IV) Les paramètres de dispersion

1- L'étendue

Définition : Un des paramètres de dispersion les plus simples est l'**étendue** de la série statistique. Il s'agit de la différence entre la plus grande modalité x_p de la série et la plus petite x_1 : étendue = $x_p - x_1$

Elle donne bien un renseignement sur la répartition des valeurs de la série statistique : plus l'étendue est grande, plus l'écart entre la plus petite et la plus grande modalité de la série est importante, et donc plus les valeurs de la série peuvent être dispersées sur un grand intervalle.

Exemple :

On a relevé les notes obtenues à 6 contrôles successifs par trois élèves E_1, E_2 et E_3 d'un même groupe :

Contrôle n°	1	2	3	4	5	6
Notes de E_1	8	10	11	9	10	12
Notes de E_2	18	4	16	9	11	2
Notes de E_3	15	10	11	1	13	10

On peut vérifier que les moyennes arithmétiques de ces trois élèves sont toutes égales à 10. Cependant cette moyenne ne reflète pas du tout les différences de comportement entre ces 3 élèves :

- En effet, il semble que l'élève E_1 fournisse un travail régulier, car toutes ses notes sont proches de 10.
- L'élève E_2, au contraire, ne travaille apparemment que par rapport à sa dernière note afin de s'assurer d'avoir 10 de moyenne sans faire plus d'effort que nécessaire.
- Et l'élève E_3 a toujours des résultats convenables, sauf à un contrôle pendant lequel il ne se sentait vraisemblablement pas bien.

Le calcul de l'étendue des séries statistiques constituées des notes des élèves E_1 et E_2 permet de mettre en évidence la différence entre les comportements scolaires de ces deux élèves :

$$\text{étendue}(E_1) = 12 - 8 = 4 \qquad \text{et} \qquad \text{étendue}(E_2) = 18 - 2 = 16$$

L'étendue de E_1 est beaucoup plus faible que celle de E_2, ce qui traduit bien le fait que les notes de E_1 sont beaucoup moins dispersées autour de la note moyenne 10 que celles de E_2.

L'étendue reste cependant un paramètre de dispersion très grossier, car elle ne permet pas de refléter le fait que les notes de l'élève E_3 sont, à une exception près, toutes voisines les unes des autres. Effectivement, l'étendue de E_3 est très élevée : Etendue (E_3) = 15 – 1 = 14.

2- Ecart interquartile

Comme la moyenne arithmétique, l'étendue présente le défaut d'être très sensible aux valeurs extrêmes et aberrantes. Si, par exemple, une seule valeur d'une série statistique est très éloignée de toutes les autres, l'étendue de cette série est très élevée alors que, à une exception près, les valeurs sont toutes regroupées.

Définition :

Pour éviter de prendre en compte les valeurs extrêmes, on peut utiliser l'**écart interquartile** : c'est la différence entre le troisième et le premier quartile.

Écart interquartile $= Q_3 - Q_1$.

Q_1 : il correspond à 25% des observations ;

Q_2 : il correspond à 50% des observations. C'est aussi la médiane Me ;

Q_3 : il correspond à 75% des observations

L'écart interquartile donne donc l'amplitude de l'intervalle qui contient 50% des valeurs "les moins extrêmes", c'est-à-dire l'intervalle qu'on obtient en enlevant à la série les 25% plus faibles valeurs et les 25% plus fortes valeurs.

Exemple:

L'écart interquartile permet de différencier les comportements des élèves E_2 et E_3 de l'exemple précédent dont on rappelle les données :

Contrôle n°	1	2	3	4	5	6
Notes de E_1	8	10	11	9	10	12
Notes de E_2	18	4	16	9	11	2
Notes de E_3	15	10	11	1	13	10

En réorganisant par ordre croissant les séries de notes de ces deux élèves, on peut déterminer les quartiles de ces séries :

Notes de E_2 (ordre croissant)	2	4	9	11	16	18	Notes de E_3 (ordre croissant)	1	10	10	11	13	15
		Q_1		Q_2	Q_3				Q_1		Q_2	Q_3	

Q_2 $(E_2) = (9+11)/2 = 10$ $\qquad$ Q_2 $(E_3) = (10+11)/2 = 10{,}5$

Lecture des quartiles pour des élèves aux comportements différents :

On a donc : $\text{écart I-Q}(E_2) = 16 - 4 = 12$ et $\text{écart I-Q}(E_3) = 13 - 10 = 3$

L'écart interquartile faible de l'élève E_3 montre bien que la plupart des notes de cet élève sont voisines les unes des autres, tandis que l'écart interquartile de l'élève E_2 confirme le fait que cet élève a des résultats en "dents de scie".

