*Mathématique : Culture, Société et Technique*
*4ème secondaire*

# Statistiques et Corrélation

| | |
|---|---|
| Distribution à un caractère : | diagramme à tige et feuilles |
| Mesure de tendance centrale : | mode, médiane, moyenne, moyenne pondérée |
| Mesures de dispersion : | étendue, diagramme de quartiles, écart moyen |
| Mesures de position : | rang centile |
| Distribution à deux caractères : | corrélation, tableau à double entrée, nuage de points |
| Interprétation quantitative : | coefficient de corrélation, droite de régression |
| Interprétation de la corrélation linéaire | |

## Corrigé

# Mesures de tendance centrale

Les mesures de tendance centrale servent à décrire le centre d'une distribution ordonnée et la position des données de la distribution par rapport à ce centre. Le **mode**, la **médiane** et la **moyenne** sont des mesures de tendance centrale.

**Mode** La mesure qui indique le centre de concentration d'une distribution.

Notation : Mo

• *Dans une distribution* ***ordonnée*** :

Le mode est la valeur qui revient le plus souvent.

Exemple: 2, 7, 7, 8, 8, 8, 9, 9, 10

Mo = 8

• *Dans une distribution de données condensées* :

Le mode est la valeur ou la modalité ayant l'effectif le plus élevé.

ex : Familles

| nombre d'enfants $x_i$ | effectif $f_i$ |
|---|---|
| 1 | 6 |
| 2 | 16 |
| 3 | 12 |
| 4 | 10 |
| 5 | 9 |
| total | 53 |

Mo = 2

• *Dans une distribution de données groupées en classes*

La classe ayant l'effectif le plus élevé est qualifiée de **classe modale**.

Le milieu de la classe modale donne une estimation de la valeur du mode.

ex : Chenil

| taille des chiens (cm) | effectif $f_i$ | milieu des classes $m_i$ |
|---|---|---|
| $[20,40[$ | 18 | 30 |
| $[40,60[$ | 19 | 50 |
| $[60,80[$ | 13 | 70 |
| **classe modale** → $[80,100[$ | 20 ← | 90 |
| $[100,120[$ | 10 | 110 |
| total | 80 | |

Mo = 90

**Médiane** : La mesure qui indique le centre de position d'une distribution.

Notation : Méd

• *Dans une distribution **ordonnée*** :

a) S'il y a un nombre impair de données, c'est la donnée du centre :

Exemple: 2, 7, 7, 8, 8, 8, 9, 9, 10

Il y a 9 données ÷ 2 = 4,5 → la médiane est la $5^{ème}$ donnée

la $5^{ème}$ donnée est 8 . réponse : méd : 8

b) S'il y a un nombre pair de données, c'est la moyenne des données du centre.

Exemple : 0, 1, 4, 8

Il y a 4 données ÷ 2 = 2 la médiane est la moyenne entre la 2ème et la 3ème donnée

$$\frac{1 + 4}{2} = 2{,}5$$

réponse : méd : 2,5

• *Dans une distribution de données groupées en classes*

La classe comportant la médiane est qualifiée de **classe médiane**.

Le milieu de la classe médiane donne une estimation de la valeur de la médiane.

ex : Chenil

| taille des chiens (cm) | effectif $f_i$ | cumulatif |
|---|---|---|
| $[20,40[$ | 18 | 18 |
| $[40,60[$ | 19 | 37 |
| **classe médiane** → $[60,80[$ | 13 | 50 |
| $[80,100[$ | 20 | 70 |
| $[100,120[$ | 10 | 80 |
| total | 80 | |

Il y a 37 chiens qui ont une taille inférieure à 60 cm.

La taille du 40ème et du 41ème chien est dans cette classe

Il y a 50 chiens qui ont une taille inférieure à 80 cm.

La médiane est la moyenne entre la 40ème et la 41ème donnée. Comme ces données sont dans la classe $[60,80[$, nous dirons que la médiane correspond au milieu de cette classe.

Méd = 70

**Moyenne** : La mesure qui indique le centre d'équilibre d'une distribution.

Notation : $\bar{x}$

• *Dans une distribution ordonnée* :

$$\text{Moyenne} = \frac{\left(\text{somme des valeurs}\right)}{\text{nombre de données}}$$

ex : Voici une distribution ordonnée comportant 15 données :

2, 2, 2, 3, 3, 4, 5, 6, 7, 8, 8, 8, 8, 10, 11.

$$\bar{x} = \frac{(2+2+2+3+3+4+5+6+7+8+8+8+8+10+11)}{15} = 5{,}8$$

• *Dans une distribution de données condensées* :

$$\text{Moyenne} = \frac{\left(\begin{array}{c}\text{somme des produits} \\ \text{des valeurs par leur effectif}\end{array}\right)}{\text{nombre de données}}$$

ex : Familles

| nombre d'enfants $x_i$ | effectif $f_i$ | $x_i \cdot f_i$ |
|---|---|---|
| 1 | 6 | 6 |
| 2 | 16 | 32 |
| 3 | 12 | 36 |
| 4 | 10 | 40 |
| 5 | 9 | 45 |
| total | 53 | 159 |

$$\bar{x} = \frac{159}{53} = 3$$

• *Dans une distribution de données groupées en classes*

$$\text{Moyenne} = \frac{\left(\begin{array}{c}\text{somme des produits des milieux}\\ \text{des classes par leur effectif}\end{array}\right)}{\text{nombre de données}}$$

ex : Chenil

| taille des chiens (cm) | effectif $f_i$ | milieu des classes $m_i$ | $m_i \bullet f_i$ |
|---|---|---|---|
| $[20,40[$ | 18 | 30 | 540 |
| $[40,60[$ | 19 | 50 | 950 |
| $[60,80[$ | 13 | 70 | 910 |
| $[80,100[$ | 20 | 90 | 1800 |
| $[100,120[$ | 10 | 110 | 1100 |
| total | 80 | | 5300 |

$$\text{Moyenne} = \frac{5300}{80} = 66.25$$

**Moyenne pondérée** :

La moyenne d'un certain nombre de valeurs n'ayant pas toutes la même importance.

Ex : Cours de géographie

| Étape | Note sur 100 | Pondération(%) |
|---|---|---|
| 1 | 75 | 20 = 0,20 |
| 2 | 72 | 30 = 0,30 |
| 3 | 88 | 50 = 0,50 |

$$\bar{x} = 75 \bullet 0{,}20 + 72 \bullet 0{,}30 + 88 \bullet 0{,}50 = 80{,}6$$

# Mesures de dispersion

Les mesures de dispersion servent à décrire l'étalement ou la concentration des données d'une distribution. L'**étendue** est une mesure de dispersion.

**Étendue** : mesure qui indique jusqu'à quel point les données sont regroupées ou éloignées les unes des autres dans une distribution.

Notation : e

• *Dans une distribution ordonnée* :

L'étendue est l'écart entre la donnée la plus élevée et la donnée la moins élevée.

ex : Voici une distribution ordonnée comportant 15 données :

2, 2, 2, 3, 3, 4, 5, 6, 7, 8, 8, 8, 8, 10, 11.

e = 11 – 2 = 9

• *Dans une distribution de données condensées* :

L'étendue est l'écart entre la donnée la plus élevée et la donnée la moins élevée.

ex : Familles

| nombre d'enfants $x_i$ | effectif $f_i$ |
|---|---|
| 1 | 6 |
| 2 | 16 |
| 3 | 12 |
| 4 | 10 |
| 5 | 9 |
| total | 53 |

e = 5 – 1 = 4

• *Dans une distribution de données groupées en classes*

L'étendue est l'écart entre la borne supérieure de la classe la plus élevée et la borne inférieure de la classe la moins élevée.

ex : Chenil

| taille des chiens (cm) | effectif $f_i$ |
|---|---|
| $[20,40[$ | 18 |
| $[40,60[$ | 19 |
| $[60,80[$ | 13 |
| $[80,100[$ | 20 |
| $[100,120[$ | 10 |
| total | 80 |

e = 120 – 20 = 100

## Rappel : Diagramme de quartiles

Les quartiles sont les valeurs qui partagent une distribution ordonnée en quatre sous-ensembles comprenant le même nombre de données appelés « quarts ».

Le **diagramme de quartiles** permet d'analyser la dispersion ou la concentration d'un ensemble de données ou de comparer deux ensembles de données de même nature.

