PROBABILITÉS ET STATISTIQUE
3. Statistiques pour décrire, explorer et comparer des données

Carmelo Vaccaro

Université de Paris-Saclay

M1 - Informatique Science des Données (ISD) - Apprentissage
2022/23: premier semestre

Aperçu

Dans ce chapitre, nous présentons des statistiques importantes, notamment la moyenne, la médiane et l'écart-type.

Nous verrons comment interpréter ces statistiques importantes.

Contenu du chapitre

1. Mesures de centre

Mesure de centre

Une mesure de centre est une valeur au centre ou au milieu d'un ensemble de données

Mesures de centre :

- moyenne ;
- médiane ;
- mode.

Moyenne

La **moyenne arithmétique** (ou simplement **moyenne**) est obtenue en additionnant les valeurs et en divisant le total par le nombre de valeurs.

Soient $x_1, \cdots, x_n$ les valeurs numériques qui constituent nos données. Alors la moyenne de ces données est égale à

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

Si on indique avec X l'ensemble de valeurs $\{x_1, \cdots, x_n\}$, alors on peut dénoter $E(X)$ la moyenne de $x_1, \cdots, x_n$ (la moyenne peut aussi être appelée **espérance**).

Exemple de calcul de moyenne

Trouver la moyenne des valeurs

$$27.531 \quad 15.684 \quad 5.638 \quad 2.997 \quad 25.433$$

La moyenne est

$$(27.531 + 15.684 + 5.638 + 27.997 + 25.433) \ / \ 5 =$$

$$102.283 \ / \ 5 = 20.456{,}6$$

Avantages, désavantages de la moyenne

Avantages : elle est relativement fiable, les moyennes des échantillons tirés de la même population ne varient pas autant que d'autres mesures de centre.
Elle prend en compte toutes les valeurs des données

Désavantages : elle est sensible à chaque valeur de données, une valeur extrême peut l'affecter de façon dramatique. On dit que ce n'est pas une mesure de centre *résistante*.

Exemple de non résistance de la moyenne

Supposons d'avoir ces données :

$$1 \quad 1 \quad 3 \quad 5$$

La moyenne est 2,5. Si on rajoute à ces données la valeur 100, la moyenne devient 22, donc l'ajout d'une seule valeur extrême peut changer la moyenne de façon importante.

Exercice

Trouver la moyenne des valeurs

51 63 36 43 34 62 73 39 53 79

Solution

La moyenne est

$$(51 + 63 + 36 + 43 + 34 + 62 + 73 + 39 + 53 + 79) \ / \ 10 = 53{,}3$$

Médiane

La **médiane** est la valeur au milieu lorsque les valeurs des données originales sont classées par ordre de grandeur croissante (ou décroissante).

Comment trouver la médiane : trier les valeurs du plus petit au plus grand, puis :

- si le nombre de valeurs est impair, la médiane est le nombre situé exactement au milieu de la liste ;
- si le nombre de valeurs est pair, la médiane est égale à la moyenne des deux nombres du milieu.

Exemple 1

Trouver la médiane des valeurs

5,40 1,10 0,42 0,73 0,48 1,10 0,66

Après avoir trié les données nous avons

0,42 0,48 0,66 0,73 1,10 1,10 5,40

Comme il y a sept valeurs et que sept est impair, la médiane est le chiffre du milieu, le 4ème chiffre, donc 0,73.

Exemple 2

Trouver la médiane des valeurs

$$5{,}40 \quad 1{,}10 \quad 0{,}42 \quad 0{,}73 \quad 0{,}48 \quad 1{,}10$$

Après avoir trié les données nous avons

$$0{,}42 \quad 0{,}48 \quad 0{,}73 \quad 1{,}10 \quad 1{,}10 \quad 5{,}40$$

Comme il y a six valeurs et que six est pair, la médiane est obtenue en additionnant les 3ème et 4ème valeurs et en divisant la somme par 2, donc $(0{,}73 + 1{,}10) \ / \ 2 = 0{,}915$.

Exercice 1

Trouver la médiane des valeurs

51 63 36 43 34 62 73 39 53 79

Solution

Après avoir trié les données nous avons

$$34 \quad 36 \quad 39 \quad 43 \quad 51 \quad 53 \quad 62 \quad 63 \quad 73 \quad 79$$

Comme il y a dix valeurs et que dix est pair, la médiane est obtenue en additionnant les 5ème et 6ème valeurs et en divisant la somme par 2, (51 + 53) / 2 = 52.