3- L'écart absolu moyen

Les deux paramètres de dispersion que l'on vient de voir (l'étendue et l'écart interquartile) n'utilisent que deux valeurs (modalités extrêmes ou quartiles) de la série statistique étudiée pour estimer la dispersion des valeurs de la série. L'information qui en résulte n'est donc pas très précise.

Définition:

Pour tenir compte de toutes les valeurs de la série dans l'estimation de la dispersion de ces valeurs, on peut calculer la moyenne des écarts entre chaque valeur de la série et la moyenne arithmétique de cette série. C'est l'**écart absolu moyen par rapport à la moyenne** qu'on notera "eamm" : pour une série statistique notée X d'effectif total N, de modalités : x_1, x_2, x_3,….,x_p et d'effectifs n_i , on a :

$$\text{eamm}(X) = \frac{1}{N}\sum_{i=1}^{p} n_i \times |x_i - \overline{X}|$$

Les valeurs absolues sont nécessaires pour que les écarts positifs ne soient pas "compensés" par les écarts négatifs.

Plus l'écart absolu moyen par rapport à la moyenne est élevé, plus il y a de valeurs éloignées de la moyenne : l'écart absolu moyen est donc bien un paramètre de dispersion, c'est-à-dire un indicateur de l'étalement des valeurs recueillies.

Remarque:

On peut, de manière similaire, définir un écart moyen par rapport à la médiane quand on juge que la médiane est un paramètre de position plus approprié à l'étude d'une série statistique que la moyenne arithmétique.

Exemple:

La répartition "modalités-effectifs" de la série des notes des élèves E_1 et E_2 des exemples précédents se synthétise par le tableau suivant :

Notes de E_1	8	9	10	11	12	Notes de E_2	2	4	9	11	16	18
Effectifs	1	1	2	1	1	Effectifs	1	1	1	1	1	1

On a vu précédemment que la moyenne de cette série statistique était : M = 10. La moyenne des écarts entre chaque note et cette note moyenne est :

$$\text{eamm}(E_1) = \frac{1}{6}(1 \times |8-10| + 1 \times |9-10| + 2 \times |10-10| + 1 \times |11-10| + 1 \times |12-10|) = 1\,\text{point}$$

Cela signifie donc que, en moyenne, les notes de l'élève E_1 sont éloignées d'1 point par rapport à sa moyenne.

Le même raisonnement concernant l'élève E_2 conduit à l'écart moyen :

$$\text{eamm}(E_2) = 5\,\text{points}$$

L'écart moyen des notes de l'élève E_2 par rapport à sa moyenne est beaucoup plus grand que l'écart moyen des notes de l'élève E_1 : cela traduit ce que l'on avait constaté dès le début, à savoir que les résultats de l'élève E_2 sont bien plus irréguliers que ceux de l'élève E_1.

4- La variance

L'écart absolu moyen vu précédemment présente l'avantage de tenir compte de toutes les valeurs d'une série statistique pour donner une indication sur la dispersion de ces valeurs. Par contre, il y a un inconvénient en termes de calculs à utiliser l'écart moyen : la formule qui le définit fait intervenir des valeurs absolues souvent délicates à manipuler.

Par conséquent, plutôt que de travailler avec des écarts en valeurs absolues, on travaille avec les carrés des écarts : les carrés de nombres étant toujours positifs, il n'y a pas de compensation entre des écarts négatifs et des écarts positifs ; de plus, la fonction carrée étant croissante sur $\mathbb{R}^+$, le carré d'un écart est d'autant plus grand que l'écart lui-même est grand.

Ainsi, en remplaçant les écarts en valeurs absolues $|x_i - \overline{X}|$ par les carrés des écarts $(x_i - \overline{X})^2$ dans la formule de définition de l'écart absolu moyen, on obtient à nouveau un indicateur qui prend une valeur positive d'autant plus grande que beaucoup de modalités x_i sont éloignées de la valeur moyenne $\overline{X}$.

Définition :

On appelle **variance** d'une série statistique, et on note Var (X), la moyenne des carrés des écarts entre les valeurs de la série et la valeur moyenne.

