A. Rappels

1. Introduction et vocabulaire

La statistique est la science qui consiste à réunir des données chiffrées, à les analyser, à les commenter et à les critiquer. Une étude statistique s'effectue sur un ensemble appelé Population, dont les éléments sont appelés Individus, et consiste à observer et étudier un même aspect sur chaque individu, appelé Caractère.
On distingue deux types de caractère :
- Les caractères qualitatifs : Ce sont les caractères dont les valeurs ne sont pas des nombres (profession, couleur des yeux, lieu de naissance, ...);
- Les caractères quantitatifs : Ce sont les caractères qui prennent des valeurs numériques ;
- Le caractère quantitatif est discret si les valeurs du caractère sont isolées (ex : nombre d'enfants, nombre de langues vivantes étudiées par un élève, nombre de repas pris au self du lycée).
- Le caractère quantitatif est continu si les valeurs du caractère sont regroupées en intervalles, appelés Classes (ex : Taille $\in [170;175[$, durée du trajet domicile lycée en minutes $\in [0;60[$, …). La « largeur » de chaque intervalle s'appelle l'amplitude de la classe.

2. Effectifs et fréquences

a) On appelle effectif d'une valeur (respectivement d'une classe) le nombre d'individus possédant le caractère de cette valeur (respectivement d'une classe) . La somme des effectifs est appelée effectif total.
On appelle fréquence d'une valeur (respectivement d'une classe) le quotient de l'effectif de cette valeur par l'effectif total de la population.
Les fréquences sont des nombres compris entre 0 et 1, souvent exprimées en pourcentage :

$$\text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}} \quad \text{ou} \quad \text{fréquence} = \frac{\text{effectif de la valeur}}{\text{effectif total}} \times 100 \text{ pour obtenir un pourcentage.}$$

La somme des fréquences est égale à 1 ou 100%.
b) Effectifs et fréquences cumulé(e)s croissant(e)s et/ou décroissant(e)s
Dans le cas d'une variable quantitative, on peut ordonner les différentes valeurs de la variable dans l'ordre croissant ou décroissant.
On peut ainsi déterminer "Quel effectif ou quelle fréquence de la population a une valeur du caractère au plus égale ou au moins égale à ...."
Ce sont les notions d'effectifs cumulés croissants ou décroissants, ou de fréquences cumulées croissantes ou décroissantes

3. Les représentations graphiques

On représente graphiquement la série statistique par différents diagrammes:
a) Pour les séries statistiques à caractère qualitatifs , on utilise souvent des diagrammes à secteurs :
Diagramme en secteurs circulaires :
Les aires des secteurs sont proportionnelles aux effectifs ou aux fréquences;
Les angles des secteurs sont proportionnels aux effectifs ou aux fréquences selon le tableau de proportionnalité :

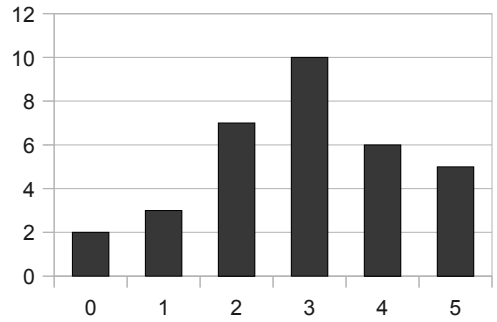
Effectif total	Effectif de la valeur
360 °	Angle du secteur

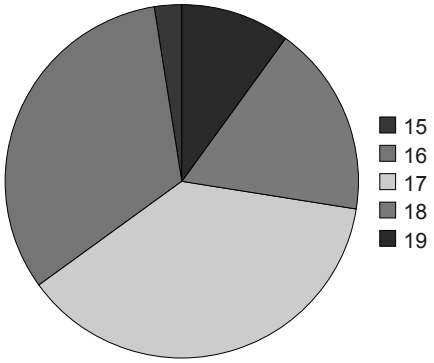
(Attention ! pour un diagramme semi-circulaire, l'effectif total correspond à un angle de 180°)

15 ans	2,5%
16 ans	32,5%
17 ans	37,5%
18 ans	17,5%
19 ans	10,00%

b) Pour les caractères discrets, on peut utiliser les diagrammes "en bâtons" :

