Année 2014-2015 - Première ES
Cité scolaire Paul Valéry
Mathématiques - F. Gaunard
http://frederic.gaunard.com

CHAPITRE 4 : STATISTIQUES

L'objectif de ce chapitre est d'apprendre à présenter, synthétiser et analyser des données statistiques.

1. Série Statistique : Rappels

1.1. **Vocabulaire.** Nous reprenons, de manière succinte le vocabulaire dont nous aurons besoin et qui a déjà été introduit précédemment dans votre cursus, notamment en Seconde.

Une *population* est un ensemble d'*individus*, qui ne sont pas nécessairement des personnes mais un élément dont on va étudier une caractéristique (taille, couleur, âge, note...).

Cette caractéristique s'appelle un *caractère*. Les *valeurs* de ce caractère peuvent être différentes d'un individu à un autre. Un caractère peut être *quantitatif* (lorsque ses valeurs sont numériques) ou *qualitatif* (lorsque ce n'est pas le cas; la couleur par exemple).

L'étude statistique que nous présentons dans ce chapitre se concentre sur les caractères quantitatifs d'une population.

Lorsque l'*effectif* d'une population est trop important, on étudit ses caractères à partir d'un *échantillon* représentatif de cette population. Par exemple, pour un sondage, on ne va pas interroger l'ensemble des Français mais un certain nombre, ou encore si l'on veut tester le bon calibrage des machines d'une usine, on va relever seulement un échantillon de la production et pas tout ce qui est produit.

On appelle donc *série statistique* une liste de nombre réels qui sont les valeurs d'un caractère pour chacun des échantillon composant l'échantillon. Le nombre d'individu dans l'échantillon s'appelle l'effectif total et est souvent noté n.

1.2. **Tableaux d'effectifs.** Pour présenter de façon pratique et lisible une telle série, on utilise un *tableau d'effectifs*. On compte le nombre d'apparations n_i de chaque valeur de la liste de manière à identifier p valeurs distinctes que l'on note $x_1, x_2, ..., x_p$ en les classant par ordre croissant le plus souvent. Puis, on affiche cela dans un tableau:

Valeurs (x_i)	x_1	x_2	...	x_p
Effectif (n_i)	n_1	n_2	...	n_p

L'effectif total de la série est naturellement égal à la somme de tous les effectifs, ce qui s'écrit comme cela:

$$n = n_1 + n_2 + ... + n_p = \sum_{i=1}^{p} n_i$$

Exemple 1.1. La série statistique suivante, que nous garderons pour plusieurs exemples au long du chapitre, représente les notes à un contrôle des 20 élèves d'une classe:

13; 10; 8; 15; 8; 13; 18; 15; 9; 10; 10; 15; 13; 13; 13; 15; 10; 18; 15; 15

Il est clair qu'il est nettement plus agréable de présenter cette liste sous la forme:

Note	8	9	10	13	15	18
Effectif	2	1	4	5	6	2

On vérifie bien que $20 = n = 2 + 1 + 4 + 5 + 6 + 2$.

On peut rajouter au tableau précédent une ligne *effectifs cumulés croissants* (souvent notés E.C.C.) dans laquelle on affichera sous chaque valeur non pas le nombre d'individus de la série statistique dont le caractère a cette valeur mais le nombre d'individu avec un caractère ayant une valeur *inférieure ou égale*.

Exemple 1.2. En reprenant la série statistique précédente, on rajoute la ligne des effectifs cumulés croissants, en faisant la somme avec la valeur de la case précédente:

Note	8	9	10	13	15	18
Effectif	2	1	4	5	6	2
Effectifs cumulés croissants	2	3	7	12	18	20

On constate que le dernier effectif cumulé croissant est *toujours égal à* l'effectif total.
On peut facilement lire de ce tableau que 12 élèves n'ont pas plus de 13 ou que 3 ont au maximum 9.

On rappelle que l'on peut également s'intéresser aux *fréquences* plutôt qu'aux effectifs: la fréquence d'apparition de chaque caractère est égale au quotient de son effectif par l'effectif total (affiché parfois sous forme décimale ou en pourcentage) et également aux fréquences cumulées croissantes. Nous aurons sûrement l'occasion de recontrer cela dans des exercices.