Exercice 2

Trouver la médiane des valeurs

$$27.531 \quad 15.684 \quad 5.638 \quad 27.997 \quad 25.433$$

Solution

Après avoir trié les données nous avons

$$5.638 \quad 15.684 \quad 25.433 \quad 27.531 \quad 27.997$$

Puisqu'il y a cinq valeurs et que cinq est impair, la médiane est le nombre qui se trouve au milieu, le 3ème nombre, donc 25.433.

La médiane est résistante

Supposons d'avoir ces données :

$$1 \quad 1 \quad 3 \quad 5$$

La médiane est $(1 + 3)/2 = 2$. Si on rajoute à ces données la valeur 100, alors les données deviennent

$$1 \quad 1 \quad 3 \quad 5 \quad 100$$

et la moyenne est le nombre au milieu, 3. Donc l'ajout d'une seule valeur extrême n'a pas changé la médiane de façon importante.

Mode

Le **mode** est la valeur qui se produit avec la plus grande fréquence.

Le mode est la seule mesure de centre qui peut être utilisée avec des données qualitatives.

Remarque

Alors que dans une liste de données (numériques) il y a toujours une et une seule moyenne et médiane, il peut y avoir un nombre quelconque de modes (même zéro modes).

Exemple.

a. 5.40 1.10 0.42 0.73 0.48 1.10. Le mode est 1,10.

b. 27 27 27 55 55 55 88 88 99. Il y a deux modes, - 27 et 55.

c. 1 2 3 6 7 8 9 10. Il n'y a aucun mode.

Exercice 1

Trouvez le mode des valeurs

Noir, Rouge, Blanc, Bleu, Noir, Noir, Bleu, Jaune, Bleu, Rouge.

Solution

Trouvez le mode des valeurs

Noir, Rouge, Blanc, Bleu, Noir, Noir, Bleu, Jaune, Bleu, Rouge.

RÉPONSE : il y a deux modes, Noir et Bleu.

Exercice 2

Trouvez le mode des valeurs

22 16 52 21 15 33 43

Solution

Trouvez le mode des valeurs

$$22 \quad 16 \quad 52 \quad 21 \quad 15 \quad 33 \quad 43$$

RÉPONSE : il n'y a pas de mode puisqu'aucune valeur n'apparaît plus d'une fois.

2. Mesures de variation

Concepts clé

L'objet de cette section est la variation dans les données.

En particulier, nous présentons les mesures de variation, telles que l'écart-type, en tant qu'outils d'analyse des données.

L'objectif n'est pas seulement de trouver les valeurs des mesures de variation, mais aussi d'interpréter ces valeurs.

