6
Statistiques

Résumé

Nous sommes noyés quotidiennement dans les statistiques : à la radio, à la télévision, sur les réseaux sociaux, etc... L'objectif du chapitre est de revoir certains objets statistiques connus comme la moyenne ou la médiane et de leur adjoindre des outils d'analyse et de comparaison comme l'écart-type.

1 Vocabulaire

Les premières études statistiques étant démographiques, de celles-ci, on a gardé le vocabulaire.

Définitions

L'ensemble sur lequel porte l'étude statistique s'appelle la **population**. Un élément de cet ensemble est un **individu**. L'aspect étudié s'appelle le **caractère**.

Exemple Au cours d'une enquête portant sur les bébés nés en 1999, on s'intéresse à leur taille en centimètres, leur mois de naissance, la couleur de leurs yeux, leur nombre d'heures quotidiennes de sommeil.
La population étudiée est **l'ensemble des bébés nés en 1999**.
Un individu est **un de ces bébés**.
Un des caractères étudiés est **le mois de naissance**.

Définitions

Si le caractère étudié prend des valeurs numériques alors on dit qu'il est **quantitatif** sinon, on dit qu'il est **qualitatif**.

Exemple Dans l'étude statistique précédente,

- les caractères quantitatifs sont : la taille, le nombre d'heures quotidiennes de sommeil.
- les caractères qualitatifs sont : le mois de naissance, la couleur des yeux.

Définition

Un caractère *quantitatif* peut être discret ou continu. On dit qu'il est **discret** s'il ne peut prendre que des valeurs isolées (ex : 0, 1, 2, ...). On dit qu'il est **continu** s'il peut prendre toutes les valeurs d'un intervalle.

Exemple Dans l'étude statistique présentée en exemple,

- le caractère quantitatif discret est : **le nombre d'heures quotidiennes de sommeil**.
- Le caractère quantitatif continu est : **la taille**.

Définition | Effectif

L'**effectif** d'une valeur est le nombre d'individus ayant cette valeur suivant le caractère étudié.
L'**effectif total** est le nombre d'individus de la population. C'est la somme des effectifs de toutes les valeurs du caractère.

Définition | Fréquence

La **fréquence** d'une valeur est le quotient de l'effectif de cette valeur par l'effectif total.

$$\text{Fréquence de } x_i = \frac{\text{effectif de } x_i}{\text{effectif total}}$$

2 Caractéristiques de position et de dispersion

Une série statistique peut contenir de très nombreuses données (plusieurs milliers parfois). Il est donc nécessaire de trouver une façon de résumer ces données. Pour cela, on calcule des éléments caractéristiques qui sont des indicateurs de position et des indicateurs de dispersion.

Exemple D'après l'INSEE, on sait qu'un 2009, le salaire mensuel moyen des français est de 1997 euros et que la moitié des français gagnait moins de 1447 euros par mois.
Les deux éléments fournis ici résument en quelque sorte l'ensemble des données. On dit que ce sont des indicateurs.

Définition | Étendue

L'**étendue** d'une série statistique est la différence entre la plus grande et la plus petite valeur de cette série.

Exemple Pour la série $5;5;6;7;9;10$, l'étendue est égale à 5.

Remarque L'étendue est un indicateur de dispersion très naïf et nous allons en chercher des plus intéressants associés à des indicateurs de positions.

2.1 Couple moyenne/écart-type

On considère la série statistique suivante : $x_1;x_2;\dots;x_n$.

Définition | Moyenne

La **moyenne** de cette série est le nombre, noté $\overline{x}$, défini par :

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Exemple La moyenne de la série $4;12;15;12;17;4;0;13;16$ est :

$$\frac{4+12+15+12+17+4+0+13+13}{9} = 10.$$

Propriété

La moyenne est **linéaire**.
C'est-à-dire, si on a la série $ax_1+b;ax_2+b;\dots;ax_n+b$ avec $a,b \in \mathbb{R}$, alors la moyenne de cette nouvelle série est $a\overline{x}+b$.

Démonstration. Notons $\overline{y}$ la moyenne de cette nouvelle série $y_1;y_2;\dots;y_n$ avec $y_i = ax_i + b$.

$$\overline{y} = \frac{y_1+\cdots+y_n}{n} = \frac{ax_1+b+\cdots+ax_n+b}{n} = a\frac{x_1+\cdots+x_n}{n} + b\frac{1+\cdots+1}{n} = a\overline{x}+b \quad \square$$

Remarque Certaines valeurs de la série peuvent revenir plusieurs fois et on résume souvent les séries statistiques avec des **tableaux d'effectifs** comme le suivant. Il permet de donner une formule plus efficace de la moyenne quand il y a beaucoup de termes identiques (avec $n = n_1+n_2+n_3+\cdots+n_p$).

