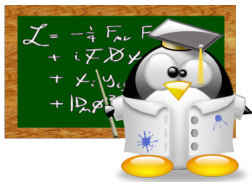




Sommaire

| | |
|---|---|
| Variance | 2 |
| Théorème de König-Huygens | 2 |
| Écart-type | 4 |
| Médiane & quartiles (rappels) | 4 |
| Diagramme en boîte (boîte à moustaches) | 5 |
| Cas des séries à caractères quantitatifs continus | 6 |
| Analyse d'une série statistique | 7 |



Prérequis

- Vocabulaire de Seconde (caractères quantitatifs)
- Moyenne, médiane, quartiles

Définition

Variance

On considère une série statistique à caractères quantitatifs discrets $(x_i ; n_i)$, $1 \leq i \leq k$, de moyenne \bar{x} .

On appelle **variance** de cette série le nombre V défini par :

$$V = \frac{n_1(\bar{x} - x_1)^2 + n_2(\bar{x} - x_2)^2 + \dots + n_k(\bar{x} - x_k)^2}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i(\bar{x} - x_i)^2}{\sum_{i=1}^k n_i}.$$

Remarque

On dit que la variance est la moyenne des carrés des écarts à la moyenne \bar{x} .

Les « écarts à la moyenne » sont les $(\bar{x} - x_i)$; les « carrés des écarts à la moyenne » sont donc les $(\bar{x} - x_i)^2$. En faisant la moyenne de ces écarts, on trouve la variance.

Exemple

Le tableau suivant donne les notes obtenues par des élèves de 1^{re} S à un devoir sur les statistiques descriptives noté sur 10.

| Notes (x_i) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---------------------|---|---|---|---|---|---|---|---|---|---|----|
| Effectifs (n_i) | 1 | 1 | 3 | 1 | 0 | 3 | 4 | 7 | 2 | 1 | 1 |

La moyenne des notes à ce devoir est :

$$\begin{aligned}\bar{x} &= \frac{0 \times 1 + 1 \times 1 + 2 \times 3 + 3 \times 1 + \dots + 9 \times 1 + 10 \times 1}{1 + 1 + 3 + 1 + 0 + 3 + 4 + 7 + 2 + 1 + 1} \\ \bar{x} &= \frac{133}{24} \\ \bar{x} &\approx 5,542.\end{aligned}$$

On a alors le tableau suivant :

| x_i | $\bar{x} - x_i$ | $n_i(\bar{x} - x_i)^2$ |
|-------|-----------------|------------------------|
| 0 | 5,542 | 30,713764 |
| 1 | 4,542 | 20,629764 |
| 2 | 3,542 | 37,637292 |
| 3 | 2,542 | 6,461764 |
| 4 | 1,542 | 0 |
| 5 | 0,542 | 0,881292 |

| x_i | $\bar{x} - x_i$ | $n_i(\bar{x} - x_i)^2$ |
|---------|-----------------|------------------------|
| 6 | -0,458 | 0,839056 |
| 7 | -1,458 | 14,880348 |
| 8 | -2,458 | 12,083528 |
| 9 | -3,458 | 11,957764 |
| 10 | -4,458 | 19,873764 |
| Somme : | | 155,958336 |

La variance de ce devoir est alors :

$$V = \frac{155,958336}{24}$$

$V \approx 6,498$

Théorème*Théorème de König-Huygens*

Soit $(x_i ; n_i)$ une série statistique à caractères quantitatifs discrets de moyenne \bar{x} et de variance V . Alors,

$$V = \frac{1}{N} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2 \quad , \quad N = n_1 + n_2 + \dots + n_k .$$

Démonstration

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



Exemple

Reprenons l'exemple précédent et ajoutons deux lignes :

| | | | | | | | | | | | | |
|---------------------|---|---|----|---|----|----|-----|-----|-----|----|-----|----------------------------|
| Notes (x_i) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| Effectifs (n_i) | 1 | 1 | 3 | 1 | 0 | 3 | 4 | 7 | 2 | 1 | 1 | |
| x_i^2 | 0 | 1 | 4 | 9 | 16 | 25 | 36 | 49 | 64 | 81 | 100 | |
| $n_i x_i^2$ | 0 | 1 | 12 | 9 | 0 | 75 | 144 | 343 | 128 | 81 | 100 | $\rightarrow \Sigma = 893$ |

$$V = \frac{893}{24} - 5,54166667^2$$

$$V \approx 6,498$$



Dans la formule du théorème de König-Huygens, il faut veiller à remplacer \bar{x} par sa valeur approchée la plus précise possible.