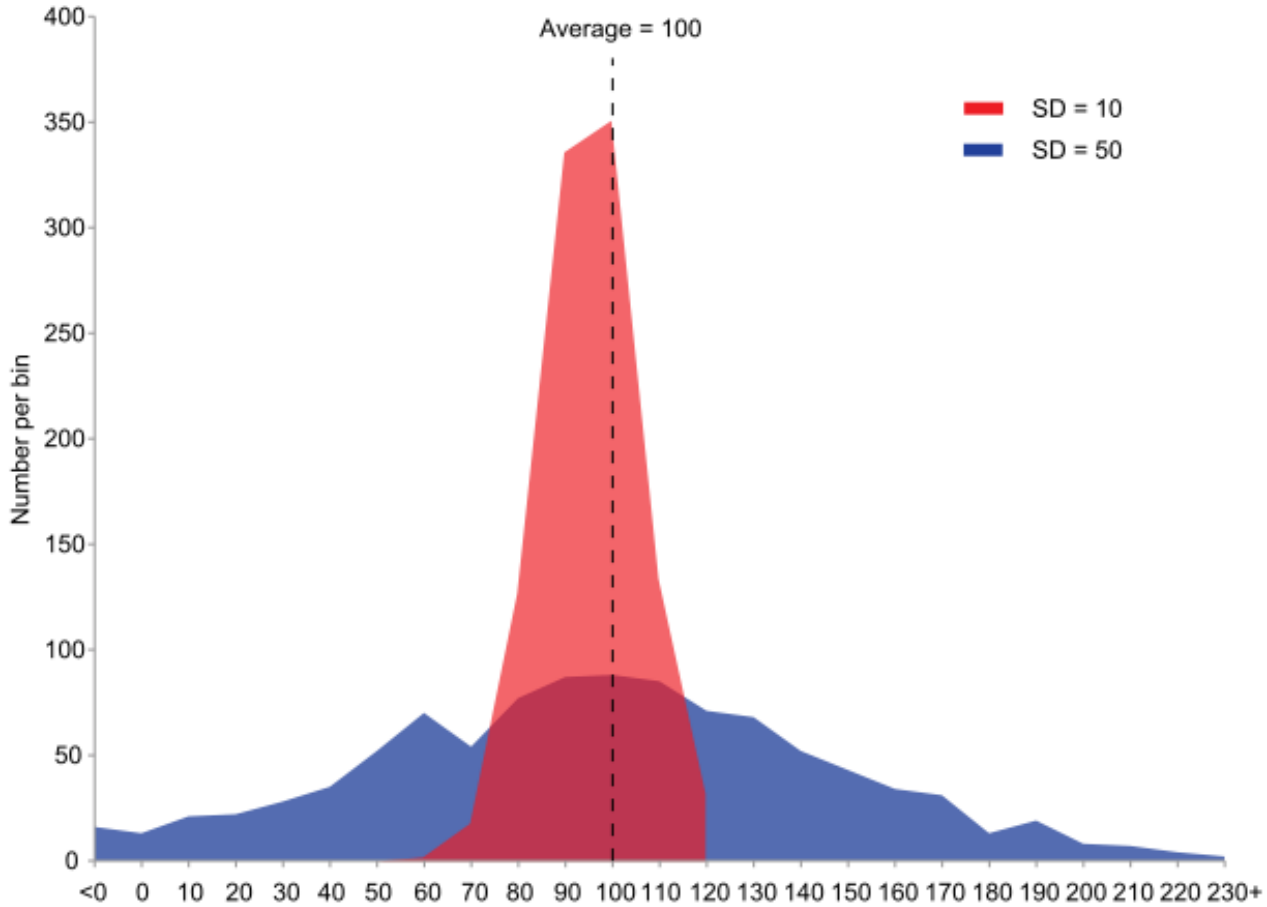
Distributions avec la même moyenne (1/2)

La seule connaissance d'une mesure de centre n'est pas suffisante à décrire les données de manière significative.

En effet, il peut y avoir des distribution avec par exemple la même moyenne mais qui sont très différentes.

Distributions avec la même moyenne (2/2)

Avoir la même moyenne ne signifie pas que les deux échantillons sont similaires. Nous avons besoin d'une mesure de l'écart entre les valeurs.



Mesure de la variation

Une mesure de la variation indique quantifie la distance entre les différents valeurs de l'ensemble des données.

Mesures de la variation :

- étendue ;
- variance ;
- écart type ;
- coefficient de variation.

Étendue

L'**étendue** d'un ensemble de valeurs de données est la différence entre la valeur de données maximale et la valeur de données minimale.

Étendue = (valeur maximale) - (valeur minimale)

L'étendue est très sensible aux valeurs extrêmes et n'est donc pas aussi utile que d'autres mesures de variation.

Exemple

Trouvez l'étendue des valeurs

$$27.531 \quad 15.684 \quad 5.638 \quad 27.997 \quad 25.433$$

Le minimum est 5.638, le maximum est 27.997, l'étendue est donc 27.997 - 5.638 = 22.359.

Exercice

Trouvez l'étendue des valeurs

51 63 36 43 34 62 73 39 53 79

Solution

Trouvez l'étendue des valeurs

$$51 \quad 63 \quad 36 \quad 43 \quad 34 \quad 62 \quad 73 \quad 39 \quad 53 \quad 79$$

RÉPONSE : La valeur minimale est 34, la valeur maximale est 79, l'étendue est donc de 79 - 34 = 45.

Variance

La **variance** d'un ensemble de valeurs est une mesure de la variation des valeurs autour de la moyenne.

La formule de la variance change en fonction du fait que les valeurs soient d'un échantillon d'une population ou de toute la population.

Formules de la variance

Soient $x_1, \cdots, x_n$ des valeurs numériques et soit k leur moyenne, c'est-à-dire

$$k = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Population. Soient $x_1, \cdots, x_n$ des valeurs numériques d'une population. Alors la variance est égale à

$$\frac{\sum_{i=1}^{n} (x_i - k)^2}{n}.$$

Échantillon. Soient $x_1, \cdots, x_n$ des valeurs numériques d'un échantillon. Alors la variance est égale à

$$\frac{\sum_{i=1}^{n} (x_i - k)^2}{n-1}.$$

Remarques sur les formules de la variance (1/3)

La différence entre la formule de la variance pour la population et celle pour un échantillon est qu'avec la population on divise par n, avec l'échantillon par n-1.