Avec les mêmes notations que précédemment, on a :

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{p} n_i \times (x_i - \overline{X})^2$$

La variance est un nombre positif qui est d'autant plus élevé que les valeurs de la série sont dispersées.

Exemple :

Reprenons encore le cas des notes des élèves E_1 et E_2 des exemples précédents :

Notes de E_1	8	9	10	11	12	Notes de E_2	2	4	9	11	16	18
Effectifs	1	1	2	1	1	Effectifs	1	1	1	1	1	1

La variance de la série des notes de l'élève E_1 est donnée par :

$$\mathrm{Var}(E_1) = \frac{1}{6}\left[1\times(8-10)^2 + 1\times(9-10)^2 + 2\times(10-10)^2 + 1\times(11-10)^2 + 1\times(12-10)^2\right]$$
$$= \frac{5}{3}\,\text{points}^2$$

De la même façon, on calcule la variance de la série des notes de l'élève E_2 :

$$\mathrm{Var}(E_2) = \frac{101}{3}\,\text{points}^2$$

On peut finalement faire la même constatation que celle faite grâce aux écarts absolus moyens : comme la variance $\mathrm{Var}(E_2)$ est beaucoup plus grande que la variance Var(E1), les notes de l'élève E_2 sont beaucoup plus dispersées que celles de l'élève E_1.

5- *Ecart-type*

La variance d'une série statistique que l'on vient de définir est un paramètre de dispersion qui joue le même rôle que l'écart absolu moyen par rapport à la moyenne, mais qui a, en plus, l'avantage d'être plus aisé à manipuler dans les calculs du fait du remplacement des valeurs absolues présentes dans la définition de l'écart moyen par des carrés. Cependant, ce remplacement a une conséquence regrettable en termes de dimension physique : en effet, si les valeurs d'une série statistique sont des grandeurs physiques dimensionnelles, alors la variance de cette série n'a pas la même dimension que les valeurs en question.

Définition :

Pour se donner un paramètre de dispersion conservant les qualités de la variance et étant de même dimension que les valeurs de la série à laquelle il se réfère, on introduit l'**écart-type**: l'écart-type $\sigma(X)$ d'une série statistique X est tout simplement la racine carrée de la variance de cette série.

Fondamental :

La fonction racine carrée étant croissante sur $\mathbb{R}$, l'interprétation de l'écart-type est équivalente à celle de la variance et donc de l'écart absolu moyen : **plus l'écart-type d'une série est grand, plus les valeurs de cette série sont dispersées**.

Formellement, on a, toujours avec les mêmes notations que précédemment :

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{p} n_i \times (x_i - \overline{X})^2} = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{p} n_i\, x_i^2\right) - \overline{X}^2}$$

Exemple :

Pour les élèves E_1 et E_2 des exemples précédents, on avait calculé :

Etudiants	E_1	E_2
Ecarts moyens	1 point	5 points
Variances	5/3 points²	101/3 points²

Les variances, à la différence des écarts absolus moyens, s'expriment ici en " points^2 " ce qui n'a guère de signification. On préfère donc comparer la dispersion des notes des deux élèves au moyen de la racine carrée des variances, c'est-à-dire des écarts-types. Or on a bien :

$$\sigma(E_1) = \sqrt{\mathrm{Var}(E_1)} \simeq 1{,}29\,\text{points} < 5{,}80\,\text{points} \simeq \sqrt{\mathrm{Var}(E_2)} = \sigma(E_2)$$

On remarque que les écarts-types restent dans les mêmes ordres de grandeur que les écarts absolus moyens : ils en fournissent en général une bonne approximation.

6- Le coefficient de variation

Définition :

Il correspond au rapport entre l'écart-type et la moyenne arithmétique. Il est noté CV.

$$\mathbf{CV} = \frac{\sigma(X)}{\overline{X}} * \mathbf{100}$$

Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande, et vice versa.

Exemple :

En partant de l'exemple ci-dessus, nous avons :

$CV(E_1) = \frac{1{,}29}{10} * 100 = 12{,}9\%$.

$CV(E_2) = \frac{5{,}80}{10} * 100 = 58\%$.

$CV(E_2) > CV(E_1)$, alors, les notes de l'élève E_2 sont plus dispersées que celles de l'élève E_1.

Ce cours est élaboré en s'inspirant du site iutenligne.net