

LES STATISTIQUES

CHAPITRE 7

m

~NOTES DE COURS~



**MATHÉMATIQUE CST - 4^E SECONDAIRE
COLLÈGE REGINA ASSUMPTA
2018 – 2019
MADAME BLANCHETTE**

**INSPIRÉ DU DOCUMENT DE NOTES DE COURS
DE AUDREY-ANN BOSSÉ (CDSL)**

NOM : _____

GROUPE : _____

1. Rappel – Mesures de tendance centrale

La **moyenne** est la mesure indiquant le centre d'équilibre d'une distribution. C'est la valeur qu'aurait chacune des données si elles étaient toutes identiques (si elles avaient toutes la même valeur).

Notation : \bar{x}

Formule :

Il est possible de généraliser la formule de la moyenne de la façon suivante :

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\text{somme des données}}{\text{nombre de données}}$$

Ce nouveau symbole Σ signifie
« faire la somme de... »

Note : À moins d'indication contraire, on doit **toujours** arrondir la moyenne à une décimale de plus que le nombre maximal de décimales de la liste des données.

Le **mode** d'une distribution est la donnée qui se répète le plus souvent. C'est le centre de concentration d'une distribution.

Le mode aide à déterminer, dans une enquête ou un sondage, la personne ou la marque la plus populaire. C'est la valeur dont l'effectif est le plus grand.

Notation : Mo

La **médiane** est une mesure de position et elle permet de localiser le centre de la distribution lorsque les données sont placées en ordre croissant.

Pour calculer la médiane, il faut :

Pour un nombre **impair** de données :

- 1) Placer les données en ordre croissant.
- 2) Trouver la donnée centrale dans la liste.

Pour un nombre **pair** de données :

- 1) Placer les données en ordre croissant.
- 2) Effectuer la moyenne entre les deux données du centre.

Notation : Md

La médiane est souvent plus significative et représentative que la moyenne lorsqu'il y a une ou des données éloignées (données extrêmes) par rapport à l'ensemble des données de la distribution.

Exemple :

Voici une distribution ordonnée comportant 16 données :

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

- a) Quel est le mode de cette distribution?

- b) Quelle la médiane de cette distribution?

- c) Quelle est la moyenne de cette distribution?

2. 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 sujet.

Voici quelques outils nous permettant de représenter et d'analyser des distributions à un caractère.

A) Diagramme à tige et à feuilles

Le **diagramme à tige et à feuilles** est utilisé pour représenter les données d'une ou de deux distributions qui sont disposées, un ou des 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;
- ✓ on écrit une légende pour faciliter la lecture du diagramme.

Exemples :

1. Les données de la distribution ci-dessous correspondent aux pulsations cardiaques de 24 personnes à la suite d'un effort physique.

70, 73, 73, 76, 78, 82, 85, 85, 87, 88, 89, 90, 92, 92,
96, 97, 99, 101, 101, 106, 106, 107, 112, 114.

Représente cette distribution à l'aide d'un diagramme à tige et à feuille.

2. Les données des deux distributions ci-dessous correspondent à la taille (en mm) de 28 saumons.

Mâles : 154, 180, 190, 225, 232, 232, 273, 306, 311, 315, 342, 343, 444, 550, 566.

Femelles : 136, 152, 180, 200, 235, 271, 271, 312, 341, 362, 399, 407, 414.

Représente cette distribution à l'aide d'un diagramme à tige et à feuilles.

B) L'écart moyen – Dispersion des données

Une mesure de dispersion sert à décrire l'étalement ou la concentration des données d'une distribution. L'écart moyen est une mesure de dispersion qui indique la moyenne des écarts de chacune des données à la moyenne d'une distribution.

$$\text{Écart moyen} = \frac{\text{somme des écarts à la moyenne (en valeur absolue)}}{\text{nombre total de données}}$$

Exemples :

1. Voici une distribution comportant 6 données : 1, 4, 5, 6, 8, 12. Détermine l'écart moyen de cette distribution.

Étape 1 : Calculer la moyenne des données

$$\bar{x} = \frac{1+4+5+6+8+12}{6} = \frac{36}{6} = 6$$

Étape 2 : Calculer l'écart à la moyenne, en valeur absolue, pour chaque donnée. Organiser les données dans un tableau peut faciliter le calcul.

Donnée	Valeur absolue de l'écart à la moyenne
1	$ 1-6 = 5$
4	$ 4-6 = 2$
5	$ 5-6 = 1$
6	$ 6-6 = 0$
8	$ 8-6 = 2$
12	$ 12-6 = 6$
Total	16

Étape 3 : Calculer la moyenne des écarts à la moyenne. (Écart moyen)

$$\text{É.M.} = \frac{16}{6} \approx 2,67$$

Étape 4 : Interpréter les résultats.

Voir encadré p.8

2. Lors d'un cours d'été en mathématiques, un enseignant a comptabilisé les résultats finaux de ces élèves : 80 %, 68 %, 77 %, 81 %, 56 % et 70 %.

- a) Combien d'élèves ont obtenu un résultat supérieur à la moyenne?

i) Moyenne

$$\bar{x} = \frac{80 + 68 + 77 + 81 + 56 + 70}{6} = \frac{432}{6} = 72\%$$

Rép: 3 élèves

- b) Détermine l'écart moyen des résultats de ce groupe en cours d'été.