Remarques sur les formules de la variance (2/3)

On peut montrer que la variance pour une population est égale à

$$\frac{\sum_{i=1}^{n}(x_i - k)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

et celle pour un échantillon à

$$\frac{\sum_{i=1}^{n}(x_i - k)^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}.$$

Remarques sur les formules de la variance (3/3)

On dénote X l'ensemble de valeurs $x_1, \cdots, x_n$ et on dénote X^2 l'ensemble de valeurs $x_1^2, \cdots, x_n^2$. Nous avons vu précédemment qu'on pouvait dénoter $E(X)$ la moyenne de $x_1, \cdots, x_n$ (dénotée k dans une slide précédente).

Donc on peut dénoter $E(X^2)$ la moyenne de $x_1^2, \cdots, x_n^2$.

Alors d'après les formules de la slide précédente on a que la variance pour une population est égale à

$$E(X^2) - \big(E(X)\big)^2$$

et celle pour un échantillon à

$$\frac{n}{n-1}\Big(E(X^2) - \big(E(X)\big)^2\Big).$$

Remarque sur les arrondis

Lorsqu'on fait des calculs numériques il faut normalement arrondir les valeurs avec un certain nombre de chiffres après la virgule. Dans ce cours on demande normalement un précision de deux chiffres après la virgule.

Il faut suivre la règle suivante : si le résultat final doit avoir une précision de deux chiffres après la virgule, les résultats intermédiaires doivent avoir un niveau de précision de trois chiffre après la virgule (c'est-à-dire un chiffre de plus par rapport à la précision du résultat final).

Exemple sur les arrondis

Calculer avec une précision de deux chiffres après la virgule la valeur de $\sqrt{\frac{10}{3}}$.

D'abord je calcule $\frac{10}{3} \approx 3.333$ (trois chiffres après la virgule) et ensuite je calcule $\sqrt{3.333} \approx 1.83$ (deux chiffres après la virgule car c'est le résultat final).

Par contre si j'arrondis $\frac{10}{3}$ avec deux chiffres j'obtiens 3.33 et $\sqrt{3.33} \approx 1.82$, qui n'est pas correcte.

Exemple

Trouver (avec une précision de deux chiffres après la virgule) la variance de population des valeurs

$$1 \quad 3 \quad 14$$

Soit $X = \{1, 3, 14\}$, donc $X^2 = \{1, 9, 196\}$.

On a $E(X) = (1 + 3 + 14)/3 = 6$, $E(X^2) = (1 + 9 + 196)/3 \approx 68{,}667$, donc la variance est égale à

$$E(X^2) - \left(E(X)\right)^2 = 68{,}667 - 6^2 \approx 32{,}667 \approx 32{,}67.$$

Maintenant on suppose que les valeurs viennent d'un échantillon. Calculer la variance.
Elle est égale à $\frac{n}{n-1}\left(E(X^2) - \left(E(X)\right)^2\right)$, avec $n = 3$, donc à

$$\frac{3}{2} \times 32{,}667 \approx 49.$$

Exercice

Trouvez la variance de population et d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

Solution

Trouvez la variance de population et d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

RÉPONSE : Soit $X = \{12, 13, 9, 10\}$, donc $X^2 = \{144, 169, 81, 100\}$.

On a $E(X) = (12 + 13 + 9 + 10)/4 = 11$,
$E(X^2) = (144 + 169 + 81 + 100)/4 = 123{,}5$, donc la variance de population est égale à

$$E(X^2) - \left(E(X)\right)^2 = 123{,}5 - 11^2 = 2{,}5.$$

La variance d'échantillon est égale à

$$\frac{4}{3} \times 2{,}5 \approx 3{,}33.$$

Remarques

- La variance mesure la variation de toutes les valeurs par rapport à la moyenne.
- La variance n'est jamais négative. Elle est zéro si et seulement si toutes les valeurs des données sont identiques.
- La valeur de la variance peut augmenter de façon importante avec l'inclusion d'une ou plusieurs valeurs aberrantes.
- L'unité de mesure de la variance est le carré de celle des données (par exemple si les données sont exprimées en mètres, la variance est exprimées en mètres carrés).

Pourquoi divise-t-on par $n-1$ dans la formule de la variance pour un échantillon ?