Notes	Effectifs
0	2
1	3
2	7
3	17
4	10
5	5





Pour les caractères continus, regroupés en intervalles, on utilise un "histogramme".
Dans les deux types de représentation graphique, le caractère est porté en abscisses et l'effectif ou la fréquence sont portés en ordonnée.
Cas particulier : Histogramme à classe d'amplitudes inégales :
Si les amplitudes des classes ne sont pas égales, alors ce sont les aires des rectangles qui doivent être proportionnelles aux effectifs des classes.
Sur l'axe des abscisses, on représente les classes.
En pratique, pour la construction de ces rectangles on procède de la manière suivante :
On décide d'une aire correspondant à une unité statistique (un effectif). On fait apparaître un tableau de proportionnalité entre aire et ordonnée qui est la hauteur du rectangle. On peut utiliser un tableau de proportionnalité entre les aires des rectangles et les effectifs. La largeur du rectangle est l'amplitude de la classe et la hauteur du rectangle est = $\frac{\text{effectif de la classe}}{\text{amplitude}}$ ou $\frac{\text{aire du rectangle}}{\text{amplitude}}$

B. Étude des séries statistiques à une variable
1. Caractéristiques de position :
La vue d'un tableau ou d'un graphique ne permet pas forcément de connaître suffisamment des données pour pouvoir en analyser les répartitions, d'autant que la consultation de tableaux peut s'avérer très longue. On cherche alors à résumer celle-ci par une caractéristique de tendance centrale, c'est à dire par un seul nombre destiné à caractériser l'ensemble d'une façon objective et impersonnelle.
a) La moyenne arithmétique
La moyenne arithmétique d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par l'effectif total. On la note $\overline{x}$ (lire x barre).
Exemple : Un élève qui a eu comme notes 4, 5, 7, 9 et 12 a une moyenne égale à : $\overline{x} = \frac{4+5+7+9+12}{5} = 7{,}4$.
Inconvénient : Le calcul peut s'avérer très lourd lors de l'énumération d'un grand nombre de données.

b) La moyenne pondérée
Exemple : Si, dans une classe, 4 élèves ont obtenu la note 8, 3 élèves ont obtenu la note 10 et 5 élèves ont obtenu la note 12, on calcule $\overline{x} = \frac{4\times8+3\times10+5\times12}{4+3+5} = 12$.
Définition : Si pour une population donnée, on a p valeurs du caractère x_1, x_2, ..., x_p d'effectifs respectifs n_1, n_2, ..., n_p alors la moyenne de cette série statistique est donnée par $\overline{x} = \frac{n_1\times x_1+n_2\times x_2+...+n_p\times x_p}{n_1+n_2+...+n_p}$.
Cas d'une variable continue : Pour calculer la moyenne d'une série statistique à caractère continu, on remplace chaque classe par son milieu, avec la part d'approximation que cela comporte.

c) Propriété de la moyenne
Propriété : Soient deux séries statistiques S_1 et S_2 d'effectifs totaux respectifs N_1 et N_2 et de moyennes respectives $\overline{x_1}$ et $\overline{x_2}$, alors la moyenne de la série S obtenue en regroupant S_1 et S_2 est donnée par : $\overline{x} = \frac{N_1\times\overline{x_1}+N_2\times\overline{x_2}}{N_1+N_2}$
Exemple : Dans une classe de 22 élèves, il y a 4 filles et 18 garçons. Lors d'un devoir, les 4 filles obtiennent 13,7 de moyenne et les 18 garçons 12,8. La moyenne de la classe est donc $\overline{x} = \frac{4\times13{,}7+18\times12{,}8}{4+18} = 12{,}96$.

