Chapitre 14

Statistiques

1/ Généralités

Une étude statistique descriptive s'effectue sur une population (des personnes, des villes, des objets...) dont les éléments sont des individus et consiste à observer et étudier un même aspect sur chaque individu, nommé caractère (taille, nombre d'habitants, consommation...).
Il existe deux types de caractères :

1/ *quantitatif* : c'est un caractère auquel on peut associer un nombre c'est-à-dire, pour simplifier, que l'on peut mesurer. On distingue alors deux types de caractères quantitatifs :
   - *discret* : c'est un caractère quantitatif qui ne prend qu'un nombre fini de valeurs. Par exemple le nombre d'enfants d'un couple.
   - *continu* : c'est un caractère quantitatif qui, théoriquement, peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels. Ses valeurs sont alors regroupées en classes. Par exemple la taille d'un individu, le nombre d'heures passées devant la télévision.

2/ *qualitatif* : comme la profession, la couleur des yeux, la nationalité. Dans ce dernier cas, « nationalité française », « nationalité allemande » etc. sont les modalités du caractère.

En général une série statistique à caractère discret se présente sous la forme :

Valeurs	x_1	x_2	.....	x_p
Effectifs	n_1	n_2	.....	n_p
Fréquences	f_1	f_2	.....	f_p

On écrira souvent : la série (x_i, n_i). (On n'indique pas le nombre de valeurs lorsqu'il n'y a pas d'ambiguïté). Souvent on notera N l'effectif total de cette série donc $N = n_1 + n_2 + \cdots + n_p$.
Lorsqu'une série comporte un grand nombre de valeurs, on cherche à la résumer, si possible, à l'aide de quelques nombres significatifs appelés paramètres.
La suite du cours présente quelques paramètres permettant de résumer des séries à caractère quantitatif qui seront illustrés à l'aide des exemples suivants :
Série 1
Une étude sur le nombre d'employés dans les commerces du centre d'une petite ville a donné les résultats suivants :

Nombre d'employés	1	2	3	4	5	6	7	8
Effectif	11	18	20	24	16	14	11	6

Série 2
Une étude sur la durée de vie en heures de 200 ampoules électriques a donné les résultats suivants :

Durée de vie en centaine d'heures	[12 ; 13[	[13 ; 14[	[14 ; 15[	[15 ; 16[	[16 ; 17[
Effectif	28	46	65	32	29

2/ Paramètres de position

a) Paramètres de position de tendance centrale

Mode - Classe modale

Définition
Le mode d'une série statistique est la valeur du caractère qui correspond au plus grand effectif.
Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, la classe modale est la classe de plus grand effectif.

Remarque : Il peut y avoir plusieurs modes ou classes modales.

Exemple : Déterminer le mode et la classe modale des séries 1 et 2.
Pour la série 1, le mode est 4.
Pour la série 2, la classe modale est l'intervalle [14; 15[.

Médiane

Définition
La médiane d'une série statistique est un réel noté M_e tel que au moins 50% des valeurs sont inférieures ou égales à M_e et au moins 50% des valeurs sont supérieures ou égales à M_e.

Dans le cas d'une série à caractère discret, la médiane s'obtient en ordonnant les valeurs dans l'ordre croissant et en prenant la valeur centrale si N est impair et la moyenne des valeurs centrales si N est pair.
Dans le cas d'une série à caractère continu, la médiane peut s'obtenir de manière graphique en prenant la valeur correspondant à $0,5$ sur le polygone des fréquences cumulées croissantes.

Exemple : Déterminer la médiane de la série 1 et la classe médiane de la série 2.
La médiane de la série 1 est 4 qui correspond à la moyenne de la soixantième et de la soixante et unième valeur.
La centième valeur de la série 2 appartient à l'intervalle [14; 15[qui est donc la classe médiane.

