moyenne***

- Moyenne simple.

Définition : La moyenne simple d'une série de valeurs est la somme des valeurs divisée par l'effectif total de la série : $\bar{x} = \frac{x_1+x_2+x_3+ \ldots+x_p}{N}$

Exemples :

On donne le poids des 8 avants d'une mêlée d'une équipe de rugby :
92kg, 102kg, 98kg, 113kg, 118kg, 97kg, 105kg, 101kg,
Quel est le poids moyen d'un avant dans cette équipe ?
Solution : Il suffit d'ajouter tous les poids et de diviser la somme par 8. $m = 103{,}25$ kg

- Moyenne pondérée

Définition : La moyenne pondérée d'une série de valeurs affectées de coefficients est donnée par la formule : $\bar{x} = \frac{n_1\times x_1+n_2\times x_2+n_3\times x_3+ \ldots+n_p\times x_p}{N}$ notée aussi $\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i \times x_i$

Exemples :

De manière classique, les notes d'un élève sont regroupées dans le tableau ci-dessous.
Calculer sa moyenne. On obtient $m = 12{,}8$

Notes	9	11	17	13
Coefficients	1	3	2	4

- Moyenne approchée à l'aide des classes.

Lorsque les valeurs d'une série statistique sont regroupées par intervalles, les formules de la proposition précédente restent valables mais x_i représente alors le centre de l'intervalle.

On se souviendra que le centre d'un intervalle d'extrémités a et b s'obtient en calculant $\frac{a+b}{2}$.

Exemple : On reprend nos ampoules de l'exemple de la première partie.

Durée de vie	[900 ; 950[	[950 ; 1000[	[1000 ; 1050[	[1050 ; 1100[	TOTAL
Effectif	75	165	36	24	300
Centre de l'intervalle	925	975	1025	1075	X
Produit	69375	160875	36900	25800	292950

La durée de vie moyenne d'une ampoule est : $m = 976{,}5$ heures

Remarques :

Résumer une série statistique par sa moyenne fait perdre des informations. On ne sait pas notamment combien d'individus sont situés au dessus de la moyenne et combien sont situés en dessous.
La moyenne obtenue lorsque les données sont regroupées en classe est une moyenne approchée. Ce procédé est utilisé lorsque l'effectif total est très grand afin de minimiser l'erreur.
Pour contourner ce problème, on dispose d'autres indicateurs que nous allons étudier.

Remarques :

- La moyenne d'une série est toujours comprise entre la plus basse valeur et la plus haute valeur des modalités.
- La moyenne d'une série statistique n'est pas nécessairement une valeur des modalités.
- La moyenne d'une série statistique n'est pas la moyenne des modalités extrêmes.

Vocabulaire : La moyenne est une caractéristique de position. Elle donne une indication sur la position précise des valeurs d'une série.

b) L'étendue

Définition : L'étendue d'une série statistique est la différence entre la plus grande et la plus petite des valeurs observées du caractère. (Attention à l'ordre croissant)

Remarque : L'étendue est une caractéristique de dispersion.

c) La médiane

Définition : Les valeurs du caractère étant rangées par ordre croissant (ou décroissant), la médiane d'une série statistique est la valeur du caractère qui partage la série en deux parties de même effectif.

Exemple :

L'organisateur d'une compétition de judo souhaite répartir les combattants en 2 poules contenant le même nombre (ou presque) de participants. La 1ère poule est dite des « légers » et la 2nde est dite des « lourds ».
Pour cela on a donc relevé le poids en kg des judokas :

62 / 98 / 78 / 95 / 68 / 59 / 74 / 81 / 102 / 71 / 80 / 61 / 65 / 72 / 65

1) Quel est le poids qui permet de séparer les judokas en deux catégories ?
Après avoir mis les poids dans l'ordre croissant, le poids médian est ici de 72 kg
2) Le judoka le plus lourd décide de se retirer. Quel est le nouveau poids médian ?
Entre 71 et 72 kg par exemple 71,18

Remarque :

En d'autres termes, la médiane est la valeur à partir de laquelle :

- L'effectif cumulé devient supérieur ou égal à la ***moitié*** de l'effectif total.
- La fréquence cumulée devient supérieure ou égale à 50%.

Exercice :

Retrouver à l'aide de la courbe des ***FCC*** la médiane.
On trouve par lecture graphique : $Med = 32$.

Exercices :

On considère les notes d'une classe de seconde.

Note	7	8	9	10	11	**12**	13	14	15	17	Total
Effectifs	3	1	4	2	3	2	4	4	3	1	27
Effectifs cumulés	3	4	8	10	13	15	19	23	26	27	

Quelle est la médiane ?
27 = 13 + 1 +13 ; on a donc 13 notes : 7, 7, …, 11, 11 inférieures à la médiane **12** et on a 13 notes supérieures : 12, 13, …, 17.
La 14ème note de la série classée par ordre croissant est la note médiane.