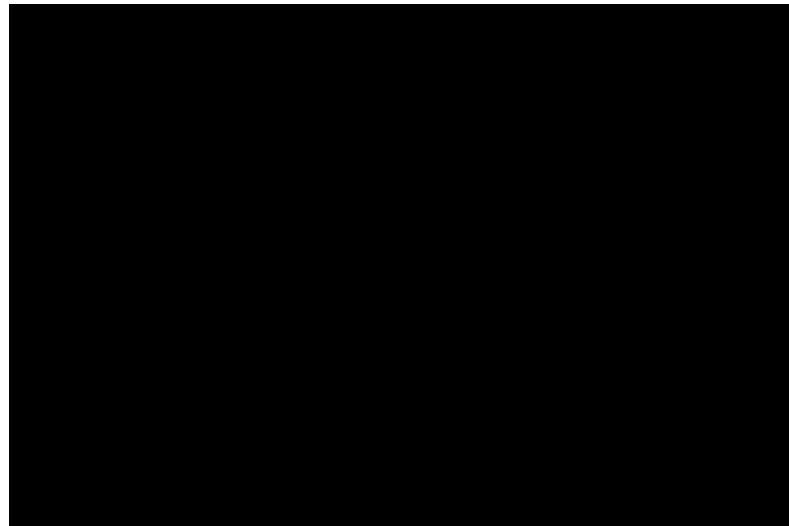
Données	Écart à la \bar{x}
80	8
68	4
77	5
81	9
56	16
70	2
TOTAL	44

$$\text{E.M.} = \frac{44}{6} \approx 7,33\%$$

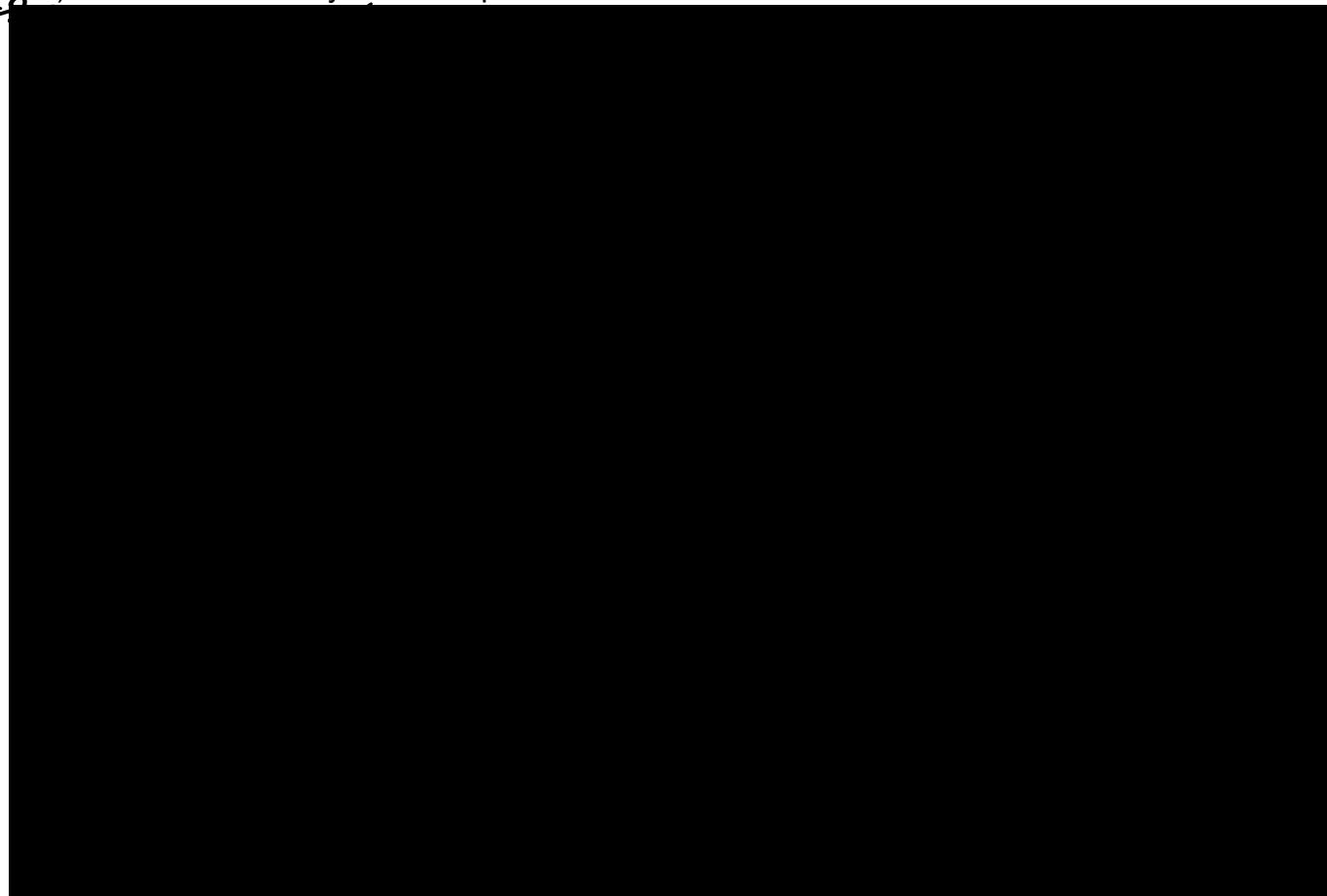
3. Voici les scores obtenus par 12 joueurs de golf :

87, 92, 98, 101, 103, 103, 107, 110, 116, 124, 128 et 139.

a) Représente cette distribution à l'aide d'un diagramme à tige et à feuilles



b) Trouve l'écart moyen et interprète le résultat.



Remarques :

- ✓ Plus l'écart moyen est petit, plus les données sont concentrées autour de la moyenne et plus la distribution est homogène. À l'inverse, plus l'écart moyen est grand, plus les données sont dispersées par rapport à la moyenne et plus la distribution est hétérogène.
- ✓ À lui seul, l'écart moyen ne fournit pas beaucoup d'information. Avant de se prononcer sur l'homogénéité ou l'hétérogénéité des données, il importe donc de considérer le caractère étudié et la moyenne des données. Ainsi, un écart moyen de 5 n'est pas très élevé pour des résultats en %. Cependant, si l'on parle d'un écart moyen de 5 pour des résultats d'un examen compté sur 10, alors un écart moyen de 5 serait très élevé.