En effet, dans le dernier calcul de l'exemple précédent, si j'avais remplacé \bar{x} par 5,542, j'aurais obtenu $V \approx 6,495$, valeur différente de celle trouvée dans le premier calcul.

Définition

Écart-type

Soit $(x_i ; n_i)$ une série statistique à caractères quantitatifs discrets de variance V .

On appelle **écart-type** de cette série le nombre défini par :

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

Remarque

L'écart-type est un **indicateur de dispersion** : il permet de se rendre compte de la dispersion des caractères étudiés autour de la moyenne.

- Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).
- Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).

Exemple

Reprenons l'exemple du devoir de la classe de 1^{re} S où la moyenne était d'environ 5,5.

Nous avons trouvé que sa variance était : $V \approx 6,5$. Ainsi, son écart-type est $\sigma \approx \sqrt{6,5} \approx 2,5$.

Les notes seront donc concentrées dans l'intervalle $[5,5 - 2,5 ; 5,5 + 2,5] = [3 ; 8]$ (bien qu'il puisse y en avoir à l'extérieur!).

Définitions

Médiane & quartiles (rappels)

Soit $(x_i ; n_i)$ une série statistique à caractères quantitatifs. Je rappelle que :

- La **médiane** (notée m_e) est la valeur du caractère pour lequel au moins 50 % des caractères lui sont inférieurs ;
- Le **premier quartile** (notée Q_1) est la valeur du caractère pour lequel au moins 25 % des caractères lui sont inférieurs ;
- Le **troisième quartile** (noté Q_3) est la valeur du caractère pour lequel au moins 75 % des caractères lui sont inférieurs.

Exemple

Reprenons les notes obtenues au devoir sur les statistiques de notre classe de 1^{re} S, et ajoutons au tableau la ligne des effectifs cumulés croissants :

| | | | | | | | | | | | |
|---------------------|---|---|---|---|---|---|----|----|----|----|----|
| Notes (x_i) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Effectifs (n_i) | 1 | 1 | 3 | 1 | 0 | 3 | 4 | 7 | 2 | 1 | 1 |
| E.c.c. | 1 | 2 | 5 | 6 | 6 | 9 | 13 | 20 | 22 | 23 | 24 |

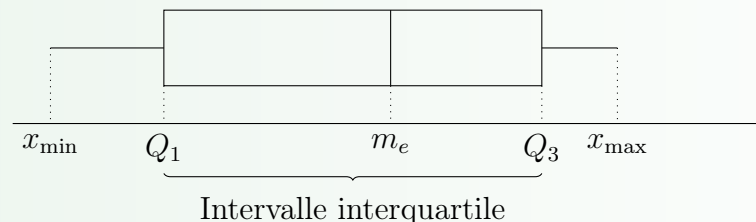
- 50 % de 24 est égal à 12 ; on atteint au moins 12 notes cumulées pour la note « 6 » ; donc $m_e = 6$;
- 25 % de 24 est égal à 6 ; on atteint au moins 6 notes cumulées pour la note « 3 » ; donc $Q_1 = 3$;
- 75 % de 24 est égal à 18 ; on atteint au moins 18 notes cumulées pour la note « 7 » ; donc $Q_3 = 7$;

Définition

Diagramme en boîte (boîte à moustaches)

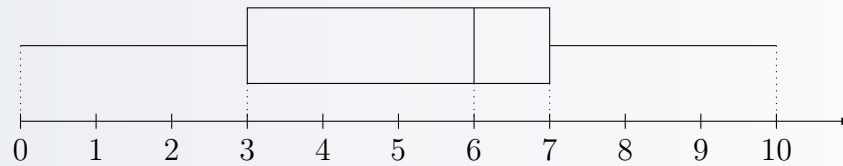
On appelle **diagramme en boîte** d'une série statistique la représentation de la dispersion des caractères de la série. Elle est composée :

- d'un rectangle de largeur $Q_3 - Q_1$ représentant ainsi l'**écart interquartile** (noté I_Q) ;
- d'un segment allant de x_{\min} à Q_1 et d'un segment allant de Q_3 à x_{\max} ;
- d'un trait vertical inclus dans le rectangle désignant la médiane.



