Rappels de statistique

1 Vocabulaire

Individu: élément sur lequel porte l'étude statistique.

Population: ensemble de tous les individus. Il est très important de toujours bien choisir la population pour faire une étude statistique puisque la fiabilité des résultats de l'étude en dépendent.

Échantillon: partie de la population. On a très rarement accès à la population dans son ensemble, on étudie donc seulement un échantillon.

Caractère: propriété observée sur l'individu auquel on s'intéresse.
Un caractère est qualitatif s'il est lié à une observation ne faisant pas l'objet d'une mesure (par exemple la région de résidence de chaque français observée lors du dernier recensement ou la couleur des yeux des chiens du quartier).
Un caractère est quantitatif s'il est mesurable. Il peut être discret si les valeurs observées sont isolées (par exemple le nombre d'enfants par famille observé aussi lors du dernier recensement). Il peut être continu s'il peut prendre, au moins théoriquement, n'importe quelle valeur d'un intervalle de nombre réel (par exemple la taille de chaque individu).

Classe: sous-ensemble de la population rassemblant les individus qui ont la même valeur pour un caractère ou qui ont des valeurs contenues dans un intervalle (par exemple, les habitants de la région PACA, ou les familles de deux enfants, ou encore les habitants dont la taille en centimètre appartient à l'intervalle [164; 170[).

Amplitude d'une classe: C'est l'écart entre la plus petite valeur d'une classe et sa plus grande valeur.

Exemple 1. Étude des tailles des élèves à l'I.U.T.

2 Série statistique à une variable

Lorsque l'on étudie un caractère dans une population (sondage, vote, étude de documents,...), on obtient un certain nombre de données brutes (tous les noms inscrits sur les bulletins de vote par exemple ou liste des notes obtenues pendant l'année scolaire). La première chose à faire est d'effectuer un dépouillement de ces données. Cette étape consiste à trier les données recueillies tout en éliminant les données suspectes ou aberrantes. On obtient alors ce que l'on appelle une série statistique.

3 Effectifs et fréquences

Lorsque l'on dispose d'un grand nombre de données, il est souvent plus commode de présenter une série statistique sous forme de tableau dans lequel on note l'effectif de chaque classe. L'effectif d'une classe est son nombre d'éléments. On dispose d'un tableau d'effectif quand on connait l'effectif n_i pour chaque valeur x_i du caractère et que l'on a mis tout cela dans un tableau :

Valeurs prises par le caractère (ou classe)	x_1	x_2	$\cdots$	x_i	$\cdots$	x_p
Effectifs correspondants	n_1	n_2	$\cdots$	n_i	$\cdots$	n_p

L'effectif total est le nombre total d'individus observés :

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

La fréquence d'une classe est la proportion d'individus appartenant à cette classe. Ainsi la i-ème classe a pour fréquence :

$$f_i = \frac{n_i}{n}.$$

Remarque 1. Dans la pratique, on exprime souvent les fréquences en pourcentages car les résultats sont plus lisibles de cette façon. Cependant, si rien n'est précisé on les exprime directement de la façon présentée ci-dessus.

On remarque que :

$$\sum_{i=1}^{p} f_i = \frac{n_1 + \cdots + n_p}{n} = 1.$$

Exemple 2. Une étude statistique sur l'âge des élèves d'une classe donne la série suivante:
19 - 15 - 18 - 17 - 17 - 15 - 16 - 16 - 15 - 19 - 16 - 15 - 19 - 16 - 18 - 16 - 16 - 17 - 18 - 19 - 15 - 17 - 17 - 16 - 18 - 19 - 17 - 17 - 18 - 19 - 15 - 19 - 18 - 18 - 16 - 16 - 17 - 16 - 18 - 18 -

Si les valeurs des caractères $x_1, \ldots, x_p$ peuvent être ordonnées, on peut considérer les fréquences cumulées croissantes (FCC) qui cumulent les fréquences associées aux valeurs du caractère inférieures ou égales à x_i. On note ces fréquences cumulées F_i et

$$F_i = \sum_{j=1}^{i} f_j \text{ pour } i = 1, \ldots, p.$$

On peut bien entendu définir de façon analogue les fréquences cumulées décroissantes (FCD) , les effectifs cumulés croissants (ECC) et décroissants (ECD).

Exemple 3. Compléter le tableau de l'exemple précédent en rajoutant les ECC, ECD, FCC, FCD.

4 Représentations graphiques

4.1 Diagramme en bâtons

Cette méthode est utilisée pour représenter les séries statistiques correspondant à un caractère quantitatif à variable discrète. La longueur des bâtons est proportionnelle à:

- aux effectifs s'il s'agit d'un tableau d'effectifs;
- aux fréquences s'il s'agit d'un diagramme de fréquence,
- aux ECC (FCC)...
- aux ECD (FCD)...

Exemple 4. Faire le diagramme des effectifs et des fréquences de l'exemple précédent.

4.2 Diagramme à bandes

Pour les séries statistiques à caractère qualitatif, on utilise souvent cette méthode. Celle-ci consiste à faire exactement comme pour les diagrammes en bâtons mais on fait des rectangles à la place des traits. Il existe de nombreuses variations de cette méthode qui permettent de représenter toutes sortes de situations.

4.3 Diagramme à secteurs

Un diagramme à secteur est un disque dont l'aire est proportionnelle aux fréquences ou aux effectifs. Il est utilisé pour les séries statistiques correspondant à un caractère quantitatif ou qualitatif à variable discrète.

Pour dessiner un tel diagramme, il faut utiliser les règles de proportionnalité. En effet, pour des fréquences 100% correspond à un disque entier soit 360°, 50% correspond donc à $\frac{50}{100} \times 360 = 180°$...