1.3. **Représentations graphiques.** A partir des tableaux précédents, on pourra décider d'afficher plusieurs types de représentations graphiques. On pourra par exemple dresser le diagramme à bâtons des effectifs ou la courbe des effectifs cumulés croissants.

Exemple 1.3. En revant à l'exemple des notes, cela donne les graphiques suivants (construits avec *GeoGebra*):

Figure 1. Diagramme à bâtons des effectifs

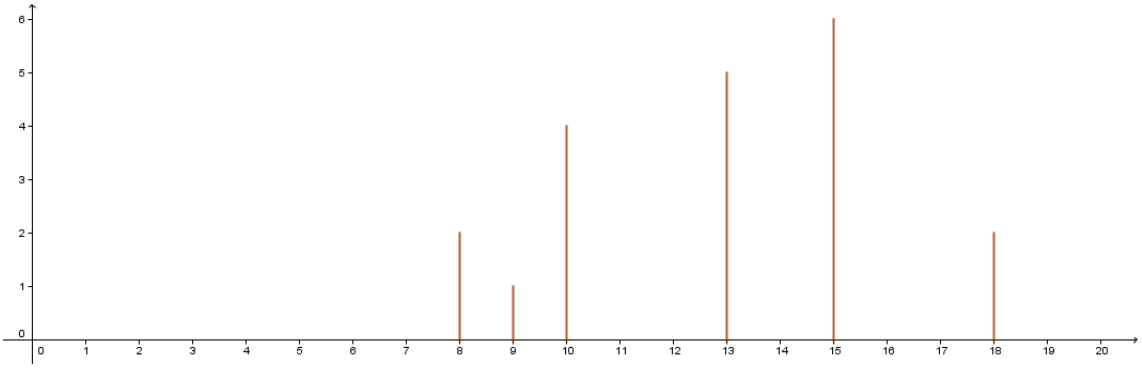
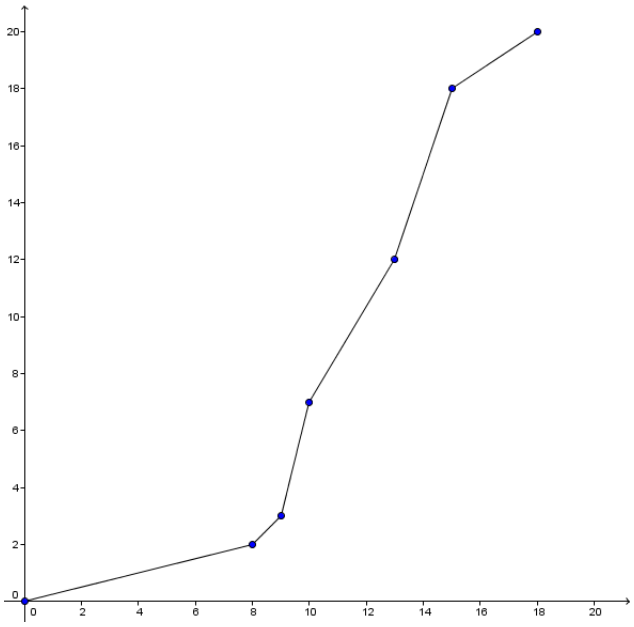


Figure 2. Courbe des effectifs cumulés croissants



2. Moyenne & Ecart-type

2.1. **Moyenne.** La *moyenne* est une mesure statistique caractérisant les éléments d'un ensemble de quantités: elle exprime la grandeur qu'aurait chacun des membres de l'ensemble s'ils étaient tous identiques sans changer la dimension globale de l'ensemble.

Pour la calculer, il faut évidemment faire attention **à tenir compte du nombre de fois que chaque valeur apparaît**. On note traditionnellement la moyenne $\overline{x}$, mais il peut arriver qu'elle soit aussi notée μ. Partant d'un tableau d'effectif, la formule est la suivante

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n_1 + n_2 + ... + n_p} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{n} = \frac{1}{n}\sum_{i=1}^{p} n_ix_i.$$

Exemple 2.1. Toujours dans l'exemple des notes, on a

$$\overline{x} = \frac{2\times 8 + 9 + 4\times 10 + 5\times 13 + 6\times 15 + 2\times 18}{20} = 12,8.$$