C) L'écart moyen dans un tableau à données condensées

Comme vous l'avez étudié l'an dernier, il existe plusieurs types de tableau servant à représenter un groupe de données. Le tableau à données condensées nous permet de connaître l'ensemble des résultats d'une situation sans avoir à les énumérer.

Exemples :

- 1) On a demandé aux élèves de 4^e secondaire quel était le nombre de frères ou de sœurs qu'ils avaient. Les résultats sont présentés dans le tableau ci-contre.

Détermine l'écart moyen du nombre de frères et de sœurs des élèves de 4^e secondaire.

Nombre de frères et sœurs	Fréquence
0	99
1	251
2	48
3	77
4	5

TOTAL

480

- 2) Une entreprise a déterminé le nombre de semaines de vacances accordées à ces employés. On a représenté les résultats dans le graphique ci-dessous. Détermine l'écart moyen du nombre de semaines de vacances des employés de cette entreprise.

1) Moyenne

$$\bar{x} = \frac{\sum x_i}{n}$$

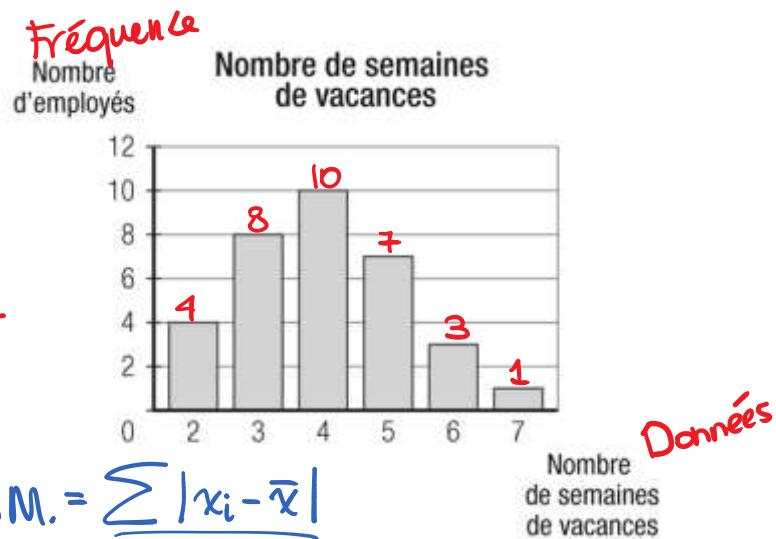
$$\bar{x} = \frac{2 \times 4 + 3 \times 8 + 4 \times 10 + 5 \times 7 + 6 \times 3 + 7 \times 1}{4 + 8 + 10 + 7 + 3 + 1}$$

$$\bar{x} = \frac{132}{33}$$

$$\bar{x} = 4$$

2) Écarts à la \bar{x}

Données	Écarts à la \bar{x}	Fréquence
2	2	4
3	1	8
4	0	10
5	1	7
6	2	3
7	3	1



$$\text{E.M.} = \sum |x_i - \bar{x}|$$

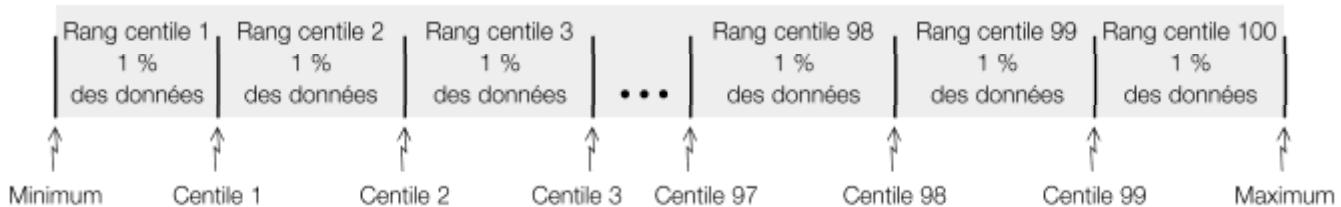
3) Écart moyen

$$\begin{aligned} \text{E.M.} &= \frac{2 \times 4 + 1 \times 8 + 0 \times 10 + 1 \times 7 + 2 \times 3 + 3 \times 1}{33} \\ &= \frac{32}{33} \\ &\approx 0,97 \end{aligned}$$

D) Le rang centile – Une mesure de position

Une mesure de position sert à situer une donnée parmi les autres données d'une distribution. Le rang centile d'une donnée est une mesure de position qui indique le pourcentage de données inférieures ou égales 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.



La formule ci-dessous permet de calculer le rang centile d'une donnée. Si le résultat n'est pas un nombre entier, on l'arrondit à l'unité supérieure.

$$\text{Rang centile de } x = \frac{\text{nombre de données} \leq x + \frac{\text{nombre de données égales à } x}{2}}{\text{nombre total de données}} \times 100$$

Exemple :

Voici une distribution comportant 158 données :

$6, 7, 8, \dots, 19$ 61 données comprises entre 8 et 19	$21, 21, 21, 23, 24, \dots, 50, 51$ 41 données comprises entre 24 et 50	$52, 55, 56, 56, 56, 57, 58, \dots, 89, 89, 90$ 36 données comprises entre 58 et 89
--	---	---

Rang centile de 21 :

$$\begin{aligned}
 R_{100}(21) &= \frac{\text{nb données} \leq 21 + \frac{\text{nb données} = 21}{2}}{\text{nb total}} \times 100 \\
 &= \frac{65 + \frac{3}{2}}{158} \times 100 \\
 &\approx 42,09 \rightarrow 43^{\text{e}} \text{ rang centile}
 \end{aligned}$$

Le rang centile de 21 est donc 43.

Ainsi, 43 % des données de la distribution lui sont inférieures ou égales.