Exemple

Le diagramme en boîte de la série des exemples précédents est :



Remarque

Cas des séries à caractères quantitatifs continus

Dans le cas où les caractères de la série sont des intervalles (des classes), on prend pour x_i le milieu de chaque classe.

Exemple

Le tableau ci-dessous représente la répartition des salaires des employés dans une PME :

| Salaires (en €) | [900 ; 1 100[| [1 100 ; 1 300[| [1 300 ; 1 500[| [1 500 ; 1 700[| [1 700 ; 1 900] |
|--------------------|---------------|-----------------|-----------------|-----------------|-----------------|
| Centre des classes | 1 000 | 1 200 | 1 400 | 1 600 | 1 800 |
| Effectifs | 10 | 30 | 50 | 40 | 10 |
| E.c.c. | 10 | 40 | 90 | 130 | 140 |

- **Moyenne** : $\bar{x} = \frac{1\,000 \times 10 + 1\,200 \times 30 + \dots + 1\,800 \times 10}{10 + 30 + 50 + 40 + 10}$
 $= \frac{198\,000}{140}$
 $\boxed{\bar{x} \approx 1\,414}$

- **Variance** : $V \approx \frac{1}{140}(10 \times 1\,000^2 + 30 \times 1\,200^2 + \dots + 10 \times 1\,800^2) - 1\,414^2$
 $\approx \frac{286\,000\,000}{140} - 1\,414^2$
 $\boxed{V \approx 43\,461}$

- **Écart-type** : $\sigma \approx \sqrt{43\,461}$
 $\boxed{\sigma \approx 208,5}$

- **Médiane** : 50 % (soit 70) est atteint pour la classe [1 300 ; 1 500[; donc $m_e \in [1\,300 ; 1\,500[$. Soient A(1 300 ; 40) et B(1 500 ; 90) dans le repère où est tracé le polygone des e.c.c.

La droite (AB) a pour équation $y = mx + p$ où $m = \frac{y_B - y_A}{x_B - x_A} = \frac{50}{200} = \frac{1}{4}$.

Donc $y_A = \frac{1}{4}x_A + p$, soit $40 = \frac{1}{4} \times 1\,300 + p$; ainsi, $p = -285$.

Finalement, on a (AB) : $y = \frac{1}{4}x - 285$. Si $y = 70$ (50 % des effectifs), alors $70 = \frac{1}{4}x - 285$, soit $x = 1\,420$.

Ainsi, $\boxed{m_e = 1\,420}$.

Exemple (suite)

- **1^{er} quartile** : 25 % (soit 35) est atteint pour la classe [1 100 ; 1 300[.

Donc $Q_1 \in [1\ 100 ; 1\ 300[$.

Soient C(1 100 ; 10) et A(1 300 ; 40) dans le repère où est tracé le polygone des e.c.c.

La droite (CA) a pour équation $y = mx + p$ où $m = \frac{y_C - y_A}{x_C - x_A} = \frac{30}{200} = \frac{3}{20}$.

Donc $y_A = \frac{3}{20}x_A + p$, soit $40 = \frac{3}{20} \times 1\ 300 + p$; ainsi, $p = -155$.

Finalement, on a (CA) : $y = \frac{3}{20}x - 155$. Si $y = 35$ (25 % des effectifs), alors

$$35 = \frac{3}{20}x - 155, \text{ soit } x \approx 1\ 266,67.$$

Ainsi, $Q_1 \approx 1\ 267$.

- **3^e quartile** : 75 % (soit 105) est atteint pour la classe [1 500 ; 1 700[.

Donc $Q_3 \in [1\ 500 ; 1\ 700[$.

Soient D(1 700 ; 130) et B(1 500 ; 90) dans le repère où est tracé le polygone des e.c.c.

La droite (BD) a pour équation $y = mx + p$ où $m = \frac{y_D - y_B}{x_D - x_B} = \frac{40}{200} = \frac{1}{5}$.