2.2. **Variance et Ecart-type.** Connaissant la moyenne d'une série statistique, on peut s'intéresser à la *dispersion des valeurs* autour de cette moyenne. En effet, considérons l'exemple näif suivant:

Exemple 2.2. Deux groupes de 4 élèves participent à un test. Leur résultats, exprimés comme des notes sur 20, sont consignés dans les tableaux suivants :

Groupe 1:

Note	9	11
Effectif	2	2

Groupe 2:

Note	1	19
Effectif	2	2

On constate alors que les moyennes des deux groupes sont toutes deux égales à 10, alors que les résultats dans chacun des deux groupes sont répartis de manière totalement différentes.

On introduit donc la *variance*. La variance est la moyenne des écarts, élevés au carré, de chaque valeur à la moyenne. On la note V et elle est donnée par la formule suivante

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

Ayant calculé la variance, on peut calculer l'*écart-type*, noté par la lettre grecque σ, qui est la racine carée de la variance et s'exprime donc selon la même unité que les valeurs de la série statistique.

$$\sigma = \sqrt{V}.$$

Exemple 2.3. Dans l'exemple des deux groupes précédents, on a, d'abord pour le premier groupe,

$$V_1 = \frac{2\times(9-10)^2 + 2\times(11-10)^2}{4} = \frac{4}{4} = 1$$

et donc

$$\sigma_1 = \sqrt{1} = 1$$

c'est à dire que l'écart-type à la moyenne est de 1, ce qu'on pouvait observer. Dans le cas du second groupe,

$$V_2 = \frac{2\times(1-10)^2 + 2\times(19-10)^2}{4} = \frac{324}{4} = 81$$

et donc

$$\sigma_2 = \sqrt{81} = 9,$$

ce qu'on aurait pu deviné aussi.

Exemple 2.4. Dans le tout premier exemple des notes de la classe de 20 élèves, le calcul donne (arrondi au centième)

$$V = 8,76 \qquad \text{et} \qquad \sigma = 2,96$$

Remarque 2.5. En pratique, même s'il est impératif de connaître la formule et la définition, on utilise le mode STATS de la **calculatrice** pour calculer variance et écart-type.

Pensez à bien prendre en compte les effectifs en vérifiant par exemple que n est bien égal à l'effectif total dans l'écran des résultats. Sinon, la calculatrice affecte automatiquement l'effectif 1 à chaque valeur, ce qui faussera naturellement tous les résultats.

Remarque 2.6. Une variance importante traduit une grande **dispersion** des données alors qu'une petite variance (et par conséquent un petit écart-type) traduit une bonne régularité des résultats.

Exercice 2.7. Durant le quinquennat d'un président de la République, on mesure tous les six mois par un sondage sa cote de popularité en interrogeant toujours le même nombre d'individus. Les résultats sont présentés dans le tableau suivant.

Sondage	1	2	3	4	5	6	7	8	9	10
Cote du popularité (en %)	52	53	49	38	35	31	29	42	45	49

Calculer la moyenne et l'écart-type de la cote de popularité de ce président au cours de ce quinquennat.

Exercice 2.8. Calculer la moyenne, la variance et l'écart-type de chacune des séries dans les cas particuliers suivants:

a) La série comporte 31 termes idéntiques égaux à 15.

b) La série comporte 20 termes égaux à 32 et 20 termes égaux à 48.

Exercice 2.9. Un artisan verrier fabrique une paque de verre coloré pour la réalisation d'un vitrail. L'épaisseur idéale pour son travail est de 2,1 mm. La série suivante donne l'épaisseur de cette plaque, en mm, mesurée en différents points.

$$2,2;\ 2,5;\ 2,1;\ 1,9;\ 2,3;\ 2,2;\ 1,8;\ 2,5;\ 1,8;\ 1,7$$

a) Calculer la moyenne de la série. Le verrier peut-il être satisfait de son travail?

b) Donner une valeur approchée au centième de la variance et de l'écart-type de la série. Quelles sont les unités de ces résultats ?

c) Quelle indication cela apporte-t-il sur la qualité de son travail ?