Moyenne

Définition
La moyenne d'une série statistique (x_i, n_i) est le réel, noté $\bar{x}$ défini par :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_px_p}{N} = \frac{\sum_{i=1}^{p} n_ix_i}{N}$$

Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.

Exemple : Déterminer la moyenne des séries 1 et 2.

$$\bar{x}_1 = \frac{11\times 1 + 18\times 2 + 20\times 3 + 24\times 4 + 16\times 5 + 14\times 6 + 11\times 7 + 6\times 8}{11+18+20+24+16+14+11+6} = 4,1$$
$$\bar{x}_2 = \frac{28\times 12,5 + 46\times 13,5 + 65\times 14,5 + 32\times 15,5 + 29\times 16,5}{28+46+65+32+29} = 14,44$$

b) Paramètres de position non centrale

Quartiles

Définition

Le premier quartile Q_1 est la plus petite valeur du caractère telle qu'au moins 25% des termes de la série aient une valeur qui lui soit inférieure ou égale.
Le troisième quartile Q_3 est la plus petite valeur du caractère telle qu'au moins 75% des termes de la série aient une valeur qui lui soit inférieure ou égale.

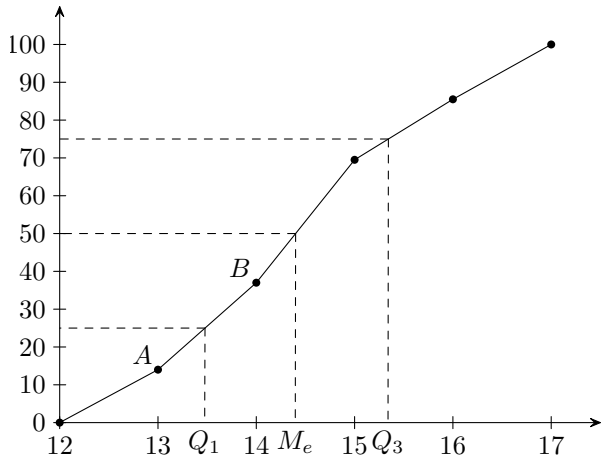
Dans le cas d'une série à caractère discret, les quartiles s'obtiennent en ordonnant les valeurs dans l'ordre croissant puis :
– Si N est multiple de 4 alors Q_1 est la valeur de rang $\frac{N}{4}$ et Q_3 est la valeur de rang $\frac{3N}{4}$.
– Si N n'est pas multiple de 4 alors Q_1 est la valeur de rang immédiatement supérieur à $\frac{N}{4}$ et Q_3 est la valeur de rang immédiatement supérieur à $\frac{3N}{4}$.

Exemple : Déterminer les quartiles de la série 1

Le nombre de données est multiple de 4 : 120. Le premier quartile est donc la 30$^{\text{e}}$ valeur et le troisième quartile la 90$^{\text{e}}$ valeur.
On a ainsi : $Q_1 = 3$ et $Q_3 = 6$.

Dans le cas d'une série à caractère continu, les quartiles peuvent s'obtenir à partir du polygone des fréquences cumulées croissantes où Q_1 est la valeur correspondant à la fréquence cumulée croissante égale $0,25$ et Q_3 est la valeur correspondant à la fréquence cumulée croissante égale $0,75$.

Exemple : Représenter le polygone des fréquences cumulées croissantes de la série 2. Déterminer graphiquement la médiane et les quartiles.



On obtient $Q_1 \simeq 13,5$, $M_e = 14,4$ et $Q_3 \simeq 15,3$

Exemple : Retrouver ces résultats par le calcul.

Le premier quartile est dans la classe $[13;14[$. En posant $A(13;14)$ et $B(14;37)$, Q_1 est le point de (AB) d'ordonnée 25.