Rang centile de 52 :

$$R_{100}(52) = \frac{\frac{\text{nb données} < 52}{n} + \frac{\text{nb données} = 52}{2}}{2} \times 100$$
$$= \frac{113 + \frac{1}{2}}{158} \times 100$$
$$\approx 71,83 \rightarrow 72^{\text{e}} \text{ rang centile}$$

Le rang centile de 52 est

donc 72.

Ainsi, 72 % des données de la distribution lui sont inférieures ou égales.

Pour repérer une donnée dont le rang centile est connu, il faut :

1. déterminer le nombre de données inférieures ou égales à la donnée recherchée en effectuant le calcul ci-dessous. Si le résultat n'est pas un nombre entier, on l'arrondit à l'unité inférieure;

$$\text{Position de la donnée} = \frac{\text{rang centile}}{100} \times \text{nombre total de données}$$

2. chercher dans la liste des données ordonnées celle qui occupe le rang trouvé.

Exemple :

Voici une distribution comportant 158 données :

$\begin{matrix} 3 \\ 65 \end{matrix}$	$\begin{matrix} 70 \\ 112^{\text{e}} \end{matrix}$	
6, 7, 8, ..., 19, 21, 21, 21, 23, 24, ..., 50, 51, 52, 55, 56, 56, 57, 58, ..., 89, 89, 90 61 données comprises entre 8 et 19	41 données comprises entre 24 et 50	36 données comprises entre 58 et 89

Donnée ayant 75 pour rang centile :

$$\#\text{donnée} = R_{100} \times \text{nb données}$$
$$= \frac{75}{100} \times 158$$
$$= 118,5 \rightarrow 118^{\text{e}} \text{ donnée}$$

La donnée ayant 75 pour rang centile correspond à la 118^e donnée de la distribution ordonnée. Cette donnée est 57.

6, 7, 8, ..., 19, 21, 21, 21, 23, 24, ..., 50, 51, 52, 55, 56, 56, 57, 58, ..., 89, 89, 90

61 données comprises entre 8 et 19

41 données comprises entre 24 et 50

36 données comprises entre 58 et 89

Donnée ayant 71 pour rang centile :

$$\begin{aligned} \text{# donnée} &= \frac{R_{100}}{100} \times n \\ &= \frac{71}{100} \times 158 \\ &\approx 112,18 \rightarrow 112^{\text{e}} \text{ donnée} \end{aligned}$$

La donnée ayant 71 pour rang centile correspond à la 112^e donnée de la distribution ordonnée. Cette donnée est 50.

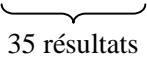
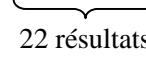
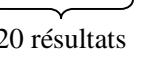
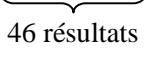
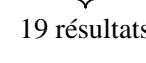
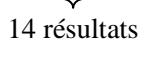
Exercices (Rang centile et écart moyen)

1. Voici les résultats en pourcentage obtenus par les élèves de deux groupes à une épreuve de mathématique. Chacun des groupes compte 20 élèves.

Résultats des élèves du groupe A										
77	87	89	97	86	91	81	90	75	91	
87	79	97	85	76	98	83	96	87	84	
Résultat des élèves du groupe B										
76	95	75	84	96	88	79	94	77	87	
98	85	75	85	88	78	77	78	75	89	

Étienne fait partie du groupe A. Il a obtenu 89 %. Olivier fait partie du groupe B. Il a obtenu le même résultat qu'Étienne, 89 %. L'association des parents de l'école remet un certain nombre de bourses aux élèves méritants. À partir du rang centile, déterminez lequel de ces deux élèves a la meilleure chance d'obtenir une bourse.

2. Voici, présentés en ordre croissant, les résultats des 255 élèves d'une école à un examen.

45	50	...	55	56	56	58	...	61	61
									
61	...	64	65	66	...	69	70	70	71
									
71	71	...	76	77	77	77	...	84	84
									
86	...	89	89	...	93	93	95	95	98
									
99	100	100							

Quel est le rang centile associé à un résultat de 70?

- | | |
|-----------------------|-----------------------|
| A) Le 28 ^e | C) Le 53 ^e |
| B) Le 48 ^e | D) Le 70 ^e |

3. Un organisme scolaire doit attribuer un rang centile à chacun des résultats obtenus par les élèves de 1^{re} secondaire inscrits en mathématique.
Ces résultats apparaissent dans le tableau ci-dessous.

Résultat	Fréquence
94	6
92	2
86	5
85	4
82	3
76	7
70	11
68	4
62	10
60	8
54	4
51	1

Quel est le rang centile d'un élève qui a obtenu un résultat de 70?

- | | |
|-------|-------|
| A) 71 | C) 59 |
| B) 67 | D) 50 |

4. Ce diagramme à tige et à feuilles présente 129 données recueillies lors d'une étude statistique.

Âge des participants et participantes à une marche de l'espoir

1	5	6	6	7	8	8	8	9	9	9
2	0	0	0	1	2	2	3	3	4	4
	7	8	9							
3	0	0	0	1	1	2	3	3	4	4
	9								6	7
4	0	2	2	2	3	3	3	4	4	4
5	0	0	1	1	2	2	2	3	4	4
6	0	1	1	2	2	3	4	4	4	5
	9	9	9	9	9	9	9	9	5	6
7	0	0	2	2	2	2	3	3	3	4
	8	9							4	5
									6	6
									7	7
									7	8

- a) Quel est le rang centile du mode de cette distribution ?
- b) Quel est le rang centile d'un marcheur ou une marcheuse âgé de 26 ans ?
- c) Quel est l'âge d'un marcheur ou une marcheuse occupant le 62^e rang centile ?

5. Lors d'une course regroupant 350 athlètes :

- Louis est arrivé 91e;
- 23% des athlètes sont arrivés avant Simone;
- François a devancé 247 personnes.