Valeur	x_1	x_2	x_3	...	x_p
Effectif	n_1	n_2	n_3	...	n_p

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \cdots + n_px_p}{n_1+n_2+n_3+\cdots+n_p}$$

On vérifie bien dans l'exemple précédent que

$$\overline{x} = \frac{1\times0 + 2\times4 + 2\times12 + 2\times13 + 1\times15 + 1\times17}{9} = 10.$$

Propriété

Si on note f_i les fréquences associées aux valeurs x_i alors la moyenne est égale à

$$\overline{x} = f_1x_1 + f_2x_2 + \cdots + f_px_p$$

Démonstration. Trivial par définition de la fréquence d'un caractère. $\square$

Définition | Variance

On définit la **variance** V d'une série statistique comme « la moyenne des écarts à la moyenne au carré » :

$$V = \frac{(x_1-\overline{x})^2 + \cdots + (x_n-\overline{x})^2}{n}.$$

Remarques ▶ On associe à notre indicateur de position, la moyenne, un indicateur de dispersion autour de la moyenne : la variance.

▶ Si on résume la série statistique avec un tableau d'effectifs comme précédemment, on peut donner une autre formule de la variance.

Valeur	x_1	x_2	x_3	...	x_p
Effectif	n_1	n_2	n_3	...	n_p

$$V = \frac{n_1(x_1-\overline{x})^2 + \cdots + n_p(x_{n_p}-\overline{x})^2}{n}$$

Définition | Écart-type

L'**écart-type** σ d'une série statistique est la racine carrée de sa variance.

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

Exemple Calculons l'écart-type de la série statistique : $5;5;6;7;9;10$.

On calcule d'abord la moyenne : $\overline{x} = \dfrac{5+5+6+7+9+10}{6} = 7$ et ensuite la variance :

$$V = \frac{(5-7)^2 + (5-7)^2 + (6-7)^2 + (7-7)^2 + (9-7)^2 + (10-7)^2}{6} = \frac{22}{6}$$

Nous obtenons l'écart-type via : $\sigma = \sqrt{V} = \sqrt{\frac{22}{6}} \simeq 1{,}915$.

Remarques ► Dans la pratique, c'est l'écart-type qui nous intéresse le plus mais nous sommes **obligés** de calculer la variance V avant pour connaitre σ. Il ne faut surtout pas les confondre.

► Le principal intérêt de l'écart-type est que si on multiplie toutes les valeurs de la série par un nombre réel positif a alors l'écart-type de la nouvelle série est $a\sigma$ alors que sa variance est a^2V.

2.2 Couple médiane/écart interquartile

Définition | Médiane

La **médiane** d'une série statistique est le nombre M_e, **valeur de la série**, tel que : 50% au moins des individus ont une valeur du caractère inférieure ou égale à M_e et 50% au moins des individus ont une valeur supérieure ou égale à M_e.

Remarque La médiane correspond à une valeur qui partage en deux parties (presque) égales la série statistique.

Exemple **Série à valeurs quantitatives discrètes :**
On considère les notes suivantes correspondant à un groupe de 20 élèves :

$$1;12;13;14;7;9;20;10;11;12;2;7;9;9;15;10;12;11;17;1$$

Voici la méthode pour obtenir la médiane.

Première étape : Je range dans l'ordre croissant ces valeurs :

$$1;1;2;7;7;9;9;9;10;10;11;11;12;12;12;13;14;15;17;20$$

Seconde étape : Il y a au total 20 valeurs. Je fais deux paquets de 10 et je trouve :

$$\text{Médiane} = 10{,}5$$

Il s'agit du milieu de la valeur moyenne entre 10 et 11.

Remarque Attention, deux cas peuvent se présenter. Soit on a un **effectif total pair** et on peut alors diviser l'ensemble de la série en deux paquets sans omettre aucune valeur, soit on a un **effectif total impair** et on ne peut former deux ensembles de valeurs que si on en omet une.

Exemple **Série à valeurs quantitatives continues :**
La technique est ici bien différente car elle s'appuie sur ce que l'on appelle le **polygône effectifs cumulés croissants** ou celui des **fréquences cumulées croissantes**.
Une enquête sur le temps de travail personnel des élèves en classe de seconde d'un lycée a donné les résultats ci-dessous.

Temps de travail (Heure)	[0 ;1[	[1 ;2[	[2 ;3[	[3 ;4[	[4 ;5]
Effectif	40	95	86	24	5

L'objectif est ici de déterminer pour cette série la médiane (nous verrons juste après que cette démarche permet aussi de déterminer les quartiles).

Première étape : On calcule les effectifs cumulés croissants, notés ECC dans le tableau ci-dessous.