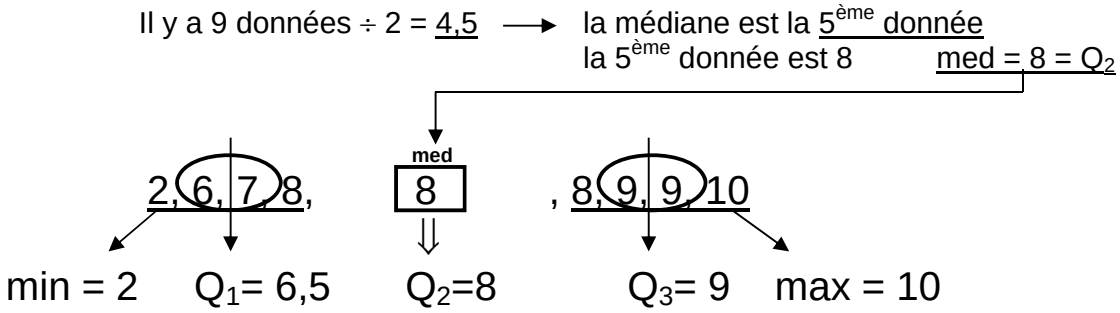
# Calcul des quartiles :

**Deuxième quartile** : médiane de toutes les données notation : $Q_2$ = med

**Premier quartile** : médiane de toutes les données inférieures à $Q_2$
notation : $Q_1$

**Troisième quartile** : médiane de toutes les données supérieures à $Q_2$
notation : $Q_3$

Exemple : 2, 6, 7, 8, 8, 8, 9, 9, 10




**Premier quart** : sous-ensemble formé par les données avant $Q_1$ : {2, 6}

**Deuxième quart** : sous-ensemble formé par les données entre $Q_1$ et $Q_2$ : {7, 8}

**Troisième quart** : sous-ensemble formé par les données entre $Q_2$ et $Q_3$ : {8, 9}

**Quatrième quart** : sous-ensemble formé par les données après $Q_3$ : {9, 10}

Il y a le **même nombre de données** dans **chaque quart**. Les quartiles ne font pas partie des quarts.

Procédure pour **ordonner une série de données** avec la **calculatrice à affichage graphique**

## ➢ Étape 1

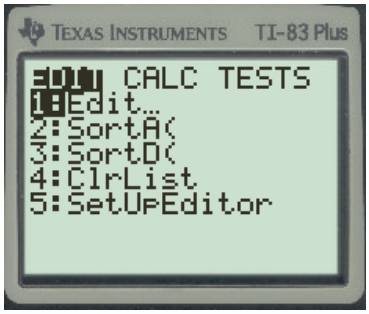
Mettre en fonction la calculatrice en appuyant sur le bouton **ON** .

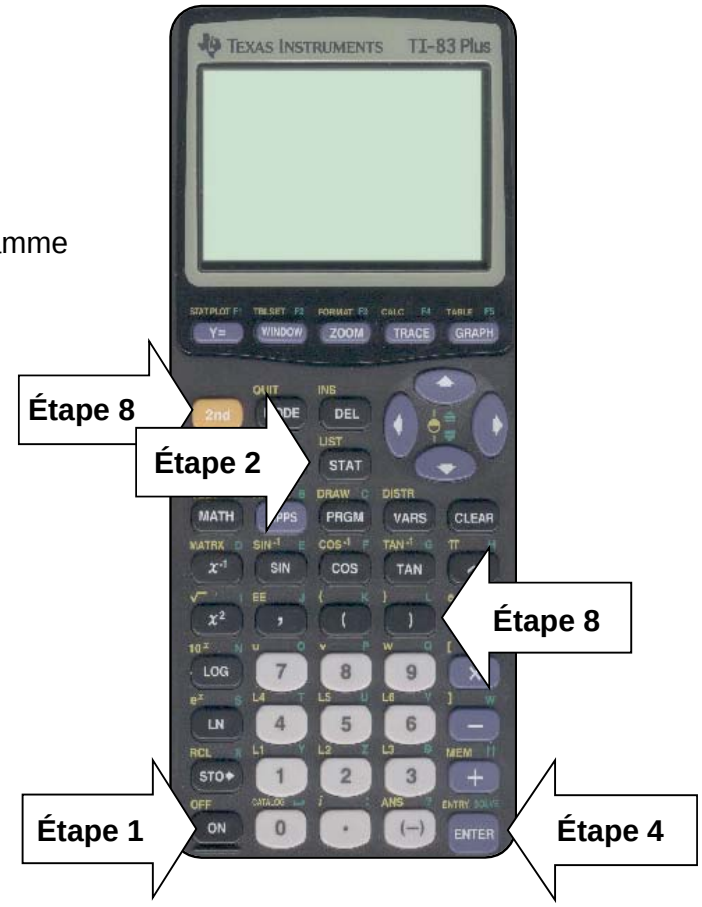
## ➢ Étape 2

Appuyer sur la touche **STAT**

## ➢ Étape 3

À l'aide des flèches, sélectionner le programme EDIT et le sous programme 1: Édit



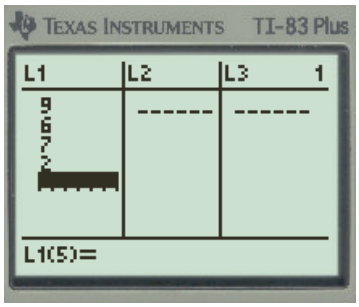

## ➢ Étape 4

Appuyer sur la touche **ENTER**

## ➢ Étape 5

Sous $L_1$, entrer les données en appuyant sur **ENTER** ou sur la flèche qui pointe vers le bas après chacune.

Exemple :



## ➢ Étape 6

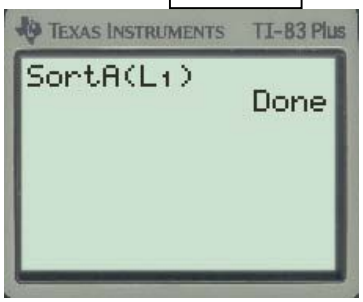
Appuyer sur la touche **STAT** comme à l'étape 2. À l'aide des flèches, sélectionner le programme EDIT et le sous programme 2: Sort A(


## ➢ Étape 7

Appuyer sur la touche **ENTER** et écrire $L_1$ en appuyant sur les touches 2nd et 1 . Ensuite fermer la parenthèse

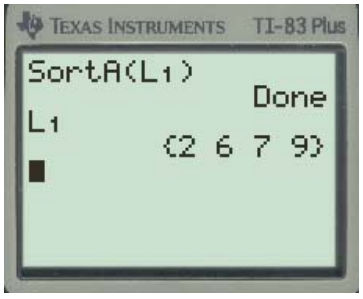

## ➢ Étape 8

Appuyer sur la touche **ENTER** , ceci apparaîtra :




## ➢ Étape 9

Écrire $L_1$ et la liste ordonnée va apparaître. Si la liste est longue, se déplacer avec les flèches pour voir toutes les données.



Procédure pour **calculer les quartiles** avec la **calculatrice à affichage graphique**

## ➢ Étape 1

Mettre en fonction la calculatrice en appuyant sur le bouton [ON] .

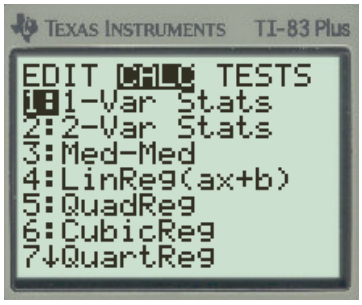
## ➢ Étape 2

Appuyer sur la touche [**STAT**]

## ➢ Étape 3

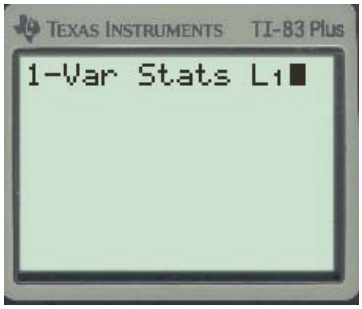
À l'aide des flèches, sélectionner le programme CALC et le sous programme 1: 1-Var Stats

Exemple :



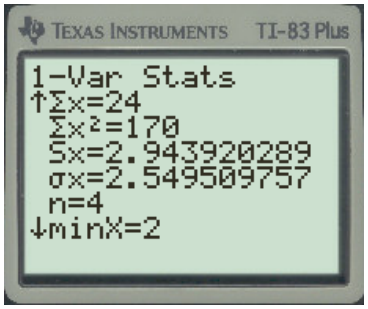

## ➢ Étape 4

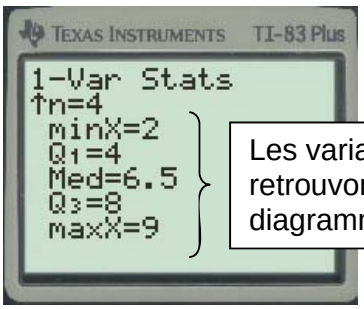
Appuyer sur la touche [**ENTER**] et écrire $L_1$ en appuyant sur les touches [$2^{nd}$] et [1] .