Sachant que le rang centile est calculé selon le temps dans lequel ils ont effectué la course, détermine le rang centile de chacun des trois athlètes.

3. 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 sujets issus d'une même situation.

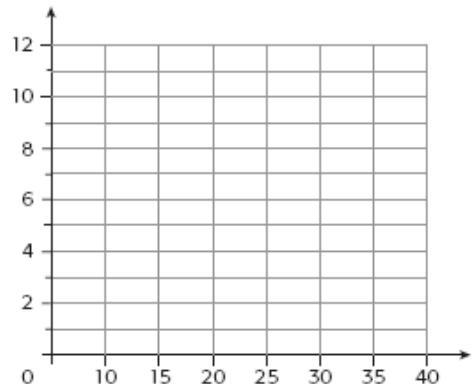
Voici quelques outils nous permettant d'analyser des distributions à deux caractères.

Exemple :

Un club d'éologie étudiant s'intéresse à la relation entre le nombre de canettes vendues à la cafétéria et le nombre de canettes recyclées. Les données recueillies sont présentées ci-dessous.

Nombre de canettes vendues	15	18	23	25	28	30	30	36
Nombre de canettes recyclées	4	2	4	5	6	8	9	11

Afin de bien se représenter la situation, construisons le nuage de points illustrant celle-ci.



A) La droite de régression

Lorsque le nuage de points d'une distribution à deux caractères présente une corrélation linéaire, la relation entre ces caractères peut être modélisée par une droite. La droite qui s'ajuste le mieux à l'ensemble des points est appelée « droite de régression ». Il existe plusieurs méthodes pour déterminer l'équation d'une telle droite.

I) La méthode de la droite de Mayer

La droite de Mayer est la droite passant par deux points moyens (P_1 et P_2) qui sont représentatifs de l'ensemble des points de la distribution.

Cette méthode est utilisée par les statisticiens lorsqu'il y a peu de données. Voici les étapes à suivre pour déterminer son équation.

Exemple 1: (6, 23) (7,26) (10,39) (13,44) (14,48) (15,55) (18,50) (19,65) (23,68) (25,72)

Étapes	Démarche
1. Ordonner les données en ordre croissant selon la première variable. <i>Remarque:</i> Pour deux valeurs égales de X , ordonner les valeurs de Y en ordre croissant.	
2. Partager la distribution en deux groupes contenant le même nombre de données. <i>Remarque:</i> Voir le tableau de la page suivante.	
3. Définir les points P_1 et P_2 en calculant la moyenne des abscisses et la moyenne des ordonnées des points de chaque groupe.	
4. Déterminer l'équation de la droite de Mayer.	L'équation de la droite de régression est :

Exemple 2

Détermine l'équation de la droite de régression à l'aide de la droite de Mayer de la distribution suivante : (2, 17) (5,14) (7,12) (7,11) (8,11) (10,10) (11,7) (13,6) (16,7) (18,7) (22,5)

Étapes	Démarche
1. Ordonner les données en ordre croissant selon la première variable. <u>Remarque:</u> Pour deux valeurs égales de X , ordonner les valeurs de Y en ordre, selon la tendance des autres points.	
2. Partager la distribution en deux groupes contenant le même nombre de données. <u>Remarque:</u> Si le nombre de données est impair, la donnée du centre est placée dans chacun des deux groupes.	
3. Définir les points P_1 et P_2 en calculant la moyenne des abscisses et la moyenne des ordonnées des points de chaque groupe.	
4. Déterminer l'équation de la droite de Mayer.	L'équation de la droite de régression est :

Exemple 3

Détermine l'équation exacte de la droite de régression à l'aide de la méthode de la droite de Mayer de la distribution suivante :

$$(3, 2) (27,17) (6,5) (14,10) (14,6) (22,21) (9,7) (19,14)$$

Exemple 4

Retournons à l'exemple présenté au départ. Détermine l'équation de la droite de Mayer qui modélise cette situation, puis estime le nombre de canettes recyclées si 40 canettes ont été vendues.

Un club d'écologie étudiant s'intéresse à la relation entre le nombre de canettes vendues à la cafétéria et le nombre de canettes recyclées. Les données recueillies sont présentées ci-dessous.

Nombre de canettes vendues	15	18	23	25	28	30	30	36
Nombre de canettes recyclées	4	2	4	5	6	8	9	11

Exercice

Afin de mieux cibler ses dépenses en publicité, une chaîne d'hôtels fait une analyse du taux d'occupation de ses chambres. Les résultats de l'analyse permettent d'établir un lien entre les frais engagés en publicité et le taux d'occupation.

Frais de publicité (en milliers de dollars)	30	27	32	25	35	22	24	35
Taux d'occupation (%)	52	45	67	55	76	48	32	72

Détermine le montant en frais de publicité qui devrait permettre d'atteindre un taux d'occupation de 80 %.

II) La méthode de la droite médiane-médiane

La droite médiane-médiane est la droite définie à partir de trois points médians, M_1 , M_2 et M_3 , représentatifs de la distribution. Cette méthode est utilisée lorsqu'il y a un grand nombre de données et/ou lorsqu'il y a des données aberrantes. Voici les étapes à suivre pour déterminer son équation.