Tps de travail	[0 ;1[	[1 ;2[	[2 ;3[	[3 ;4[	[4 ;5]
Effectif	40	95	86	24	5
ECC	40	40+95=135	135+86=221	221+24=245	250

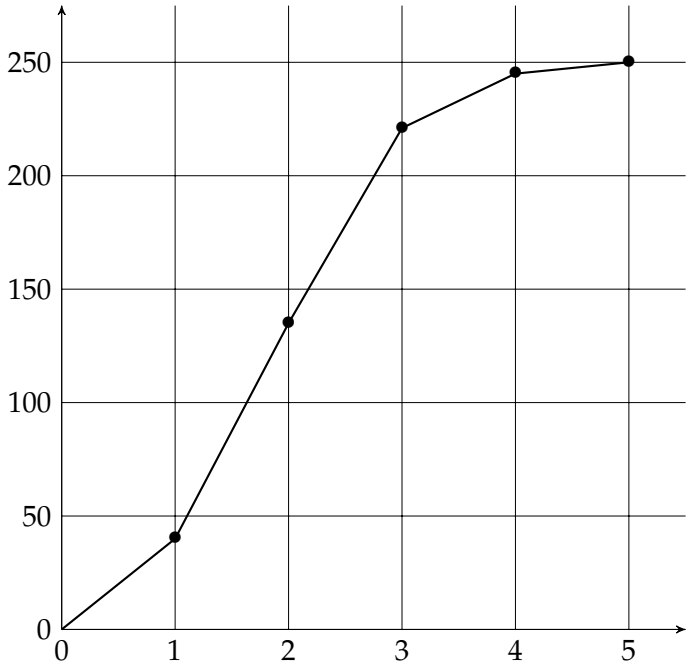
Pour obtenir des effectifs, il suffit d'ajouter les effectifs des valeurs inférieures.

Ainsi l'ECC pour la classe [2 ;3[est égale à 40 + 95 + 86 = 221.

Cet ECC signifie que 221 élèves travaillent moins de 3 heures.

Pour les fréquences cumulées croissantes, on procède de la même manière à partir des fréquences.

Seconde étape : On dessine à l'aide de ces données le polygône des ECC. Celui associé à la série que l'on considère est donnée ci-dessous.



Troisième étape : Il suffit alors de lire sur l'axe des abscisses l'abscisse du point du polygone correspondant à 125 (soit la moitié de l'effectif total qui est 250) sur l'axe des ordonnées. On obtient ainsi la médiane qui vaut ici environ $1,9$.

Définitions | Quartiles

Le **premier** et le **troisième quartile** d'une série statistique sont les nombres Q_1 et Q_3 obtenus comme suit :

- le premier quartile Q_1 est la plus petite valeur de la série telle qu'au moins 25 % des valeurs lui soient inférieures ou égales.
- le premier quartile Q_3 est la plus petite valeur de la série telle qu'au moins 75 % des valeurs lui soient inférieures ou égales.

Exemple On considère la série statistique suivante :

Valeur	30	45	50	60	61
Effectif	2	3	2	2	2

L'effectif total de cette série est égal à 11. De ce fait, le premier quartile est $Q_1 = 45$ et le troisième quartile est $Q_3 = 60$ alors que la médiane est égale à $M_e = 50$.

Remarques ▶ Le rang du 1^{er} quartile d'une série de N valeurs est le plus petit entier supérieur ou égal à $\frac{1}{4}N$.

▶ Le rang du 3^e quartile d'une série de N valeurs est le plus petit entier supérieur ou égal à $\frac{3}{4}N$.

Remarque On peut aussi, suivant les situations, procéder à partir du polygone des ECC ou des FCC. Avec le polygone des ECC, il suffit de lire sur l'axe des abscisses les abscisses de points du polygone dont l'ordonnée est égale à $\frac{1}{4}$ de l'effectif total pour le premier quartile et $\frac{3}{4}$ pour le troisième quartile.

2.3 Les indicateurs de dispersion

Définitions | Intervalle et écart interquartile

L'intervalle $[Q_1; Q_3]$ est appelé l'**intervalle interquartile**. Le nombre $Q_3 - Q_1$ est appelé l'**écart interquartile**.

Exemple Toujours dans le même exemple, $Q_1 = 5$ et $Q_3 = 9$ donc l'écart interquartile $Q_3 - Q_1$ est égal à 4.

3 Représentation graphique d'une série statistique

Une série statistique peut être représenter par différents graphiques parmi lesquels on trouve :

Le nuage de points : chaque donnée est représentée par un point.

Le diagramme en bâtons : chaque donnée est représentée par un bâton vertical.

L'histogramme : chaque donnée est représentée par un rectangle dont l'aire est proportionnelle à l'effectif.

Les deux premiers types de représentation sont appropriés au caractère quantitatif discret alors que le dernier est réservé au caractère quantitatif continu.