Remarque :

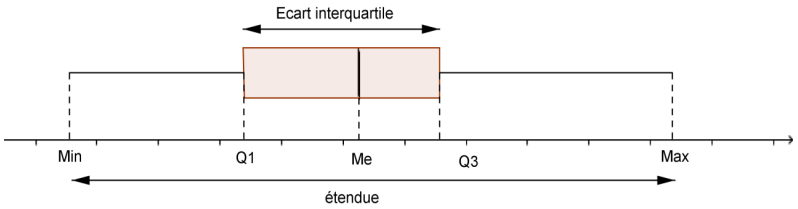
Ici, la médiane appartient à la série, ce n'est pas toujours le cas. Quand la série a un effectif total pair, comme pour les filles, on prend, en général, la moyenne des deux valeurs médianes (centrales). La médiane, comme la moyenne, est ***une caractéristique de position*** de la série.

d) Les quartiles

Définition :
Les quartiles d'une série statistiques sont les valeurs théoriques qui séparent la série en quatre groupes de même effectif.
Ainsi, Le premier quartile, noté Q_1, est la plus petite valeur de la série tel que 25% des données soient inférieures ou égales à Q_1.
Le troisième quartile, noté Q_3, est la plus petite valeur de la série tel que 75% des données soient inférieures ou égales à Q_3.

Remarques :

- Contrairement à la médiane, les quartiles sont toujours des valeurs de la série.
- La connaissance des *ECC* est d'une aide précieuse dans la recherche rapide des quartiles.



Vocabulaire :

L'intervalle $[Q_1; Q_3]$est appelé ***écart interquartile***. Il contient au moins 50 % des valeurs de la série.

Remarques :

- L'écart interquartile mesure la dispersion des valeurs autour de la médiane ; plus l'écart est petit, plus les valeurs de la série appartenant à l'intervalle interquartile sont concentrées autour de la médiane.
- Contrairement à l'***étendue*** qui mesure l'écart entre la plus grande et la plus petite valeur, l'écart interquartile élimine les valeurs extrêmes qui peuvent être douteuses, cependant il ne tient compte que de 50% de l'effectif …
- On peut correctement résumer une série statistique par le couple **(*médiane* ; *intervalle interquartile)***
- La représentation ci-dessus est un résumé statistique de la série appelée diagramme en boite et moustache ou diagramme de Tukey.

Exemple 1 :

On reprend la liste des poids des judokas. Donner les Quartiles.
62 / 98 / 78 / 95 / 68 / 59 / 74 / 81 / 102 / 71 / 80 / 61 / 65 / 72 / 65

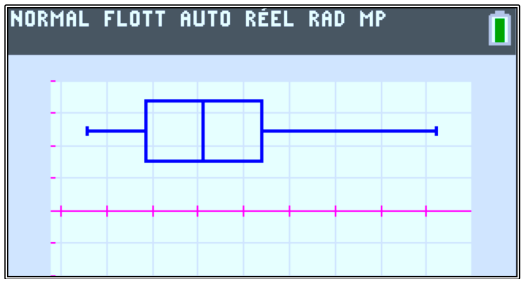
On obtient : $Q_1 = 65$ kg et $Q_3 = 81$ kg

Exemple 2 :

On donne les longueurs en mètres des lancers de javelots d'une classe.

Longueurs en m	37	39	40	41	42	43	46	48	49	TOTAL
Effectif	4	3	4	3	2	5	3	1	1	26
ECC	4	7	11	14	16	21	24	25	26	

1) Quel est l'effectif total de la série ? il y a eu 26 lancers
2) Où se situe la médiane ? La médiane est la demi-somme des 13ème et 14ème valeur soit $Med = 41$ m
3) Quelle est la valeur correspondant à Q_1 ? 26 / 4 = 6,5 Q_1 correspond donc à la 7ème valeur de la série soit donc $Q_1 = 39$
4) Quelle est la valeur correspondant à Q_3 ? 3x 26 / 4 = 19,5. Q_3 correspond donc à la 20ème valeur de la série soit donc $Q_3 = 43$
5) Calculer l'étendue de la série. L'étendue est 49 – 37 = 12 m.
6) Représenter graphiquement les quartiles.



e) *La variance et l'écart-type*

Définition :

La ***Variance*** d'une série statistique est la moyenne des carrés des écarts à la moyenne.
En d'autres termes, on a : $V(x) = \frac{1}{N}\sum_{i=1}^{n} n_i \times (x_i - \bar{x})^2$
L'écart-type est la racine carrée de la variance.

Remarques :

- La variance est un nombre toujours positif.
- La formule $V(x) = \frac{1}{N}\sum_{i=1}^{n} n_i \times {x_i}^2 - \bar{x}^2$ est plus facilement utilisable dans le tableau. Cette formule est appelée formule de Koenig
- Plus l'écart-type est grand, plus les valeurs de la série sont dispersées autour de $\bar{x}$
- L'écart-type a la même unité que les valeurs de la série.
- On peut résumer la série statistique par le couple (moyenne, écart-type). Il utilise toutes les valeurs de la série mais est sensible aux valeurs extrêmes.