Exemple : (2,14) (5,16) (7,12) (7,12) (8,11) (9,9) (11,7) (13,6) (15,7) (18,7) (22,2) (23,5)

Étapes	Démarche
1. Ordonner les données en ordre croissant selon la première variable. <u>Remarque:</u> Pour deux valeurs égales de X , ordonner les valeurs de Y en ordre croissant.	$\begin{array}{c c} x & y \\ \hline 2 & 14 \\ 5 & 16 \\ 7 & 12 \\ 7 & 12 \\ 8 & 11 \\ 9 & 9 \\ 11 & 7 \\ 13 & 6 \\ 15 & 7 \\ 18 & 7 \\ 22 & 2 \\ 23 & 5 \\ \hline & 16 \\ & 14 \\ & 12 \\ & 12 \\ & 11 \\ & 9 \\ & 7 \\ & 6 \\ & 7 \\ & 7 \\ & 5 \\ & 2 \end{array}$ $x_{m1} = 6 \quad y_{m1} = 13$ $x_{m2} = 10 \quad y_{m2} = 8$ $x_{m3} = 20 \quad y_{m3} = 6$
2. Partager la distribution en trois groupes. Le premier et le troisième groupe doivent compter le même nombre de données. <u>Remarque:</u> Les trois groupes doivent avoir le plus possible le même nombre de données.	
3. Définir trois points, M_1 , M_2 et M_3 , dont les coordonnées x sont les médianes des abscisses de chaque groupe et les coordonnées y sont les médianes des ordonnées de chaque groupe.	
4. Déterminer les coordonnées du point P qui est le point moyen de M_1 , M_2 et M_3 .	$x_p = \frac{6+10+20}{3} = 12 \quad y_p = \frac{13+8+6}{3} = 9$
5. Trouver l'équation de la droite médiane-médiane, sachant : <ul style="list-style-type: none"> • qu'elle est parallèle à la droite qui passe par les points M_1 et M_3 ; • qu'elle passe par le point P. 	<p>a) <u>Taux de variation (M_1 et M_3)</u></p> $+14 \quad (6, 13) \rightarrow 7 \quad a = \frac{\Delta y}{\Delta x} = \frac{-7}{14} = -\frac{1}{2}$ <p>b) <u>Valeur initiale (pt P)</u></p> $y = -\frac{1}{2}x + b$ $9 = -\frac{1}{2} \cdot 12 + b$ $9 = -6 + b$ $15 = b$ <p>L'équation de la droite de régression est :</p> $y = -\frac{1}{2}x + 15$

Exemple 2

Trouve l'équation de la droite médiane-médiane pour les coordonnées suivantes :

(3, 2) (9, 6) (19, 14) (22, 21) (28, 17) (14, 6) (14, 10).

x	y
3	2
9	6
14	6
14	10
19	14
22	21
28	17

1) Coordonnées

M₁(6, 4)

M₂(14, 10)

M₃(25, 19)

Taux de variation

$$P\left(\frac{6+14+25}{3}, \frac{4+10+19}{3}\right)$$

P(15, 11) - Valeur initiale

2) Équation dte régression

a) Taux de var.

$$+19 \downarrow \begin{pmatrix} (6, 4) \\ (25, 19) \end{pmatrix} +15$$

$$a = \frac{\Delta y}{\Delta x} = \frac{15}{19}$$

b) Val. initiale

$$y = \frac{15}{19}x + b$$

$$11 = \frac{15}{19} \cdot 15 + b$$

$$11 = \frac{225}{19} + b$$

$$-\frac{225}{19} = \frac{-225}{19}$$

$$-\frac{16}{19} = b$$

$$y = \frac{15}{19}x - \frac{16}{19}$$

Exemple 3

Détermine l'équation de la droite de régression de la distribution ci-dessous selon la méthode médiane-médiane.

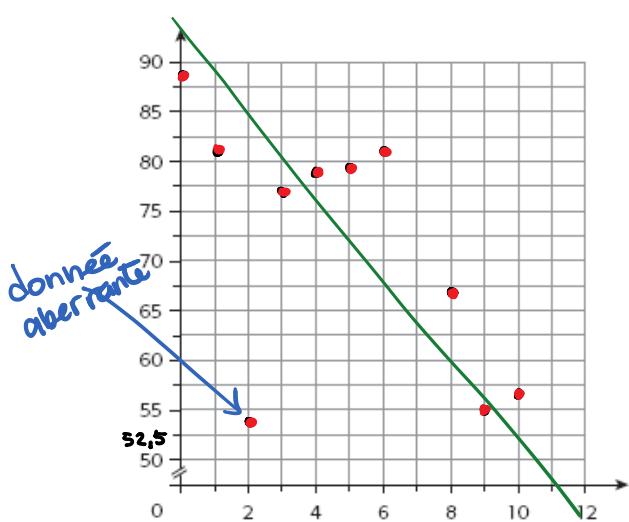
(3,5) (3,6) (4,5) (5,6) (5,8) (6,6) (7,2) (8,9)

Exemple 4

Une enseignante de mathématique désire savoir si le nombre de jours d'absence de ses élèves influe réellement sur leurs résultats scolaires. Le tableau ci-dessous présente la relation entre le nombre de jours d'absence par année de 10 élèves choisis de façon aléatoire et la note moyenne qu'ils ont obtenue dans leur bulletin de fin d'année.

Nombre de jours d'absence	8	4	2	0	9	5	3	1	10	6
Note moyenne au bulletin de fin d'année	67	79	54	88	55	79	77	81	57	81

- a) Représente ces données à l'aide d'un nuage de points.



- b) Détermine la note moyenne au bulletin d'un élève qui a été absent 6 jours.

x	y
0	88
1	81
2	54
3	77
4	79
5	79
6	81
7	67
8	67
9	55
10	57
11	55

1) Coordonnées

$$M_1(1, 81)$$

$$M_2(4,5, 79)$$

$$M_3(9, 57)$$

$$P\left(\frac{29}{6}, \frac{217}{3}\right)$$

moyenne
des x
médians

moyenne
des y
médians

2) Droite de régression

a) Taux de variation

$$\rightarrow \Delta y = (1,81) - (9,57) = -24$$

$$\Delta x = 8 - 0 = 8$$

b) Val. initiale

$$y = -3x + b$$

$$\frac{217}{3} = -3 \cdot \frac{29}{6} + b$$

$$\frac{217}{3} = -\frac{29}{2} + b$$

$$\frac{521}{6} = b$$

$$y = -3x + \frac{521}{6}$$

PdM p.299

3) Calculer la note

$$y = -3 \cdot 6 + \frac{521}{6}$$

$$y \approx 68,83\%$$

B) Corrélation

Étudier la **corrélation** entre deux variables statistiques, c'est décrire le lien entre deux caractères quantitatifs d'une distribution. Il est possible de **qualifier le type, le sens et l'intensité** d'une corrélation entre deux variables.