La raison pour laquelle on divise par $n-1$ dans la formule de la variance pour un échantillon est qu'ainsi la variance de l'échantillon est un estimateur sans biais de la variance de la population.

En effet, supposons d'avoir les données d'une population et de calculer la variance avec la formule vue plus haut (où on divise par n). Puis on considère tous les échantillons de la population et on calcule la variance de chacun d'eux avec la formule où on divise par $n-1$.

Alors la moyenne de ces variances (pour tous les échantillons) est exactement la variance de la population. Par contre si on calculait la variance d'un échantillon en divisant par n et pas par $n-1$, alors la moyenne des variances de tous les échantillons ne serait pas égale à la variance de la population, mais elle serait plus petite.

Écart type

L'**écart type** est la racine carrée de la variance.

Soient $x_1, \cdots, x_n$ des valeurs numériques et soit k leur moyenne.

Population. Soient $x_1, \cdots, x_n$ des valeurs numériques d'une population. Alors l'écart type est égal à

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - k)^2}{n}}.$$

Échantillon. Soient $x_1, \cdots, x_n$ des valeurs numériques d'un échantillon. Alors l'écart type est égal à

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - k)^2}{n-1}}.$$

Dimension de l'écart type

On a vu que l'unité de mesure de la variance d'un ensemble de données est le carré de l'unité de mesure des données. Comme l'écart type est la racine carrée de la variance, son unité de mesure est la même que celle des données.

Remarques sur les formules de l'écart type

On dénote X l'ensemble de valeurs $x_1, \cdots, x_n$ et on dénote X^2 l'ensemble de valeurs $x_1^2, \cdots, x_n^2$. Comme l'écart type est la racine carrée de la variance alors l'écart type pour une population est égale à

$$\sqrt{E(X^2) - \left(E(X)\right)^2}$$

et ce pour un échantillon à

$$\sqrt{\frac{n}{n-1}\left(E(X^2) - \left(E(X)\right)^2\right)}.$$

Exemple

Trouver l'écart type de population des valeurs

$$1 \quad 3 \quad 14$$

Comme l'écart type est la racine carrée de la variance et comme on avait trouvé que la variance était égale à 32,667, alors l'écart type est égal à $\sqrt{32{,}667} \approx 5{,}72$.

Maintenant on suppose que les valeurs viennent d'un échantillon. Calculer l'écart type.

Comme l'écart type est la racine carrée de la variance et comme on avait trouvé que la variance était égale à 49, alors l'écart type est égal à $\sqrt{49} = 7$.

Exercice

Trouvez l'écart type de population et d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

Solution

Trouvez l'écart type de population et d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

RÉPONSE : Soit $X = \{12, 13, 9, 10\}$, donc $X^2 = \{144, 169, 81, 100\}$.

Comme l'écart type est la racine carrée de la variance et comme on avait trouvé que la variance était égale à 2.5, alors l'écart type de population est égale à $\sqrt{2.5} \approx 1.58$.

L'écart type d'échantillon est égal à la racine carrée de la variance d'échantillon, qui était égale à $3{,}333 \approx 3{,}33$, donc il est égal à 1,83.

Coefficient de variation

Le **coefficient de variation** d'un ensemble de données d'un échantillon ou d'une population, montre de combien varient des données par rapport à la moyenne. Il est normalement exprimé en pourcentage.

Soit k la moyenne des données et s l'écart type. Alors le coefficient de variation est égal à

$$\frac{s}{|k|} \times 100\%.$$

Comme le coefficient de variation est égal à l'écart type normalisée avec la moyenne, il permet de comparer les écarts type entre deux échantillons qui ont des moyennes différentes.

Exemple

Trouver le coefficient de variation d'échantillon des valeurs

$$1 \quad 3 \quad 14$$

La moyenne est 6, l'écart type d'échantillon est 7, donc le coefficient de variation est $\frac{6}{7} \times 100\% = 86\%$.

Remarque sur la précision du coefficient de variation

Comme le coefficient de variation est exprimé en pourcentage, alors un niveau de précision de deux chiffres après la virgule correspond à un niveau de précision de zéro chiffres après la virgule.

Par exemple dans la slide précédente on avait $\frac{6}{7} = 0{,}86$ (avec une précision de deux chiffres après la virgule), qui en pourcentage devient 86%, donc zéro chiffres après la virgule.