## ➢ Étape 5

Appuyer sur la touche [**ENTER**] . Le calcul de plusieurs variables statistiques va apparaître à l'écran, se déplacer avec les flèches pour voir celles qui nous intéressent.




Les variables que nous retrouvons dans les diagrammes de quartiles.

Procédure pour **tracer les diagrammes de quartiles** avec la **calculatrice à affichage graphique**

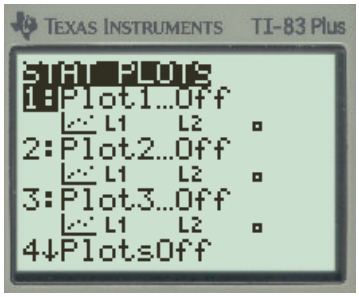
Écrire la série de données sous $L_1$, en suivant les étapes 1 à 5 de la page 10.

## ➢ Étape 1

Appuyer sur les touches [2nd] et [Y=]

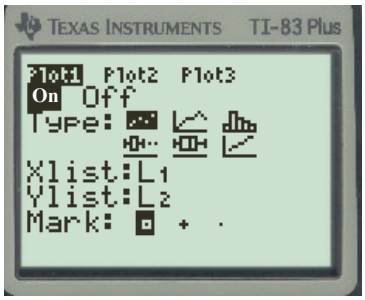
## ➢ Étape 2

Sélectionner 1



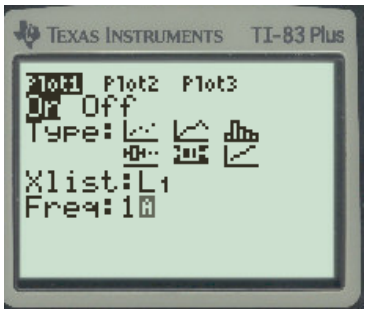

## ➢ Étape 3

Appuyer sur la touche **[ENTER]**. Pour le graphique 1 (Plot 1) sélectionner [On] avec les flèches.




## ➢ Étape 4

Se servir des flèches et de [Enter] pour sélectionner votre écran comme celui-ci.

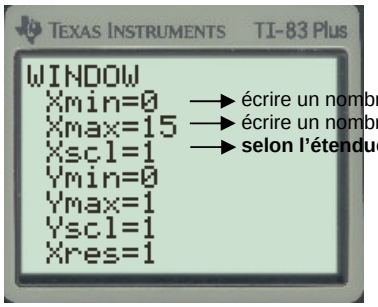


## ➢ Étape 5

S'assurer que les autres graphiques de [2nd] [y =] sont à [off] et que l'écran de [y =] est vide.

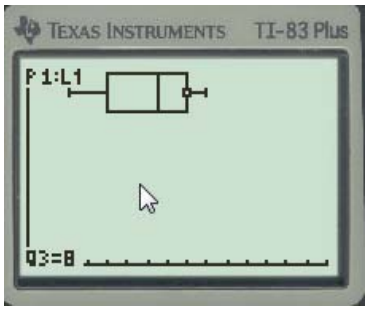
## ➢ Étape 6

Appuyer sur la touche [WINDOW] pour ajuster la graduation des axes comme ceci :




## ➢ Étape 7

Appuyer sur la touche [GRAPH]. La fonction [TRACE] permet de lire le graphique, en déplaçant le curseur avec les flèches, on voit le minimum, les quartiles et le maximum.

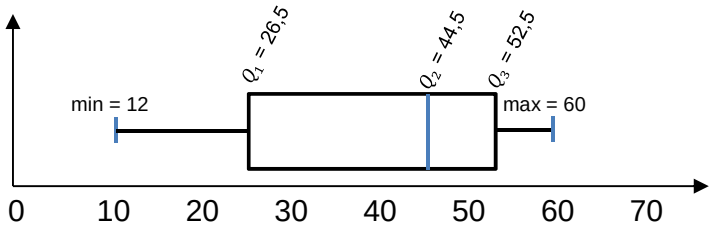


Exercices :

1. Construis un diagramme de quartiles avec la série de données suivante :

   12, 15, 16, 18, 18, 25, 28, 30, 34, 34, 40, 44, 45, 48, 50, 50, 51, 52, 53, 53, 55, 56, 58, 60.

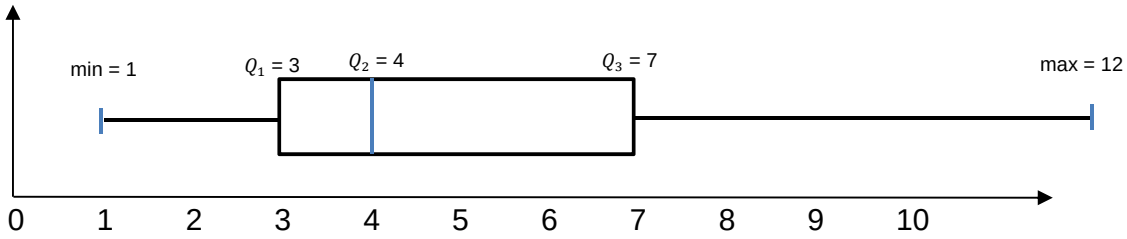
   Reproduis le graphique obtenu en identifiant la valeur des variables statistiques.




2. Construis un diagramme de quartiles avec la série de données suivante :

   2, 8, 12, 6, 4, 2, 4, 10, 9, 2, 4, 4, 6, 10, 1, 8, 5, 3, 2, 4, 10, 3, 5, 7, 3, 2, 3, 7, 6, 4.

   Reproduis le graphique obtenu en identifiant la valeur des variables statistiques.




3. Trouve la **médiane**, **$Q_1$**, **$Q_3$**, **ei** et **e** de chaque série :

| | | min | $Q_1$ | $Q_2$ | $Q_3$ | max | ei | e |
|---|---|---|---|---|---|---|---|---|
| a) | 18, 18, 28, 22, 24, 29, 19, 25, 27, 27, 26. | 18 | 19 | 25 | 27 | 29 | 8 | 11 |
| b) | 14, 1, 5, 28, 10, 31, 28, 30, 32, 12, 19, 25. | 1 | 11 | 22 | 29 | 32 | 18 | 31 |
| c) | 1, 2, 19, 22, 24, 25, 26, 27, 27, 50, 75. | 1 | 19 | 25 | 27 | 75 | 8 | 74 |

## Distribution à un caractère

Une distribution à un caractère correspond à l'ensemble des données recueillies au cours d'une étude statistique portant sur un seul caractère (une seule variable statistique).

**Diagramme à tige et feuilles**
Le diagramme à tige et à feuille est utilisé pour représenter les données d'une ou de deux distributions (à un caractère) qui sont disposées d'un ou de deux côtés d'une colonne, appelée tige. Dans un tel diagramme :

- chaque ligne est associée à une classe ;
- chaque donnée est décomposée en deux parties se trouvant sur une même ligne : la partie constituée de ses premiers chiffres formant la tige et la partie constituée de ses derniers chiffres formant une feuille.

Ex :
a) Une seule distribution :

| Titre | |
|---|---|
| 0 | 2 – 5 |
| 1 | 0 – 1 |
| 2 | 3 – 9 – 9 |
| 3 | |
| 4 | 4 – 8 |

début des nombres formant les résultats

fin des nombres formant les résultats

Les résultats sont : 2, 5, 10, 11, 23, 29, 29, 44, 48

b) Deux distributions :

| Nom de la 1ère distribution | | Nom de la 2ème distribution |
|---|---|---|
| 9 – 8 – 8 – 7 – 2 | 0 | 1 – 2 – 6 – 7 - 9 |
| 8 – 7 – 5 – 4 – 1 | 1 | 0 – 3 – 4 – 8 - 9 |
| 6 – 3 – 3 | 2 | 0 – 0 – 5 - 6 |
| 4 – 3 – 3 – 0 | 3 | 1 – 1 – 2 |
| | 4 | |

fin des nombres de la 1ère distribution

début des nombres

fin des nombres de la 2ème distribution

Les résultats de la 1ère distribution sont : 2, 7, 8, 8, 9, 11, 14, 15, 17, 18, 23, 23, 26, 30, 33, 33, 34

Les résultats de la 2ème distribution sont : 1, 2, 6, 7, 9, 10, 13, 14, 18, 19 20, 20, 25, 26, 31, 31, 32

**Écart moyen**

L'écart moyen est une mesure de dispersion qui indique la moyenne des écarts de chacune des données à la moyenne d'une distribution.

$$\textbf{Écart moyen} = \frac{\textbf{Somme des écarts à la moyenne}}{\textbf{Nombre total de données}}$$

Voici une distribution comportant 8 données : 1, 4, 5, 6, 8, 8, 9, 11.