Le coefficient directeur de (AB) est égal à $\dfrac{37-14}{14-13} = 23$ et l'ordonnée à l'origine se calcule en utilisant, par exemple, les coordonnées de A.
On obtient l'équation suivante : $(AB) : y = 23x - 285$

Q_1 étant le point de (AB) d'ordonnée 25, il est solution de l'équation $25 = 23Q_1 - 285$. On obtient $Q_1 = \dfrac{310}{23} \simeq 13,5$

Avec des raisonnements analogues, on obtient $M_e = 14,4$ et $Q_3 \simeq 15,3$

3/ Paramètres de dispersion

a) Étendue

Définition

L'étendue est la différence entre la plus grande valeur du caractère et la plus petite.

Remarque : L'étendue est très sensible aux valeurs extrêmes.

Exemple : Déterminer l'étendue des séries 1 et 2

L'étendue de la série 1 est $8 - 1 = 7$.

L'étendue de la série 2 est $17 - 12 = 5$.

b) Écart interquartile

Définition

L'intervalle interquartile est l'intervalle $[Q_1; Q_3]$.

L'écart interquartile est le nombre $Q_3 - Q_1$. C'est la longueur de l'intervalle interquartile.

Remarque : Contrairement à l'étendue, l'écart interquartile élimine les valeurs extrêmes, ce peut être un avantage. En revanche il ne prend en compte que 50% de l'effectif, ce peut être un inconvénient.

Exemple : Déterminer l'intervalle interquartile et l'écart interquartile des séries 1 et 2.

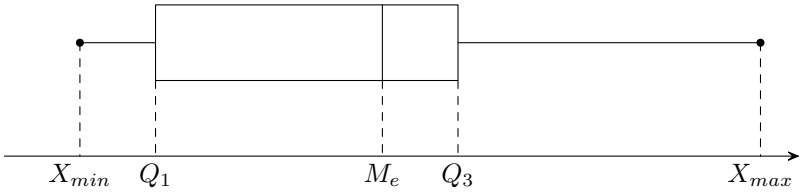
Pour la série 1 : l'intervalle interquartile est $[3; 6]$. L'écart interquartile est donc 3.

Pour la série 2 : l'intervalle interquartile est $[13,5; 15,3]$. L'écart interquartile est donc 1,8.

c) Diagramme en boîte

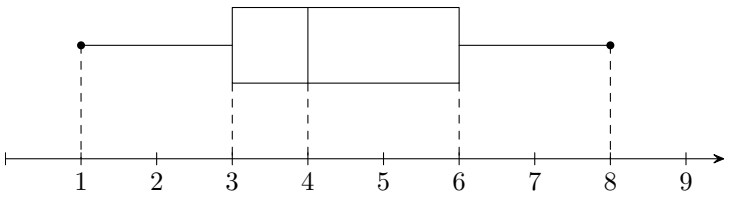
On construit un diagramme en boîte de la façon suivante :

– les valeurs du caractère sont représentées sur un axe (vertical ou horizontal) ;
– on place sur cet axe, le minimum, le maximum, les quartiles et la médiane de la série ;
– on construit alors un rectangle parallèlement à l'axe, dont la longueur est l'interquartile et la largeur arbitraire.



Remarque : Ce diagramme permet non seulement de visualiser la dispersion d'une série mais aussi de comparer plusieurs séries entre elles.

Exemple : Construire le diagramme en boîte de la série 1



d) Variance et écart-type

Pour mesurer la dispersion d'une série, on peut s'intéresser à la moyenne des distances des valeurs à la moyenne. On utilise plutôt les carrés des distances qui facilitent les calculs.

Définition

On appelle variance d'une série quelconque à caractère quantitatif discret le nombre :

$$V = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

On appelle écart-type de cette série le nombre $\sigma = \sqrt{V}$.

Dans le cas d'une série à caractère quantitatif continu dont les valeurs sont regroupées en classes, x_i désigne le centre de chaque classe.