Exemple 5. Dans une société d'assurances, les salaires mensuels payés aux employés sont résumés ci-dessous. Faire un diagramme à secteurs.

Salaire en euros	effectifs	fréquences	angle au centre
[800;850[	2		
[850;900[	5		
[900;950[	12		
[950;1000[	36		
[1000;1500[	30		
[1500;2000[	15		
	N=	100	360

4.4 Histogramme

Une série statistique dont les valeurs sont regroupés par classe est généralement représenté par un histogramme.
Il y a deux cas qui se présentent:

Soit toutes les classes ont la même amplitude et dans ce cas, le principe est le même que pour le diagramme en bâtons sauf qu'au lieu de faire un trait qui monte, on dessine des rectangles dont la base est la classe et dont la hauteur est proportionnelle aux effectifs ou aux fréquences.

Exemple 6. Si l'on a le tableau suivant:

Salaire en euros	effectifs	fréquences
[800;1000[	55	
[1000;1200[	18	
[1200;1400[	8	
[1400;1600[	6	
[1600;1800[	5	
[1800;2000[	8	
	N=	100

Soit toutes les classes n'ont pas la même amplitude et dans ce cas, on procède de la façon suivante:

- On commence par calculer l'amplitude de chaque classe et on prend (en général) la plus petite amplitude comme amplitude de base;
- On détermine le nombre d'intervalle de base de chaque classe,
- On détermine ensuite la hauteur de chaque rectangle en divisant l'effectif (ou la fréquence) de la classe par le nombre d'intervalles de base de la classe.

Exemple 7. Tracer l'histogramme pour le premier exemple des salaires en euros.
Tracer l'histogramme pour l'exemple suivant:

Ancienneté (ans)	[0;5[	[5;15[	[15;20[	[20;30[	[30;35[	[35;40[
Effectif	15	22	54	64	22	30
Nombre d'intervalles de base						
Effectif/nb intervalles de base						

4.5 Polygone des fréquences et des ECC (ECD, FCC,FCD)

Le polygone des fréquences se trace après avoir tracé le diagramme des fréquences. Il s'agit de relier les sommets des bâtons (ou le sommet au milieu de chaque rectangle) par des segments de droite.

Exemple 8. Tracer le polygone des fréquences sur le premier histogramme précédent.

Le polygone des effectifs cumulés croissants est obtenu en plaçant des points de coordonnées (plus grande valeur de la classe; ECC correspondant). Le polygone des effectifs cumulés décroissants est obtenu en plaçant des points de coordonnées (plus petite valeur de la classe; ECD correspondant).

Exemple 9. Tracer le polygone des ECC sur le premier histogramme précédent.

5 Caractéristiques de position

Mode (ou classe modale): Valeur correspondant à l'effectif maximal. Il est important de remarquer que le mode n'est pas forcément unique. On parle alors de série statistique multi-modale.

Valeurs maximale et minimale de la série: On note ces valeurs x_{max} et x_{min}.

Moyenne:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{p} n_i x_i}{n} = \sum_{i=1}^{p} f_i x_i.$$

Médiane: La médiane est la valeur telle que 50 % des observations de l'échantillon lui sont inférieures (et 50% lui sont supérieures).
Si le nombre d'observations est pair : la médiane est la moyenne entre les observations $\frac{n}{2}$ et $\frac{n+2}{2}$.
Si le nombre d'observations est impair : la médiane est l'observation $\frac{n+1}{2}$.

Premier et troisième quartiles: Le premier Q_1 et le troisième quartile Q_3 sont les valeurs qui partagent la série avec un quart de l'effectif avant Q_1 et trois quarts de l'effectif avant Q_3.

Décile, centiles: On définit de la même façon les déciles et les centiles qui partagent la population en 10 ou en 100.

Exemple 10. Dans un bureau de postes, le montant des retraits en euros est réparti de la façon suivante:

Montant en euro	effectif n_i	ECC	ECD	Fréquence	Centre des classes x_i	$n_i \times x_i$
[0;500[	28					
[500;1000[	28					
[1000;1500[	28					
[1500;2000[	28					
[2000;2500[	28					
[2500;3000[	28					
	N=			100		

6 Mesures de dispersion

Variance: La variance est la quantité suivante

$$V(x) = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{p} n_i x_i^2}{n} - \overline{x}^2.$$

Écart type: L'écart type est simplement la racine carrée de la variance: $\sigma = \sqrt{V(x)}$.

Étendue: L'étendue est la différence entre la plus grande valeur du caractère et la plus petite: $x_{max} - x_{min}$.

Écart interquartiles: L'écart interquartile est le nombre $Q_3 - Q_1$. Il permet de mesurer la dispersion du caractère autour de la médiane.

Coefficient de variation: Le coefficient de variation est donné par la formule: $c(x) = \frac{\sigma(x)}{\overline{x}}$. C'est une façon de normaliser l'écart type et donc de pouvoir comparer la dispersion de séries qui ne sont pas représentées à la même échelle.

6.1 Boite à moustache

On résume souvent les caractéristiques de position et de dispersion dans une boite à moustache:

Exemple 11. On a relevé les notes de 24 élèves dans une classe lors d'un examen noté sur 100: 67 74 76 77 72 65 65 68 74 76 64 57 59 54 77 76 59 77 79 67 72 72 68 78.
Déterminer la médiane, les quartiles et dessiner la boite à moustache de cette série.
Comparer ces résultats à ceux d'une autre classe où la note minimale est 47, la note maximale est 85, la médiane est 70, le premier quartile 67 et le troisième 76.