La moyenne de cette distribution est 6,55

Un écart correspond à la valeur absolue de la différence entre deux valeurs.

Calcul de l'écart moyen :

| Donnée | Écart à la moyenne |
|---|---|
| 1 | \| 1 – 6,5 \| = 5,5 |
| 4 | \| 4 – 6,5 \| = 2,5 |
| 5 | \| 5 – 6,5 \| = 1,5 |
| 6 | \| 6 – 6,5 \| = 0,5 |
| 8 | \| 8 – 6,5 \| = 1,5 |
| 8 | \| 8 – 6,5 \| = 1,5 |
| 9 | \| 9 – 6,5 \| = 2,5 |
| 11 | \| 11 – 6,5 \| = 4,5 |
| $\bar{x} = \mathbf{6,5}$ | *Écart moyen* = ***2,5*** |

**Écart moyen :** $\frac{(5,5 + 2,5 + 1,5 + 0,5 + 1,5 + 1,5 + 2,5 + 4,5)}{8} = 2,5$

Exercices :

## 1. Deux données manquantes

La moyenne d'une distribution statistique est 27.

L'écart moyen de cette distribution est 5,5.

Cette distribution statistique est composée de 8 données. Voici 6 de ces 8 données.

**20 21 23 32 33 34**

Quelles sont les deux données manquantes?

Démarche :

| donnée | Écart à la moyenne |
|---|---|
| 20 | \| 20 – 27 \| = 7 |
| 21 | \| 21 – 27 \| = 6 |
| 23 | \| 23 – 27 \| = 4 |
| 32 | \| 32 – 27 \| = 5 |
| 33 | \| 33 – 27 \| = 6 |
| 34 | \| 34 – 27 \| = 7 |
| x | \| x – 27 \| |
| y | \| y – 27 \| |
| $\bar{x} = 27$ | |

1) La moyenne est 27 :

$$\frac{(20 + 21 + 23 + 32 + 33 + 34 + x + y)}{8} = 27$$

$$\frac{(163+x+y)}{8} = 27$$

$$(163 + x + y) = 216$$

$$x + y = 53$$

2) L'écart moyen est 5,5 :

$$\frac{(7+6+4+5+6+7)+ |\,x - 27\,| + |\,y - 27\,|}{8} = 5{,}5$$

$$\frac{(35) + |\,x - 27\,| + |\,y - 27\,|}{8} = 5{,}5$$

$$35 + |\,x - 27\,| + |\,y - 27\,| = 44$$

$$|\,x - 27\,| + |\,y - 27\,| = 9$$

3) Cherchons les nombres x et y par essais et erreurs :

$9 \div 2 = 4{,}5$

$|\,x - 27\,| = 4$ $\qquad$ $|\,y - 27\,| = 5$

$x - 27 = 4 \;\; ou \;\; x - 27 = -4$ $\qquad$ $y - 27 = 5 \;\; ou \;\; y - 27 = -5$

$x = 31$ ou $x = 23$ $\qquad$ $y = 32$ $ou$ $y = 22$

Comme : $31 + 22 = 53$

Réponse : Les deux données manquantes sont 31 et 22.

## 2. Deux données manquantes

La moyenne d'une distribution statistique est 16.

L'écart moyen de cette distribution est 3,4.

Cette distribution statistique est composée de 10 données. Voici 8 de ces 10 données.

**11 12 13 13 16 17 22 24**

Quelles sont les deux données manquantes?

Démarche :

| donnée | Écart à la moyenne |
|---|---|
| 11 | \| 11 – 16 \| = 5 |
| 12 | \| 12 – 16 \| = 4 |
| 13 | \| 13 – 16 \| = 3 |
| 13 | \| 13 – 16 \| = 3 |
| 16 | \| 16 – 16 \| = 0 |
| 17 | \| 17 – 16 \| = 1 |
| 22 | \| 22 – 16 \| = 6 |
| 24 | \| 24 – 16 \| = 8 |
| x | \| x – 16 \| |
| y | \| y – 16 \| |
| $\bar{x} = 16$ | |

1) La moyenne est 16 :

$$\frac{(11 + 12 + 13 + 13 + 16 + 17 + 22 + 24 + x + y)}{10} = 16$$

$$\frac{(128+x+y)}{8} = 16$$

$$(128 + x + y) = 160$$

$$x + y = 32$$

2) L'écart moyen est 3,4 :

$$\frac{(5+4+3+3+0+1+6+8)+ |x-16| + |y-16|}{10} = 3,4$$

$$\frac{(30) + |x-16| + |y-16|}{10} = 3,4$$

$$30 + |x-16| + |y-16| = 34$$

$$|x-16| + |y-16| = 4$$

3) Cherchons les nombres x et y par essais et erreurs :

$4 \div 2 = 2$

$|x-16| = 2$ $\qquad$ $|y-16| = 2$

$x - 16 = 2 \ ou \ x - 16 = -2$ $\qquad$ $y - 16 = 2 \ ou \ y - 16 = -2$

$x = 18$ ou $x = 14$ $\qquad$ $y = 18$ $ou$ $y = 14$

Comme : $18 + 14 = 32$

Réponse : Les deux données manquantes sont 18 et 14 .

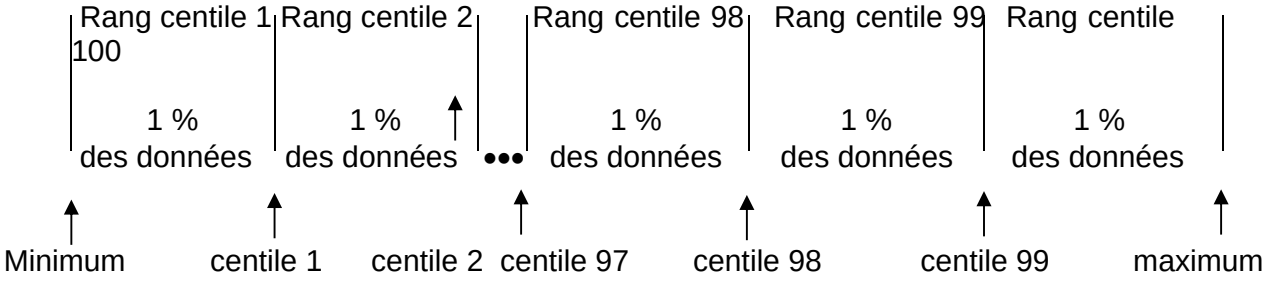
## Mesure de position

Une mesure de position sert à situer une donnée parmi les autres données d'une distribution.

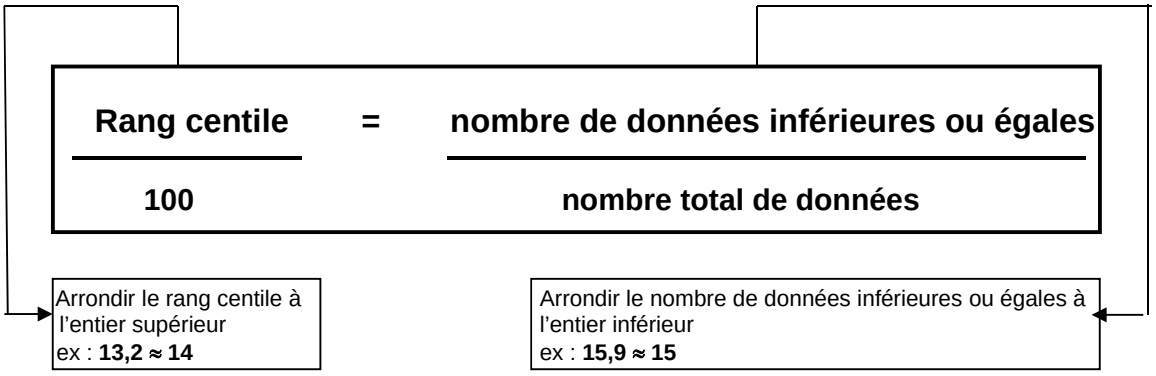
**Rang centile**

Le rang centile d'une donnée est une mesure de position qui indique le pourcentage de données inférieures ou égales à cette donnée dans la distribution.