Définition : On dit que S_1 et S_2 sont des sous-séries statistiques (ou séries statistiques extraites) de S.
Propriété : Soit S une série statistique, de valeurs du caractère notées x_i affectées des coefficients ou effectifs n_i , et de moyenne $\overline{x}$. Soient a et b deux réels quelconques. Alors la série S' , de valeurs du caractère $ax_i + b$ affectées des mêmes effectifs n_i, a pour moyenne $a\,\overline{x} + b$.
Exemple : Dans une classe, 4 élèves ont obtenu la note 8, 3 élèves ont obtenu la note 10 et 5 élèves ont obtenu la note 12. La moyenne est donc $\overline{x} = \frac{4\times8+3\times10+5\times12}{4+3+5} = 10{,}17$.
Si l'enseignant décide de transformer les notes sur 40, et de les augmenter de un point (sur 40), la moyenne de la nouvelle série statistique sera $2\,\overline{x} + 1 = 21{,}34$.
d) Le mode ou la valeur modale
Définition: Le mode est la valeur du caractère de plus grand effectif.
La valeur modale ou classe modale est la classe de plus grand effectif.

e) La médiane

Définition: La médiane d'une série statistique est la valeur du caractère qui partage l'effectif total en deux parties égales, c'est à dire telle qu'il y ait autant d'observations ayant une valeur supérieure ou égale à la médiane que d'observations ayant une valeur inférieure ou égale à la médiane.

Exemple : Un groupe d'élève a obtenu les notes suivantes : 6, 7, 8, 9 et 20 .

Leur moyenne est donc $\bar{x} = \frac{6+7+8+9+20}{5} = 10$. Cette moyenne n'est pas très représentative de la répartition des notes, car tous les élèves sauf un, ont une note strictement inférieure à 10. La note médiane est égale à 8 : Il y a autant d'élèves qui ont 8 ou plus que d'élèves qui ont 8 ou moins.

Remarque : Soit n l'effectif total. Si n est impair, on divise n par 2 et on prend l'entier supérieur ; $n = 2p + 1$ alors la médiane correspond à la $p + 1$ ème valeur.
Si n est pair, on divise n par 2; $n = 2p$ alors la médiane correspond à la moyenne arithmétique entre la p ème et la $p + 1$ème valeur.

Cas d'une variable continue : Si le caractère est continu, on détermine la valeur du caractère correspondant à la fréquence cumulée 50% (ou à l'effectif cumulé de n), en utilisant le tableau ou le polygone des effectifs ou fréquences cumulé(e)s et en effectuant une interpolation linéaire.

2. Caractéristiques de dispersion :

a) Les quartiles, déciles et centiles

Définition : Les quartiles sont les valeurs du caractère qui partagent l'effectif total en 4 parties égales. Plus précisément : Le quartile Q_1 est la plus petite valeur du caractère pour laquelle 25 % des valeurs de la série statistique lui sont inférieures ou égales. De même, le quartile Q_3 est la plus petite valeur du caractère pour laquelle 75 % des valeurs de la série statistique lui sont inférieures ou égales.
Il y a donc trois quartiles, le 2ème quartile correspondant à la médiane.
Là encore, le procédé de calcul des quartiles est différent selon qu'il s'agit de variables discrètes en nombre pair ou impair ou de variables continu.

Définition : Les déciles et les centiles sont les valeurs du caractère qui partagent l'effectif total en respectivement 10 et 100 parties égales.
Plus précisément : Le décile D_1 est la plus petite valeur du caractère pour laquelle 10 % des valeurs de la série statistique lui sont inférieures ou égales. On définit de même le décile D_9 . On remarque que le 5ème décile est égal à la médiane et que le 50ème centile est égal à la médiane.

b) Intervalles interquartile

L'intervalle interquartile est une caractéristique de dispersion simple. C'est l'intervalle $[Q_1 ; Q_3]$ qui contient 50% de la population.
L'écart interquartile est égal à $Q_3 - Q_1$.

3. Diagrammes en boîte

Afin de représenter différentes caractéristiques d'une série statistique, on a recours, entres autres, aux représentations dites "diagrammes en boîte" ou "boites à moustaches" ou "diagrammes à pattes".
Il s'agit d'un diagramme d'une boîte rectangulaire dont les extrémités sont Q_1 et Q_3 , des traits extérieurs à cette boîte terminés par des segments qui leur sont perpendiculaires (les moustaches) reliant la valeur minimale Min à la valeur maximale Max, et un trait dans la boîte correspondant à la médiane Me; voir dessin ci-contre.