Exercice

Trouvez le coefficient de variation d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

Solution

Trouvez le coefficient de variation d'échantillon des valeurs

$$12 \quad 13 \quad 9 \quad 10$$

RÉPONSE : La moyenne est 11, l'écart type d'échantillon est 1,826, donc le coefficient de variation est

$$\frac{1,826}{11} \times 100\% = 16{,}6\%.$$

3. Mesures de position relative et boîtes à moustaches

Concepts clé

Dans cette section nous verrons les **mesures de position relative**, qui sont des nombres indiquant la position de la valeur d'une donnée par rapport aux autres dans un ensemble de données.

Ces mesures de position relative peuvent être utilisées pour comparer des valeurs provenant de différents ensembles de données, ou pour comparer des valeurs au sein d'un même ensemble de données.

Le concept le plus important est la **cote Z**. Nous aborderons également les **centiles** et les **quartiles**, ainsi qu'un nouveau graphique statistique appelé **boîte à moustache**.

Cote Z

La **cote Z** d'une donnée est le nombre d'écarts type que la valeur de la donnée est supérieure ou inférieure à la moyenne.

Soient $X = \{x_1, \cdots, x_n\}$ un ensemble de données, soient k la moyenne et s l'écart type (de population ou échantillon en fonction du fait que X soit les données d'une population ou d'un échantillon). Alors la cote Z de x_i est égale à

$$\frac{x_i - k}{s}.$$

Cote Z : remarques

Supposons que x_i est égale exactement à la moyenne, c'est-à-dire $x_i = k$. Alors la cote Z de x_i est égale à 0.

Supposons maintenant que la moyenne est égale à 10, l'écart type à 5 et que $x_i = 25$. Alors d'après la formule la cote Z de x_i est égale à $\frac{25-10}{5} = 3$, donc x_i est supérieur à la moyenne de trois fois l'écart type.

Toujours avec la moyenne égale à 10 et l'écart type à 5 supposons que $x_i = 5$. Alors la cote Z de x_i est égale à $\frac{5-10}{5} = -1$, donc x_i est inférieur à la moyenne d'une fois l'écart type.

Il faut remarquer que la cote Z d'une valeur est positive si et seulement si la valeur est supérieure à la moyenne.

Exemple

Supposons d'avoir des données dont la moyenne est 173,58 et l'écart type est 7,67. Calculer la cote Z d'une donnée égale à 193,5.

D'après la formule la cote Z est

$$\frac{193{,}5 - 173{,}58}{7{,}67} = 2{,}60.$$

Exercice

Calculer la cote Z d'une donnée égale à 107,5 dans un ensemble de données de moyenne 78,27 et écart type 11,94.

Solution

Calculer la cote Z d'une donnée égale à 107,5 dans un ensemble de données de moyenne 78,27 et écart type 11,94.

La cote Z est

$$\frac{107{,}5 - 78{,}27}{11{,}94} = 2{,}45.$$

Une interprétation de la cote Z

La cote Z permet de comparer l'éloignement de la moyenne dans des ensembles de données différentes.

Assumons que la distribution de tailles des hommes a une moyenne de 173,58 cm et un écart type de 7,67 cm. Assumons aussi que la distribution de poids des hommes a une moyenne de 78,27 kg et un écart type de 11,94 kg. Quelle valeur est plus extrême, une taille de 193,5 cm ou un poids de 107,5 kg ?

On a vu dans l'exemple précédent que la cote Z pour la taille de 193,5 cm était de 2,60, tandis que dans l'exercice précédent on a vu que la cote Z pour le poids de 107,5 kg était de 2,45. Cela veut dire qu'il est moins insolite qu'un homme pèse 107,5 kg plutôt qu'il soit 193,5 cm de taille.

Une justification avec le coefficient de variation

Utilisons le coefficient de variation pour justifier pourquoi un poids de 107,5 kg est moins insolite qu'une taille de 193,5 cm.

Le coefficient de variation pour les tailles est de $7{,}67/173{,}58 = 4\%$, alors que le coefficient de variation pour les poids est de $11{,}94/78{,}27 = 15\%$. Donc les poids ont une variation presque 4 fois plus grande que les tailles et cela justifie pourquoi un poids 107,5 kg est moins insolite qu'une taille de 193,5 cm malgré à première vue on aurait l'impression qu'un poids de 107,5 kg est plus extrême qu'une taille de 193,5 cm.

Valeurs inhabituelles

Dans un ensemble de données nous définissons une valeur *inhabituelle* si sa cote Z est supérieure à 2 ou inférieure à -2, c'est-à-dire si sa distance de la moyenne est supérieure à 2 écarts type.