À l'aide de 99 valeurs appelées centiles, il est possible de partager une distribution ordonnée en 100 sous-ensembles contenant chacun 1% des données. Le rang de chaque sous-ensemble constitue le rang centile de chacune des données qu'il contient.




Calcul du **rang centile**




Le rang centile permet de comparer la position de deux données provenant de distributions différentes.

Notation : $R_c$ ou $R_{(78)}$ : le rang centile de la donnée 78.

Exemple : Soit la série : 30, 34, 35, 35, 38, 38, 40, 40, 44, 44,
45, 45, 47, 48, 49, 50

Quel est le rang centile de la donnée 44 ?

$$\frac{R_{(44)}}{100} = \frac{10}{16}$$

← Il y a 10 données ≤ à la donnée 44

$R_{(44)} = 100 \bullet 10 \div 16 = 62{,}5 \approx 63$

réponse : $R_{(44)} = 63$

Quel est le rang centile de la donnée 48 ?

$$\frac{R_{(48)}}{100} = \frac{14}{16}$$

← Il y a 14 données ≤ à la donnée 48

$R_{(48)} = 100 \bullet 14 \div 16 = 87{,}5 \approx 88$

réponse : $R_{(48)} = 88$

Exercices :

1. À partir de cette distribution : 30, 34, 35, 35, 38, 38, 40, 40, 44, 44,
45, 45, 47, 48, 49, 50.

Quelle donnée a le rang centile 30 ($R_c$= 30) ?

$$\frac{30}{100} = \frac{d}{16}$$

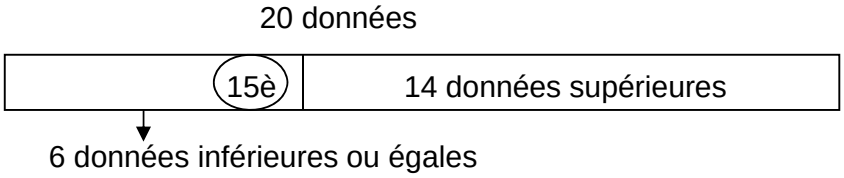
→ nbr de données ≤

$d = 30 \cdot 16 \div 100 = 4{,}8$ → Il y a 4 données ≤ à la donnée que je cherche dans la série.

réponse : 35

$R_{(35)} = 30$

2. Quel est le rang centile, de la 15ème donnée parmi 20 ?

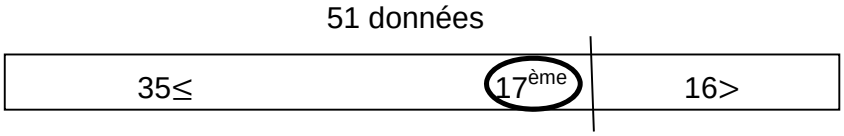



$$\frac{R_{(15è)}}{100} = \frac{6}{20}$$

$R_{(15è)} = 100 \bullet 6 \div 20 = 30$

réponse : 30

3. Quel est le rang centile, de la 17ème donnée parmi 51 ?




$$\frac{R_{(17è)}}{100} = \frac{35}{51}$$

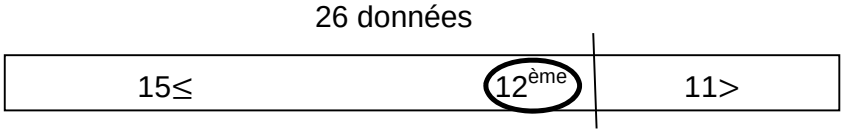
$R_{(17è)} = 100 \bullet 35 \div 51 \approx 68{,}7$

réponse : 69

4. À quel rang (1er , 2ème, 3ème …) a terminé quelqu'un qui a 61 comme rang centile si au total il y a 26 données.

$$\frac{61}{100} = \frac{d}{26} \rightarrow \text{nbr de données} \leq$$

$d = 61 \cdot 26 \div 100 = 15{,}86 \rightarrow$ Il y a 15 données ≤ à la donnée que je cherche dans la série.




réponse : 12ème

5. À quel rang (1er , 2ème, 3ème …) a terminé quelqu'un qui a 92 comme rang centile si au total il y a 5200 données.

$$\frac{92}{100} = \frac{d}{5\,200} \longrightarrow \text{nbr de données} \leq$$

$d = 92 \cdot 5\,200 \div 100 = 4\,784$ ⟶ Il y a 4 784 données ≤ à la donnée que je cherche dans la série.

5 200 données

| 4 784 ≤ | 417ème | 416 > |
|---|---|---|

réponse : 417ème

6. Voici les notes des élèves de M. Tremblay lors d'un examen de statistiques :

75, 76, 76, 77, 77, 79, 80, 81, 82, 83, 84, 84, 84, 84, 84, 85, 86, 87, 88, 88, 88, 88, 89, 89, 89, 89, 91, 91, 91

Quel est le rang centile de la donnée 89 ?

$$\frac{R_{(89)}}{100} = \frac{26}{29}$$ ← Dans la série. il y a 26 données ≤ à la donnée 89

$R_{(89)} = 100 \bullet 26 \div 29 \approx 89{,}66$ réponse : 90

7. Toujours à partir de la distribution des notes d'examen du groupe de M. Tremblay ,

Quelle donnée a le rang centile 75 ?

$$\frac{75}{100} = \frac{d}{29} \longrightarrow \text{nbr de données} \leq$$

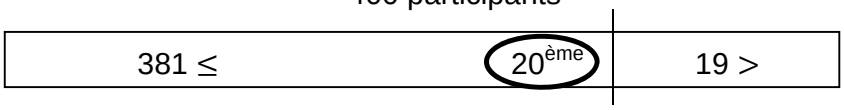
$d = 75 \cdot 29 \div 100 = 21{,}75$ ⟶ Il y a 21 données ≤ à la donnée que je cherche dans la série.

réponse : 88

$R_{(88)} = 75$

8. Lors d'un concours, Jacques s'est classé 20ème sur 400 participants. Quel est son rang centile ?

400 participants




$$\frac{R_{(20è)}}{100} = \frac{381}{400}$$

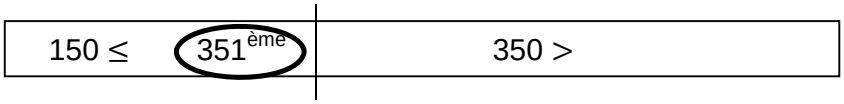
$R_{(20è)} = 100 \bullet 381 \div 400 \approx 95{,}25$ réponse : 96

9. Lors d'un concours opposant 500 concurrents, Sylvie s'est classée avec un rang centile de 30. Quelle était sa position (1ère , 2ème, 3ème …) ?

$$\frac{30}{100} = \frac{d}{500}$$

$d = 30 \cdot 500 \div 100 = 150$ ⟶ Il y a 150 concurrents ≤ à Sylvie.

500 concurrents



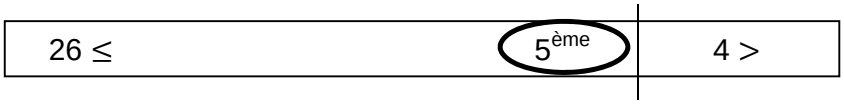

réponse : 351ème

10. Thomas a obtenu un rang centile de 88 lors du dernier examen d'histoire. Quelle était sa position, s'il y a 30 élèves dans la classe ?

$$\frac{88}{100} = \frac{d}{30}$$

$d = 88 \cdot 30 \div 100 = 26{,}4$ ⟶ Il y a 26 élèves ≤ à Thomas.

30 élèves

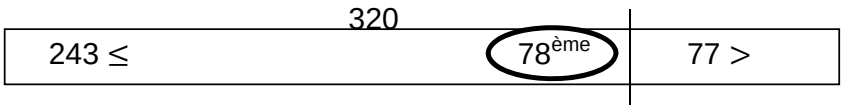



réponse : 5ème

11. Brutus a obtenu un rang centile de 76 lors de la dernière compétition de Sumo. Quel était son rang dans la série (position) s'il y avait 320 participants ?