Exercice 2.10. Une entreprise, qui produit du chocolat, fabrique des tablettes de 100 grammes. Au début de l'année 2010, elle décide de prélever un échantillon dans sa production afin d'en vérifier la masse. Les résultats sont consignés dans le tableau ci-dessous:

Masse (en grammes)	96	97	98	99	100	101	102	103
Effectifs	5	6	9	13	32	16	5	4

On suppose qu'il n'y a pas de problème particulier si au moins 95% des valeurs sont dans l'intervalle $[\overline{x}-2\sigma;\overline{x}+2\sigma]$. Que peut-on dire de cet échantillon ?

Exercice 2.11. (Cet exercice nécessite l'utilisation des notions du chapitre *Second Degré*) On considère une série statistique de 3 nombres: $3a$, $(-a-1)$ et $(a+4)$, où a est un réel. Calculer alors la moyenne et la variance de cette série, en fonction de a. Peut-on avoir une variance égale à 4 ? Même question avec $\frac{14}{3}$.

3. Médiane, Quartiles et Position

Il apparaît qu'il est intéressant de pouvoir représenter la *répartition* des valeurs d'une série statistiques. Pour ce faire, nous introduisons plusieurs valeurs caractéristiques importantes.

L'*étendue* de la série est la différence entre la valeur maximale de cette série et la valeur minimale

$$e = v_{\max} - v_{\min}.$$

3.1. Médiane.

La *médiane*, notée M, est une valeur qui permet de "couper" l'ensemble des valeurs en deux parties égales : mettant d'un côté une moitié des valeurs, qui sont toutes inférieures ou égales à M et de l'autre côté l'autre moitié des valeurs, qui sont toutes supérieures ou égales à M.

En pratique, son calcul dépend de la *parité* de l'effectif total. En effet, si n est **impair**, la médiane sera alors la **valeur centrale** de la série.

Si n est **pair**, on prendra la **moyenne des deux valeurs centrales** de la série.

Exemple 3.1. On relève le prix des baguettes de pain dans diverses boulangeries du quartier

Prix (en euros)	0,85	0,90	0,95	1	1,05	1,10	1,20
Effectif	3	5	2	3	2	1	1

L'effectif total est $3+5+2+3+2+1+1=17$ qui est impair. La valeur centrale est la 9ème valeur de la série (8 valeurs inférieures, 8 valeurs supérieures) et elle vaut 0,95. Ainsi, $M = 0,95$.

Exemple 3.2. On revient au tout premier exemple des notes dans la classe de 20 élèves. L'effectif étant pair, on prendra la moyenne de la 10ème valeur et de la 11ème valeur.

$$M = \frac{13+13}{2} = 13.$$

Dans cet exemple, les deux valeurs centrales sont les mêmes mais ce n'est pas toujours le cas du tout!

3.2. **Quartiles.** Un *quartile* est chacune des trois valeurs qui divisent les données triées en quatre parts égales, de sorte que chaque partie représente un quart (25%) de l'échantillon de population.

Le premier quartile, noté Q_1, est la plus petite valeur de la série telle qu'au moins un quart (25%) des valeurs de la série lui soient inférieures ou égales.

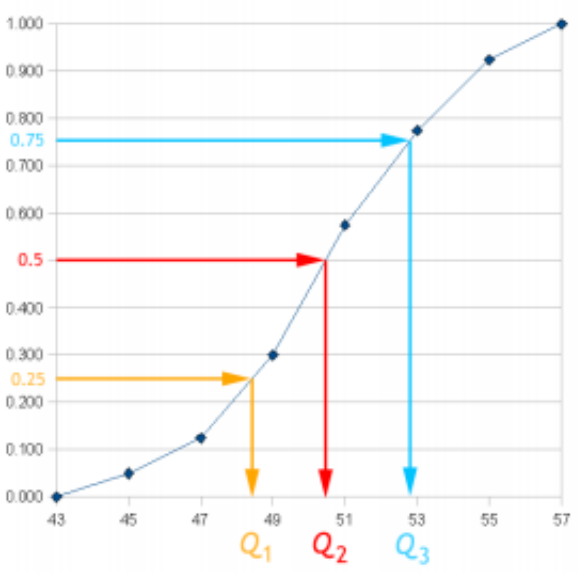
Le troisième quartile, noté Q_3, est la plus petite valeur de la série telle qu'au moins trois quarts (75%) des valeurs de la série lui soient inférieures ou égales.