Centiles

Le concept de centile généralise celui de médiane : la médiane d'un ensemble de données S est une valeur qui sépare la moitié des valeurs les plus petites de la moitié des valeurs les plus grandes.

Supposons d'avoir un ensemble de données numériques que nous appelons S. Le *centiles* de S sont 99 valeurs (pas nécessairement différents), notées $P_1, P_2, \cdots, P_{99}$ tels que pour $i = 1, \cdots, 99$ le centile P_i sépare le $i\%$ des valeurs les plus petites du $(100 - i)\%$ des valeurs les plus grandes de S.

Les centiles de S sont (comme la médiane) soit des valeurs de S, soit la somme divisée par 2 d'une valeur de S et de celle qui la suit si on trie les valeurs de S.

En outre pour chaque valeur d'un ensemble de donnée on peut déterminer son centile.

Centiles : exemples

Soit $S = \{1, 2, \cdots, 99\}$. Alors

$$P_1 = 1,\ P_2 = 2, \cdots, P_{99} = 99.$$

Soit $S = \{1, 2, \cdots, 99, 100\}$. Alors

$$P_1 = 1{,}5,\ P_2 = 2{,}5, \cdots, P_{99} = 99{,}5.$$

Soit $S = \{1, 2, \cdots, 49, 50\}$. Alors

$$P_1 = 1,\ P_2 = 1{,}5,\ P_3 = 2,\ P_4 = 2{,}5, \cdots, P_{98} = 49{,}5,\ P_{99} = 50.$$

Soit $S = \{1, 2, \cdots, 49\}$. Alors

$$P_1 = P_2 = 1,\ P_3 = P_4 = 2, \cdots, P_{98} = P_{99} = 49.$$

Centiles : remarques

- Comme le montre le quatrième exemple de la slide précédente, les centiles peuvent ne pas être tous distincts.

- Le centile P_{50} est la médiane.

- Il existe différentes façons différentes de calculer les centiles d'une liste de données. Si la liste de données est assez grande tous donnent les mêmes valeurs. Au contraire les différentes façons peuvent donner des résultats différentes.

Quartiles

Supposons d'avoir une ensemble de données numériques que nous appelons S. Le *quartiles* de S sont 3 valeurs, notées Q_1, Q_2, Q_3 égaux respectivement aux centiles P_{25}, P_{50} et P_{75}.

Donc $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$. Comme P_{50} est la médiane de S, alors Q_2 est la médiane.

Quelques autres statistiques

Écart interquartile (ou EI) $= Q_3 - Q_1$

Écart semi-interquartile $= (Q_3 - Q_1)/2$

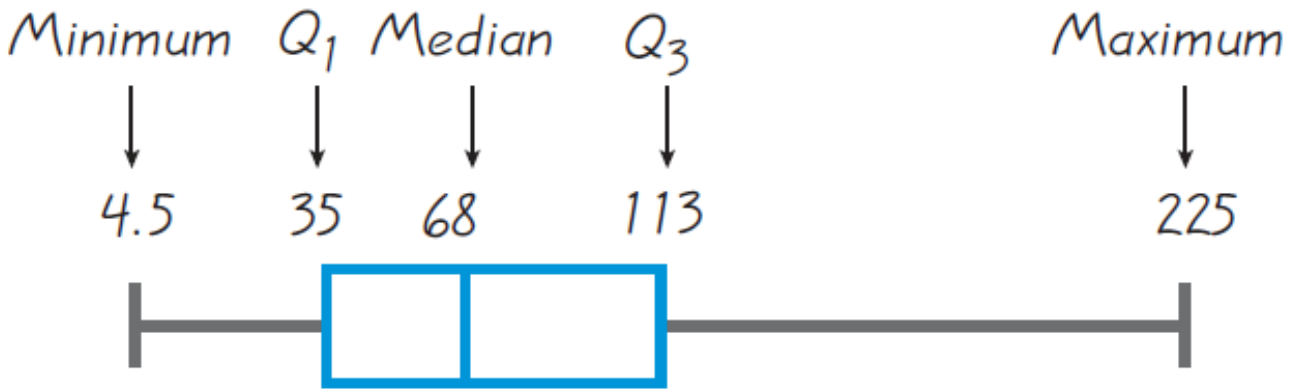
Mi-quartile $= (Q_3 + Q_1)/2$

Intervalle de centile 10-90 $= P_{90} - P_{10}$

Résumé en cinq nombres et boîte à moustaches

Pour un ensemble de données, le *résumé en 5 nombres* est l'ensemble des 5 nombres suivants : le premier quartile Q_1, la médiane (ou le deuxième quartile Q_2), le troisième quartile Q_3, et la valeur maximale.