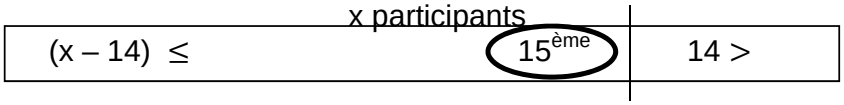
$$\frac{76}{100} = \frac{d}{320}$$

$d = 76 \cdot 320 \div 100 = 243{,}2$ ⟶ Il y a 243 données ≤.




réponse : 78ème

12. Jean s'est classé 15ème à une compétition et son rang centile était de 85. Combien y avait-il de participants à cette compétition ?




$$\frac{85}{100} = \frac{(x - 14)}{x} \quad \text{ou} \quad \frac{15}{100} = \frac{14}{x}$$

$85x = 100\,(x - 14)$ $\qquad$ $15x = 1\,400$

$85x = 100x - 1\,400$ $\qquad$ $x = 93{,}333$

$-15x = -1\,400$

$x = 93{,}333$

réponse : 93 participants

# Distribution à deux caractères

Une **distribution à deux caractères** correspond à l'ensemble des couples de données recueillies au cours d'une étude statistique portant sur deux caractères issus d'une même situation.

Dans une étude statistique, on donne le nom de **variable statistique** à tout caractère dont les données peuvent être différentes.

Dans l'étude de deux variables, une première façon de représenter la situation est de :

## 1) Construire un tableau à double entrée

Ce tableau permet de mettre en évidence des informations relatives aux variables étudiées.

A) À partir de variables qualitatives

*Exemple :* Voici les résultats d'une étude portant sur la répartition des décès des 18-50 ans selon le sexe et la cause.

Les variables à l'étude : - la cause du décès (x)
- le sexe (y)

**Répartition des décès selon le sexe et la cause**

| Cause | Hommes | Femmes | Total |
|---|---|---|---|
| Accidents | **254** | **77** | **331** |
| Suicides | **189** | **26** | **215** |
| Autres maladies que les cancers | **59** | **27** | **86** |
| Cancers | **41** | **37** | **78** |
| **Total** | **543** | **167** | **710** |

→ y

↓ x

Remarques :

1) Donner un **titre** à votre tableau.

2) Afficher le total pour chacune des lignes et colonnes.

3) Inscrire la taille de l'échantillon dans le coin inférieur droit.

4) Les x sont dans la première colonne et les y sont dans la première ligne.

5) Regarder le total de certaines lignes ou colonnes ou certaines cases en particulier pour mettre en évidence des informations.

B) À partir de variables quantitatives

C'est à partir d'ici qu'on peut introduire le concept de **Corrélation**. La corrélation caractérise le lien qui peut exister entre différents caractères d'une population.

*Exemple :* Voici les couples d'une étude statistique où **x** représente **l'âge des moutons** et **y** représente **la production de laine du mouton (en kg) pour une année**.

(3, 5,0), (3, 5,1), (3, 5,3), (3, 5,3), (3, 5,7), (3, 5,8),
(4, 5,2), (4, 5,5), (4, 5,8), (4, 5,8), (4, 5,9), (4, 5,9),
(5, 5,4), (5, 6,1), (5, 6,2), (5, 6,3), (5, 6,4), (5, 6,5),
(6, 5,5), (6, 6,3), (6, 6,4), (6, 6,9), (6, 7,1), (6, 7,4),
(7, 6,8), (7, 7,1), (7, 7,2), (7, 7,4), (7, 7,5), (7, 8,2),
(8, 6,5), (8, 7,4), (8, 7,6), (8, 8,1), (8, 8,1), (8, 8,2),
(9, 8,2), (9, 8,3), (9, 8,3), (9, 8,5),
(10, 7,5), (10, 8), (10, 8,1), (10, 8,1),
(11, 6,8), (11, 7,0), (11, 7,2),
(12, 7,3).

Comme il y a un grand nombre de données en y, nous formerons des classes lors de la représentation de cette situation à l'aide d'un tableau à double entrée.


y

**Production de laine selon l'âge des moutons**

x

| Production (en kg) / Âge (en année) | [4, 5[ | [5, 6[ | [6, 7[ | [7, 8[ | [8, 9[ | Total |
|---|---|---|---|---|---|---|
| 3 | | 6 | | | | 6 |
| 4 | | 6 | | | | 6 |
| 5 | | 1 | 5 | | | 6 |
| 6 | | 1 | 3 | 2 | | 6 |
| 7 | | | 1 | 4 | 1 | 6 |
| 8 | | | 1 | 2 | 3 | 6 |
| 9 | | | | | 4 | 6 |
| 10 | | | | 1 | 3 | 4 |
| 11 | | | 1 | 2 | | 3 |
| 12 | | | | 1 | | 1 |
| Total | | 14 | 11 | 12 | 11 | 48 |

+

Pour déterminer s'il y a de la corrélation entre les 2 variables à l'étude, nous regardons si la **majorité des données suivent l'une des deux diagonales du tableau**. On dit alors que la corrélation est linéaire positive ou négative.

La conclusion de la situation de la page précédente :

Il y a une corrélation linéaire positive entre les 2 variables.

La production de laine augmente avec l'âge des moutons.

**Problème :** Maxime et Brigitte participent à des compétitions de BMX. Après plusieurs compétitions, ils ont enregistré les résultats suivants :

| Masse du vélo ( en kg) | Hauteur du saut ( en cm) |
|---|---|
| 8,6 | 25,9 |
| 8,7 | 25,75 |
| 9,1 | 25,625 |
| 9,3 | 25,5 |
| 9,5 | 25,25 |
| 10 | 24,625 |
| 10,2 | 24,5 |
| 10,5 | 24,4 |
| 10,7 | 24,25 |
| 10,9 | 24 |
| 11,3 | 23,75 |

Construis le tableau à double entrée et dis-moi s'il y a une relation entre les deux variables

y (-)

x (+)

| Hauteur du saut (cm) / Masse du vélo (en kg) | [23,5 ; 24[ | [24 ; 24,5[ | [24,5 ; 25[ | [25 ; 25,5[ | [25,5 ; 26[ | total |
|---|---|---|---|---|---|---|
| [8,5 ; 9[ | | | | | 2 | 2 |
| [9 ; 9,5[ | | | | | 2 | 2 |
| [9,5 ; 10[ | | | | 1 | | 1 |
| [10 ; 10,5[ | | | 2 | | | 2 |
| [10,5 ; 11[ | | 3 | | | | 3 |
| [11 ; 11,5[ | 1 | | | | | 1 |
| total | 1 | 3 | 2 | 1 | 4 | 11 |

Conclusion :

Il y a une corrélation linéaire négative entre les 2 variables.

Plus la masse du vélo augmente moins la saut est élevé.

## 2- Construire un nuage de points

C'est un graphique déterminant l'**intensité** de la corrélation. (points non reliés correspondant à chacun des couples de la distribution)

Ainsi, plus les points tendent à former une droite, plus la corrélation est forte (plus la corrélation est dite *linéaire*).

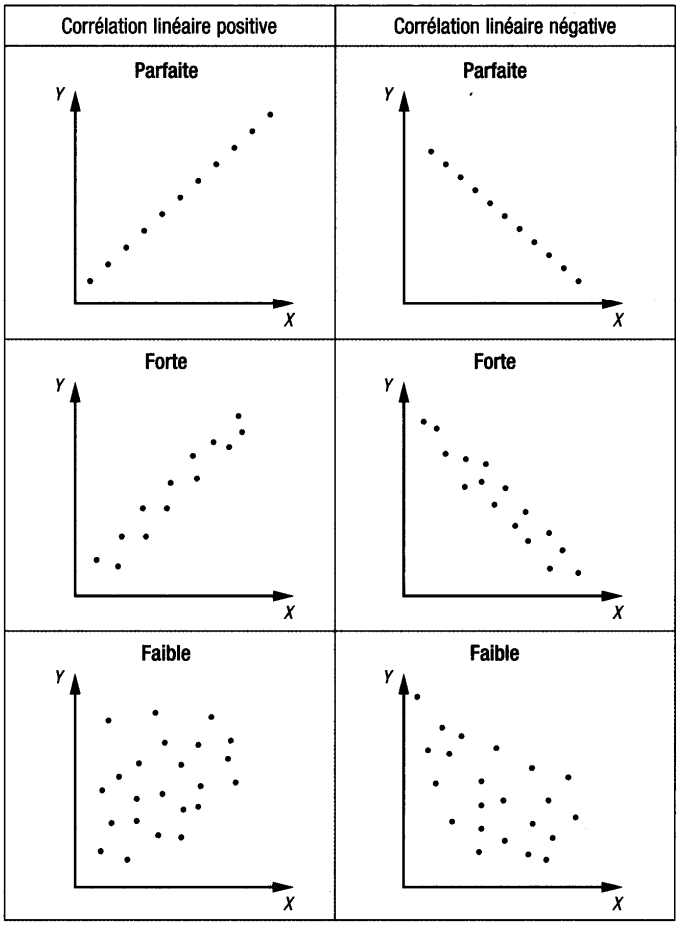
On peut ainsi ***qualifier*** la corrélation, voici ses pricipales caractéristiques :