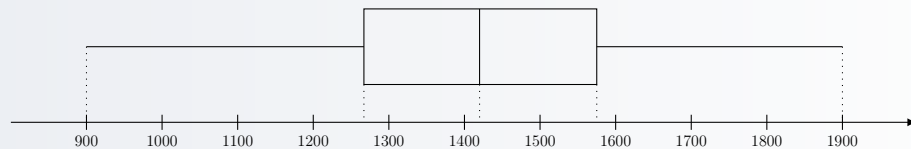
Donc $y_B = \frac{1}{5}x_B + p$, soit $90 = \frac{1}{5} \times 1\ 500 + p$; ainsi, $p = -210$.

Finalement, on a (BD) : $y = \frac{1}{5}x - 210$. Si $y = 105$ (75 % des effectifs), alors

$$105 = \frac{1}{5}x - 210, \text{ soit } x = 1\ 575.$$

Ainsi, $Q_3 = 1\ 575$.

- **Diagramme en boîte** :



Méthode

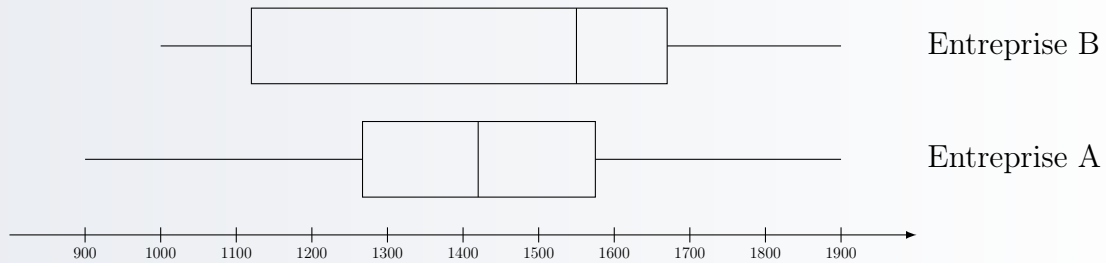
Analyse d'une série statistique

Pour analyser une série statistique, on utilise :

- le couple $(m_e ; I_Q)$, qui permet d'avoir une idée de la répartition des valeurs de la série autour de la médiane ;
- le couple $(\bar{x} ; \sigma)$, qui permet d'avoir une idée de la répartition des valeurs de la série autour de la moyenne.

Exemples

- 1 On a représenté ci-dessous le diagramme en boîte de la répartition des salaires dans deux entreprises :



Nous pouvons alors voir que l'intervalle interquartile correspondant à l'entreprise B est plus grand que celui correspondant à l'entreprise A ; les salaires de l'entreprise B sont donc plus hétérogènes que ceux de l'entreprise A.

La médiane des salaires de l'entreprise B est plus élevée que celle de l'entreprise A et correspond presque au Q_3 de l'entreprise A : au moins 50 % des salariés de l'entreprise B gagnent moins d'environ 1 550 € alors que dans l'entreprise A, ils sont au moins 75 % à gagner moins que cette somme.

Les salaires de l'entreprise A semblent donc ici plus homogènes que ceux de l'entreprise B.

- 2 Dans l'entreprise Λ , la moyenne des salaires est $\bar{x}_1 = 1\,500$ € ; l'écart-type est $\sigma_1 = 120$ €. Dans l'entreprise Θ , la moyenne des salaires est $\bar{x}_2 = 1\,700$ € ; l'écart-type est $\sigma_2 = 660$ €.

A priori, il fait bon travailler dans l'entreprise Θ vu que la moyenne des salaires est plus élevée que celle de l'entreprise Λ . Cependant, on remarque que l'écart-type des salaires de l'entreprise Θ est plus grand que l'autre.

Pour l'entreprise Λ , l'intervalle $[\bar{x}_1 - \sigma_1 ; \bar{x}_1 + \sigma_1]$ est $[1\,380 ; 1\,620]$;

Pour l'entreprise Θ , l'intervalle $[\bar{x}_2 - \sigma_2 ; \bar{x}_2 + \sigma_2]$ est $[1\,040 ; 2\,360]$.

On constate que dans l'entreprise Θ , les salaires sont plus dispersés autour du salaire moyen que dans l'entreprise Λ .



On ne peut rien conclure de bien constructif à partir de ces données.

Le fait que les salaires soient plus dispersés dans une entreprise que dans une autre ne doit pas aboutir à une conclusion précise. Les statistiques nous permettent uniquement d'avoir une vue globale d'une situation.