- Le **type** de corrélation correspond au modèle mathématique qui décrit le mieux le lien entre les variables (linéaire, quadratique, exponentiel, ...).
- Une corrélation est dite **positive** ou **négative** selon le **sens** de variation des variables.
 - **Positive** : Lorsque les valeurs d'une variable augmentent (ou diminuent), les valeurs de l'autre variable augmentent (ou diminuent) aussi.
 - **Négative** : Lorsque les valeurs d'une variable augmentent (ou diminuent), les valeurs de l'autre variable diminuent (ou augmentent).
- Une corrélation est dite **nulle, faible, moyenne, forte ou parfaite** selon **l'intensité** du lien entre les variables.

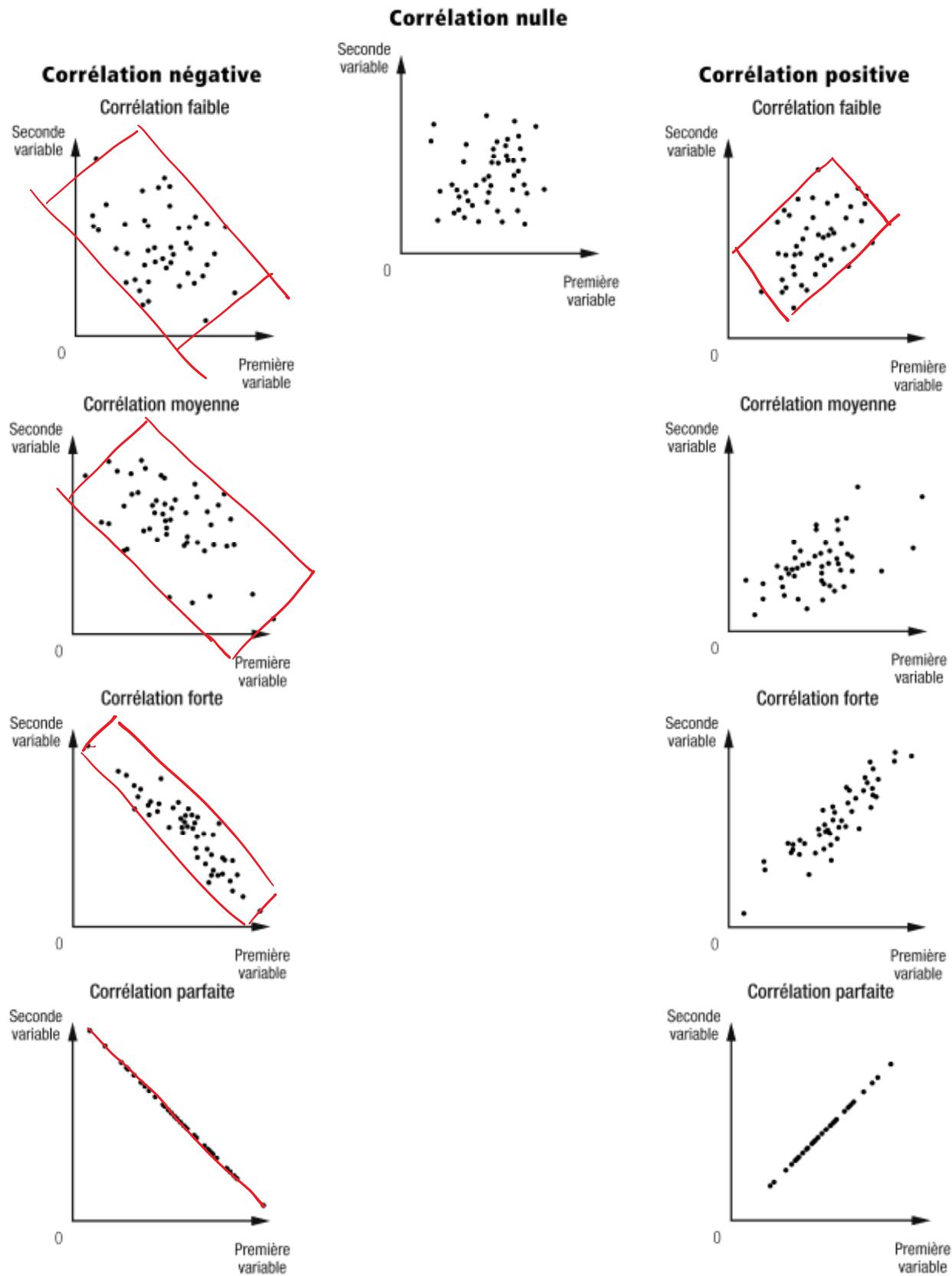
C) NUAGE DE POINTS

Un nuage de points permet de représenter une distribution à deux variables et de qualifier le type, le sens et l'intensité de la corrélation qui peut exister entre les deux variables.

Dans un nuage de points :

- l'une des variables est associée à l'axe des abscisses et l'autre variable, à l'axe des ordonnées;
- chaque couple de la distribution est représenté par un point;
- la corrélation est dite linéaire lorsque les points tendent à former une droite oblique;
- la corrélation est dite nulle si les points sont distribués au hasard, et de plus en plus forte au fur et à mesure que les points se rapprochent d'une droite oblique.

Voici les principales caractéristiques qu'on attribue à une corrélation linéaire :



D) Coefficient de corrélation

Il est possible de quantifier l'intensité de la corrélation linéaire entre deux variables statistiques à l'aide d'un nombre de l'intervalle $[-1, 1]$. Ce nombre est appelé le coefficient de corrélation et on le désigne par la lettre r .