La corrélation est **positive** → Si les variables varient dans le **même sens.**
( si x augmente alors y augmente )

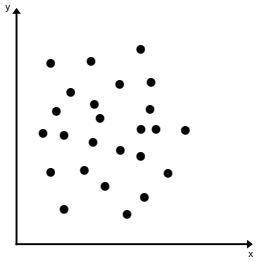
**négative** → Si les variables varient dans le **sens contraire**.
( si x augmente alors y diminue )

**nulle** → Si les points sont distribués au hasard.

De façon générale, on se réfère à ces nuages :

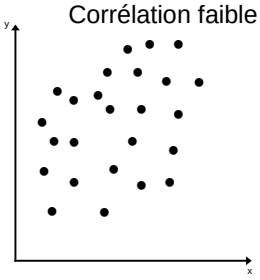


**Corrélation nulle**



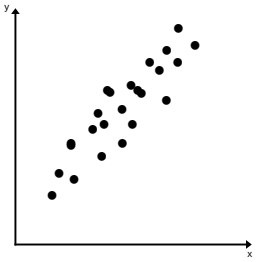
**Corrélation positive**

Corrélation faible

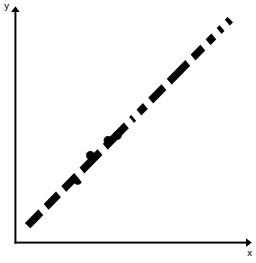


Corrélation moyenne

Corrélation forte

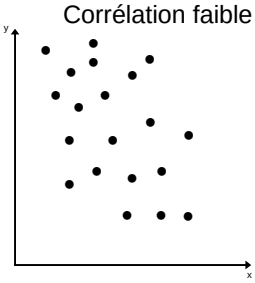


Corrélation parfaite

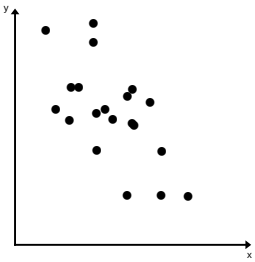


**Corrélation négative**

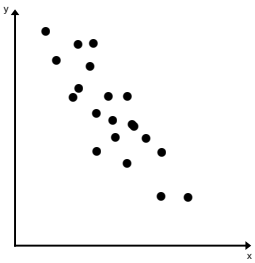
Corrélation faible



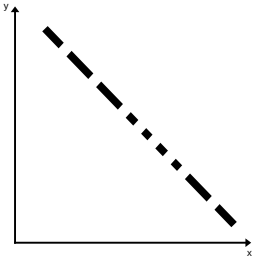
Corrélation moyenne



Corrélation forte



Corrélation parfaite



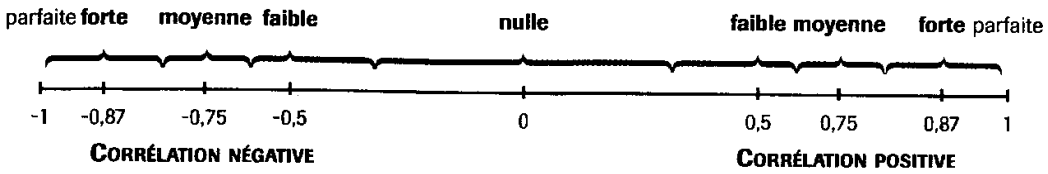
**Remarques :**Les nuages ci-dessus ne sont que des croquis car un vrai nuage de points doit avoir un **titre**. De plus, les **axes** doivent être **identifiées** et **graduées.**

## Interprétation quantitative de la corrélation

### Coefficient de corrélation

On cherche maintenant à ***quantifier*** la corrélation. Le coefficient de corrélation ( noté **r** ) est un nombre utilisé pour quantifier le degré de corrélation entre 2 variables quantitatives. Ce nombre sera toujours compris dans l'intervalle [-1, 1].

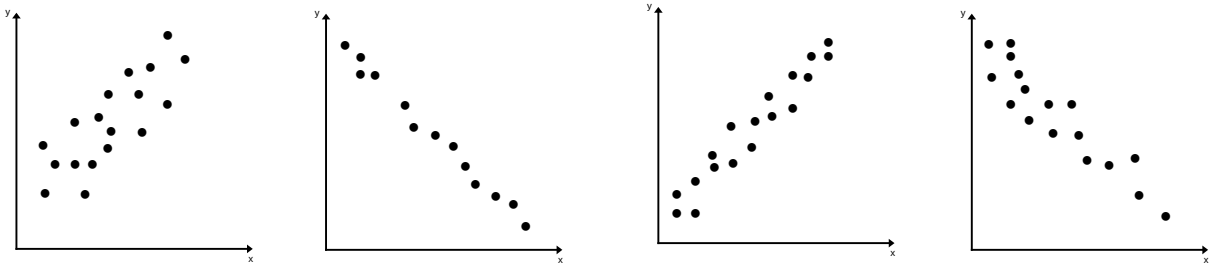
Donc **r** prendra sa valeur en fonction de l'échelle de la corrélation suivante:




Ainsi , on pourra décrire en mots la corrélation entre les 2 variables :

Exemple :

À quel nuage de points est associé chacun des coefficients de corrélation linéaire suivants? -0,98 -0,86 0,61 0,94



0,61 -0,98 0,94 -0,86

Décrire en mots la corrélation :

| | | | |
|---|---|---|---|
| positive | négative | positive | négative |
| faible-moyenne | très-forte | très-forte | forte |

## Estimation de la corrélation :

### A) Méthode d'estimation graphique (méthode du rectangle)

À partir du nuage de points,

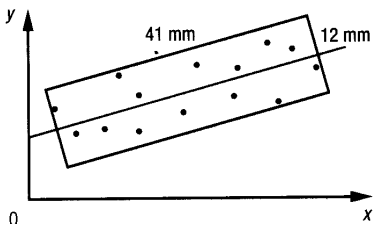
1) On trace une droite qui passe au centre de l'ensemble de points. ( celle-ci doit suivre la tendance générale des points)

2) On trace ensuite un rectangle qui contiendra tous les points (sauf les données aberrantes)

3) Par la suite, on utilise alors la formule suivante pour estimer la corrélation :

$$r \approx \pm \left(1 - \frac{\text{mesure du petit côté}}{\text{mesure du grand côté}}\right)$$

Exemple :




$r \approx 1 - \frac{12}{41} \approx 0{,}71$

La corrélation entre les deux variables est donc positive et moyenne.

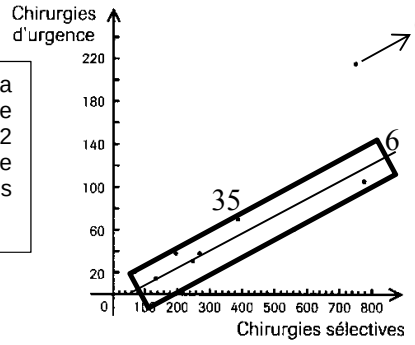
Remarques :

1) La précision de cette méthode est de plus ou moins 0,15

2) Lorsque la corrélation est négative ( rectangle vers le bas )
Nous devons nous-mêmes ajouter le – devant la valeur obtenue

Exemple : Calculer le coefficient de corrélation.

Si le graphique n'a pas la même graduation pour les 2 axes, utiliser le même nombre d'intervalles en x et en y.




$r = +(1 - \frac{6}{35}) \approx 0{,}83$

La corrélation entre les 2 variables est positive moyenne-forte.