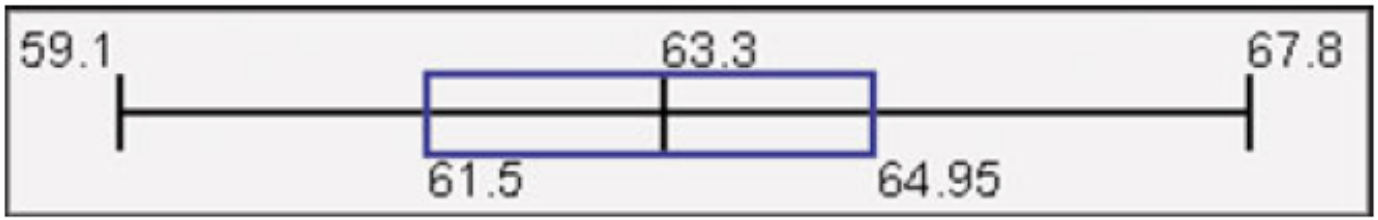
Pour un ensemble de données, la *boîte et moustaches* est un graphique constitué d'une ligne allant de la valeur minimale à la valeur maximale, et d'une boîte avec des lignes tracées au niveau du premier quartile, Q_1, de la médiane et du troisième quartile, Q_3.



Boites à moustaches et symétrie

Les boites à moustaches nous donnent des informations sur la distribution et la dispersion des données.

La première des deux boites à moustaches ci-dessous vient d'une distribution de données qui est symétrique, alors que la deuxième vient d'une qui est inclinée vers la droite.



Valeurs aberrantes

Une *valeur aberrante* est une valeur qui se situe très loin de la grande majorité des autres valeurs d'un ensemble de données.

Une valeur aberrante peut avoir un effet dramatique sur la moyenne et l'écart-type. Il peut aussi avoir un effet dramatique sur l'échelle de l'histogramme, de sorte que la véritable nature de la distribution soit totalement masquée.

C'est la raison pour laquelle souvent on fait un histogramme d'un ensemble de données après avoir retiré les valeurs aberrantes.

Valeurs aberrantes : définition

Nous pouvons définir les valeurs aberrantes de la façon suivante (rappelons que l'écart interquartile, noté EI, est égal à $Q_3 - Q_1$) :

Une valeur de donnée est une valeur aberrante si elle est supérieure à Q_3 ou inférieure à Q_1 d'une quantité supérieure à 1,5 EI.

Cela veut dire qu'une valeur est aberrante si elle est supérieure à

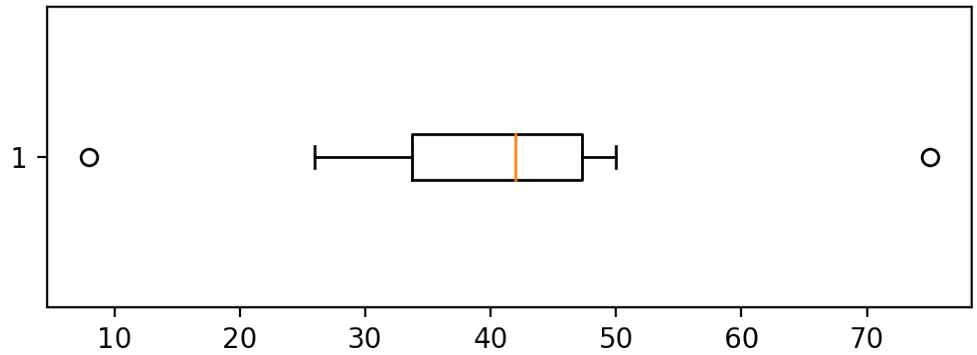
$$Q_3 + 1{,}5EI = Q_3 + 1{,}5(Q_3 - Q_1) = 2{,}5Q_3 - 1{,}5Q_1$$

ou inférieure à

$$Q_1 - 1{,}5EI = Q_1 - 1{,}5(Q_3 - Q_1) = 2{,}5Q_1 - 1{,}5Q_3.$$

Boîte à moustaches modifiée

Une *boîte à moustaches modifiée* est construite à partir de la boîte à moustaches en rajoutant les valeurs aberrantes.



Les deux cercles vides représentent des valeurs aberrantes.

Fin.