La différence $Q_3 - Q_1$ s'appelle *écart interquartile*, c'est un critère de dispersion de la série.

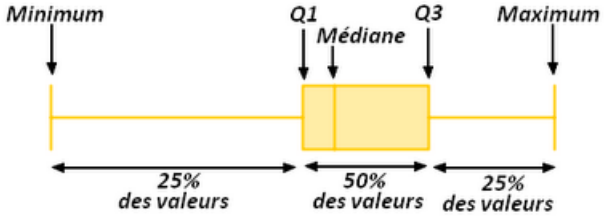
Remarque 3.3. Le deuxième quartile est en fait la médiane de la série.

Exemple 3.4. Reprenons les baguettes de pain. L'effectif total est 17. Un quart de 17 est égal à 4,25 ainsi, le premier quartile sera la cinquième valeur afin d'avoir au moins un quart de valeurs inférieures. En comptant, on trouve $Q_1 = 0,90$. Trois quarts de 17 font 12,75, on prendra donc la treizième valeur pour Q_3, c'est à dire $Q_3 = 1$. L'écart interquartile vaut donc ici $1 - 0,90 = 0,10$.

Remarque 3.5. On pourra également dans certains contextes utiliser la courbe des fréquences cumulées croissantes pour déterminer médiane et quartiles:



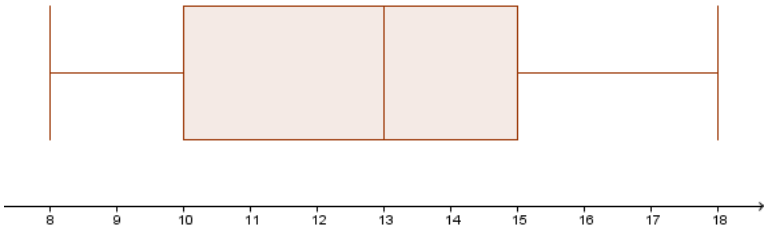
3.3. **Boîte à moustaches.** La *boîte à moustaches* est un moyen rapide de figurer le profil essentiel d'une série statistique quantitative et résume l'étude en faisant apparaître sur un même diagramme les valeurs extrêmes, les quartiles et la médianes de la manière suivante:



Remarque 3.6. On peut calculer médiane et quartiles avec la calculatrice ainsi que faire afficher la boîte à moustache par la calculatrice.

Pensez à bien faire figurer une **graduation** sous votre boîte à moustaches. Dans le cas contraire, sa lecture n'est pas possible et elle ne sert donc à rien!

Exemple 3.7. On représente la boîte à moustaches de l'exemple initial des notes (figure réalisée avec *GeoGebra)*



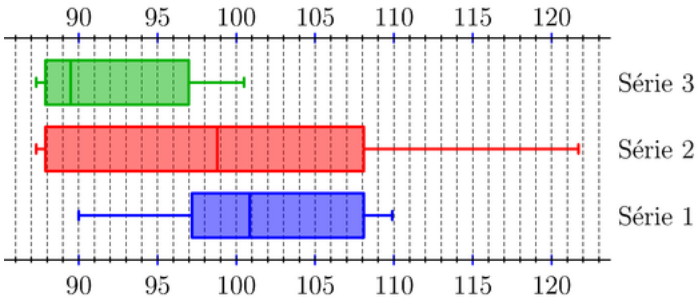
On peut donc lire directement que trois quarts des élèves ont une note supérieure à 10.

Exercice 3.8. On s'intéresse au temps total de transport de 133 employés d'une usine pendant une semaine. On représente le tableau des effectifs ci-dessous.

Temps total (en heures)	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Effectifs	1	2	3	6	8	10	15	24	16	13	12	11	9	3

Après avoir déterminé la médiane et les quartiles de cette série statistique, tracer la boîte à moustaches correspondante

Exercice 3.9. A l'aide de la figure, répondre aux questions suivantes:



a) Quelle est la médiane de la série 1, arrondie au dixième?
b) Quelle série a au moins 50% de ses valeurs comprises entre 88 et 98?
c) Quelles sont les valeurs extrêmes de la série 2?
d) Vrai ou faux? 75% des valeurs de la série 3 sont supérieures ou égales à 89, 5.