**B)** Procédure pour **trouver le coefficient de corrélation** avec la **calculatrice à affichage graphique**

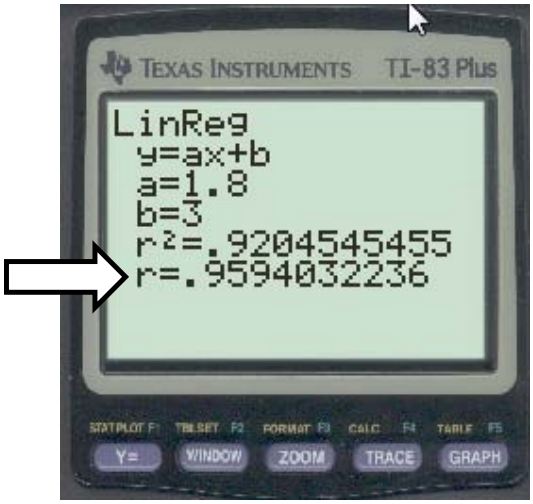
Enlever les données aberrantes de la série de données, s'il y en a.
Écrire la série de données sous $L_1$ *et* $L_2$, en suivant les étapes 1 à 5 de la page 10.

### ➢ Étape 1

Pour obtenir le coefficient de corrélation **r** de façon définitive, aller dans **Catalog (** $\boxed{2^{nd}}$**,** $\boxed{\mathbf{0}}$ , sélectionner **DiagnosticOn** en descendant avec les flèches. Appuyer 2 fois sur la touche $\boxed{\mathbf{ENTER}}$. Done sera inscrit à l'écran.

### ➢ Étape 2

Appuyer sur la touche $\boxed{\mathbf{Stat}}$. Sélectionner CALC avec les flèches. Descendre à la ligne **LinReg (ax + b)** . Appuyer 2 fois sur la touche $\boxed{\mathbf{ENTER}}$. Le coefficient de corrélation « **r »** apparaîtra.




## La droite de régression

C'est la droite qui représente le mieux l'ensemble des points du nuage de points. Elle est de la forme y = ax + b

C'est aussi la droite qui permet de **prédire la ou les valeurs de l'une des variables à partir des valeurs de l'autre.**

Ainsi le coefficient de corrélation permet de savoir jusqu'à quel point cette prédiction est fiable. Plus le coefficient est élevé, plus la prédiction sera valable.

Par exemple, si le r = 0,95 alors la droite de régression sera un bon outil de prédiction puisque le nuage de points a la forme d'une droite.

On peut obtenir cette droite **par estimation** ou **avec la calculatrice**

a) **Par estimation** :

Il s'agit d'une droite qui passe par deux points représentatifs de la distribution. Voici une méthode pour définir cette droite :

• **Méthode de la droite de Mayer**:

1. Ordonner les couples de la distribution d'après leurs abscisses.
2. Diviser l'ensemble des couples en deux groupes, égaux si possible.
3. Déterminer la moyenne des abscisses et la moyenne des ordonnées dans chacun des deux groupes afin de former les couples moyens $P_1$ ($x_1$, $y_1$) et $P_2$ ($x_2$, $y_2$).
4. La droite de régression est celle qui passe par les points $P_1$ et $P_2$.

$Ex_1$ :

| x | y | |
|---|---|---|
| 8 | 25 | |
| 9 | 28 | |
| 12 | 41 | $P_1$ ( 12 , 38 ) |
| 15 | 46 | |
| 16 | 50 | |
| 17 | 57 | |
| 20 | 52 | |
| 21 | 67 | $P_2$ ( 22 , 64 ) |
| 25 | 70 | |
| 27 | 74 | |

r = 0,97

↓

La corrélation est très forte, la droite de régression nous permettra de prédire la valeur d'une variable à partir de l'autre.

Trouve l'équation de la droite qui passe par $P_1$ et $P_2$ :

$$a = \frac{64 - 38}{22 - 12} = \frac{26}{10}$$

$$y = 2{,}6x + b$$

$$38 = 2{,}6 \cdot 12 + b$$

$$38 = 31{,}2 + b$$

$$6{,}8 = b$$

$$y = 2{,}6x + 6{,}8$$

Réponse : $y = 2{,}6x + 6{,}8$

Quelle sera la valeur de y si x=10 ?

$$y = 2{,}6x + 6{,}8$$

$$y = 2{,}6 \cdot 10 + 6{,}8$$

$$y = 26 + 6{,}8 = 32{,}8$$

Réponse : 32,8

$Ex_2$ : Quelle est l'équation de la droite de régression de cette distribution ?

| x | y | |
|---|---|---|
| 2 | 16 | |
| 5 | 14 | |
| 7 | 12 | $P_1$ ( 6,33 , 12,33) |
| 7 | 12 | |
| 8 | 11 | |
| 9 | 9 | |
| 11 | 7 | |
| 13 | 6 | |
| 15 | 7 | $P_2$ ( 17 , 5,67 ) |
| 18 | 7 | |
| 22 | 5 | |
| 23 | 2 | |

$$a = \frac{5{,}67 - 12{,}33}{17 - 6{,}33} = \frac{-6{,}66}{10{,}67} = -0{,}62$$

$$y = -0{,}62x + b$$

$$12{,}33 = -0{,}62 \cdot 6{,}33 + b$$

$$12{,}33 = -3{,}92 + b$$

$$16{,}25 = b$$

$$y = -0{,}62x + 16{,}25$$

r = -0,62

↓

La corrélation est moyenne, nous pouvons prédire la valeur d'une variable à l'aide de la droite de régression.

Réponse : $y = -0{,}62x + 16{,}25$

Quelle sera la valeur de x si y=10 ?

$$y = -0{,}62x + 16{,}25$$

$$10 = -0{,}62 \cdot x + 16{,}25$$

$$-6{,}25 = -0{,}62 \cdot x$$

$$10{,}08 = x$$

Réponse : 10,08

b) **Avec la calculatrice à affichage graphique**

La procédure pour **trouver l'équation de la droite de régression** est la même que celle utilisée pour **trouver le coefficient de corrélation** (voir page 34).

Vous devez substituer les paramètres $\boldsymbol{a}$ et $\boldsymbol{b}$ dans l'équation $\boldsymbol{y = ax + b}$.

La droite de régression permet de prédire la ou les valeurs de l'une des variables à partir des valeurs de l'autre, et le coefficient de corrélation permet de savoir jusqu'à quel point cette prédiction est fiable.

Exemple :

On s'intéresse à la relation, des 10 derniers mois, entre les montants alloués à la publicité ( X ) en milliers de $ et les ventes en milliers de $ d'une entreprise de cellulaire.

| x (en milliers de $) | y (en milliers de $) |
|---|---|
| 6 | 23 |
| 7 | 26 |
| 10 | 39 |
| 13 | 44 |
| 14 | 48 |
| 15 | 55 |
| 18 | 50 |
| 19 | 65 |
| 23 | 68 |
| 25 | 72 |

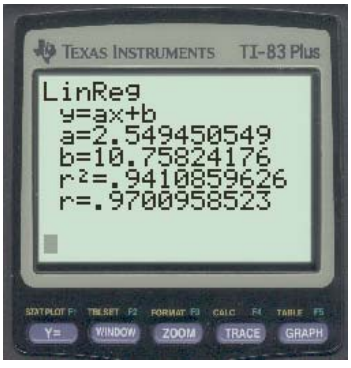
a) Calcule le coefficient de corrélation à l'aide de la calculatrice.

Réponse : r = 0,97

b) La corrélation est-elle significative ?

Corrélation positive très forte, nous pouvons utiliser la droite de régression pour prédire la valeur d'une variable à partir de la valeur de l'autre.

c) Calcule l'équation de la droite de régression avec la calculatrice :




Réponse : $y = 2{,}55x + 10{,}76$

d) Question d'estimation :

Si l'entreprise investit 16 000 $ en publicité, quel volume de vente peut-elle espérer obtenir ?

x=16 ↓

$y = 2{,}55x + 10{,}76$

$y = 2{,}55 \cdot 16 + 10{,}76$

$y = 40{,}8 + 10{,}76$

$y = 51{,}56$

Réponse : Le volume des ventes sera estimé à 51 560$

**Note:** Toujours appuyer notre prédiction avec la droite de régression.

## Interprétation de la corrélation

La corrélation prouve l'existence d'un lien mais pas nécessairement un lien de cause à effet, elle n'explique **ni le pourquoi ni le comment** des choses.

L'interprétation dépend des valeurs trouvées, des circonstances et des contextes. Il faut donc

**éviter les conclusions hâtives** et **considérer toutes les hypothèses.**

Certaines corrélations peuvent s'expliquer par le simple fait du hasard ou par l'influence d'une 3ième variable.