Propriété

On peut calculer la variance de la façon suivante :

$$V = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2$$

Démonstration

$$\begin{aligned} V &= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2n_i x_i \bar{x} + n_i \bar{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\bar{x} \times \frac{1}{N}\sum_{i=1}^{n} n_i x_i + \bar{x}^2 \times \frac{1}{N}\sum_{i=1}^{n} n_i = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2 \end{aligned}$$

Exemple : Déterminer les variances et écart-types des séries 1 et 2.

Pour la série 1 : $V_1 = \dfrac{571}{150} \simeq 3,8$ et $\sigma_1 = \sqrt{V_1} \simeq 1,95$

Pour la série 2 : $V_2 = 1,5264$ et $\sigma_1 = \sqrt{V_2} \simeq 1,24$

Propriété

La fonction $g : t \longmapsto \dfrac{1}{N}\sum_{i=1}^{p} n_i(x_i - t)^2$ admet un minimum atteint en $t = \bar{x}$ (la moyenne de la série) et ce minimum vaut V (la variance de la série).

Démonstration

Pour tout $t \in \mathbb{R}$:

$$\begin{aligned} g(t) &= \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - t)^2 = \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 - 2n_i x_i t + n_i t^2 \\ &= \left(\frac{1}{N}\sum_{i=1}^{p} n_i\right) t^2 - 2 \times \left(\frac{1}{N}\sum_{i=1}^{n} n_i x_i\right) t + \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 = t^2 - 2\bar{x}t + \frac{1}{N}\sum_{i=1}^{p} n_i x_i^2 \end{aligned}$$

g est un polynôme du second degré qui atteint son minimum en $\dfrac{2\bar{x}}{2} = \bar{x}$. Ce minimum est $g(\bar{x}) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = V$

4/ Influence d'une transformation affine

Propriété

Soit (x_i, n_i) une série statistique de médiane M_x, de quartiles Q_{1x} et Q_{3x}, de moyenne $\bar{x}$, de variance V_x et d'écart-type σ_x.
Si, pour tout i, $y_i = ax_i + b$, où a et b sont des réels, alors les paramètres de la série (y_i, n_i) sont :
- Médiane : $M_y = aM_x + b$
- Quartiles : Si $a > 0$, $Q_{1y} = aQ_{1x} + b$ et $Q_{3y} = aQ_{3x} + b$
- Moyenne : $\bar{y} = a\bar{x} + b$
- Variance : $V_y = a^2 V_x$
- Écart-type : $\sigma_y = |a|\sigma_x$

Exemple : Déterminer les paramètres de la série 2 où la durée de vie est exprimée en minutes au delà de 12 heures.

Les nouvelles valeurs de la série sont obtenues en appliquant aux valeurs de départ la transformation affine $x \mapsto 60x - 720$.
On obtient alors :
$M_y = 60M_x - 720 = 60 \times 14,4 - 720 = 144$
$Q_{1y} = 60Q_{1x} - 720 = 60 \times \dfrac{310}{23} - 720 \simeq 88,7$ et $Q_{3y} = 60Q_{3x} - 720 = 60 \times \dfrac{245,5}{16} - 720 \simeq 200,6$
$\bar{y} = 60\bar{x} - 720 = 146,4$
$V_y = 60^2 V_x = 5495,04$
$\sigma_y = 60\sigma_x \simeq 74,13$

5/ Résumé d'une série statistique

On résume souvent une série statistique par une mesure de tendance centrale associée à une mesure de dispersion. les plus utilisées sont les suivantes :
- Le couple *médiane ; écart interquartile.*
  Il est insensible aux valeurs extrêmes et permet de comparer rapidement deux séries (par exemple grâce au diagramme en boîte) mais sa détermination n'est pas toujours pratique car il faut classer les données et il n'est pas possible d'obtenir la médiane d'un regroupement de séries.
- Le couple *moyenne ; écart-type.*
  Il est sensible aux valeurs extrêmes mais se prête mieux aux calculs. On peut notamment obtenir la moyenne et l'écart-type d'un regroupement de séries connaissant la moyenne et l'écart-type des séries de départ.