Coefficient de corrélation		Signification
Négatif	Positif	
Près de 0	Près de 0	Indique une corrélation linéaire nulle entre les deux variables.
Près de -0,5	Près de 0,5	Indique une corrélation linéaire faible entre les deux variables.
Près de -0,75	Près de 0,75	Indique une corrélation linéaire moyenne entre les deux variables.
Près de -0,87	Près de 0,87	Indique une corrélation linéaire forte entre les deux variables.
Égal à -1	Égal à 1	Indique une corrélation linéaire parfaite entre les deux variables.

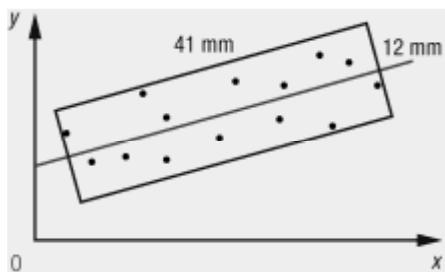
Il existe plusieurs méthodes pour approximer le coefficient de corrélation linéaire d'une distribution à deux variables.

L'une d'elles est une méthode d'estimation graphique faisant intervenir un rectangle dans un nuage de points. Cette méthode consiste à :

1. représenter par un nuage de points la distribution à deux variables;
2. tracer une droite représentative de la majorité des points;
3. construire sur le nuage de points le rectangle de plus petites dimensions englobant tous les points significatifs et dont deux des côtés sont parallèles à la droite;
4. approximer le coefficient de corrélation linéaire entre les deux variables à l'aide de la formule suivante :

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

Exemple 1 : Trouve le coefficient de corrélation du nuage de points ci-dessous, puis donne la signification de ce coefficient.



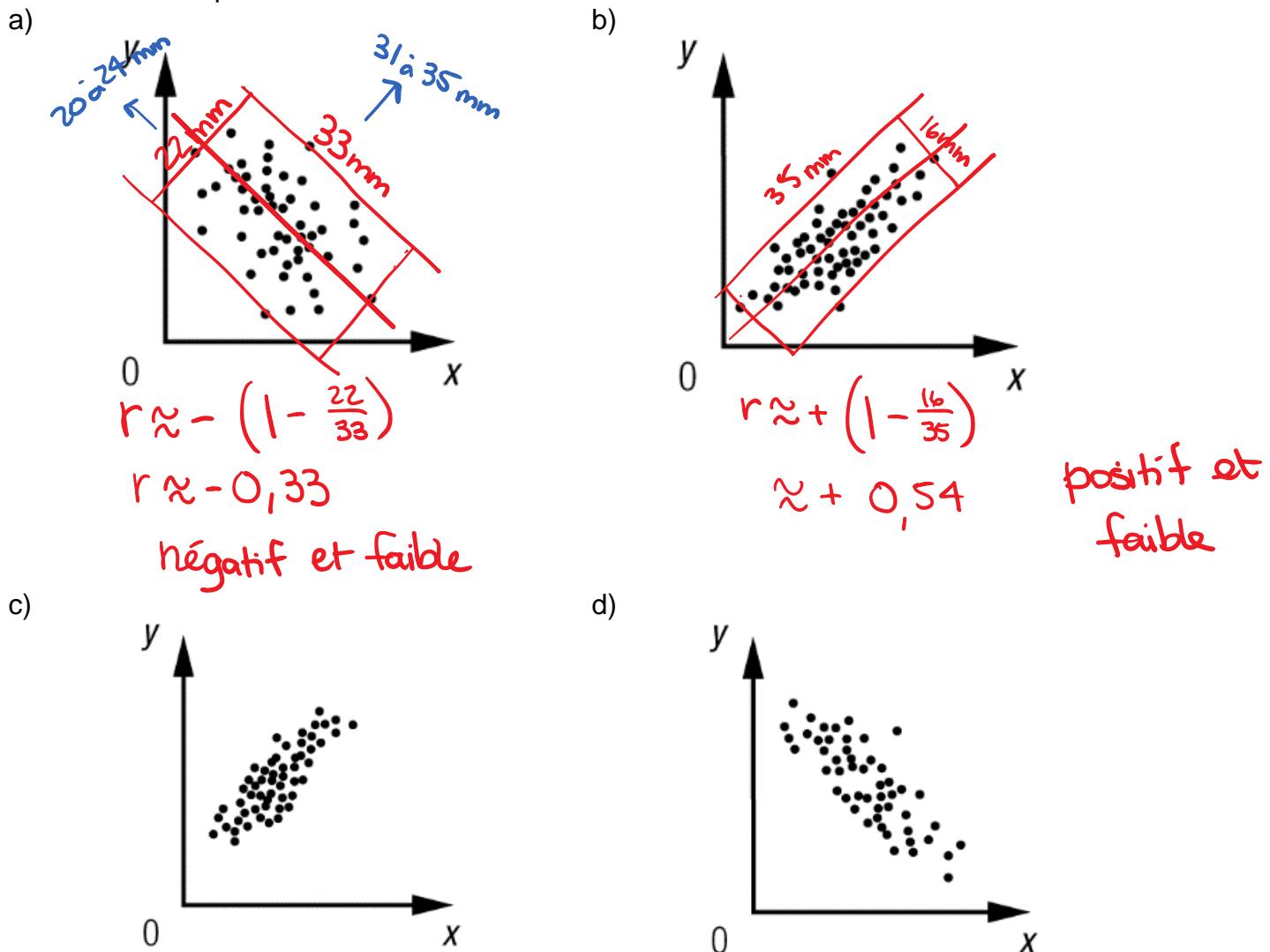
$$r \approx + \left(1 - \frac{\text{mes. petit côté}}{\text{mes. grand côté}} \right)$$

$$r \approx + \left(1 - \frac{12}{41} \right)$$

$$\approx +0,71$$

Positive et moyenne

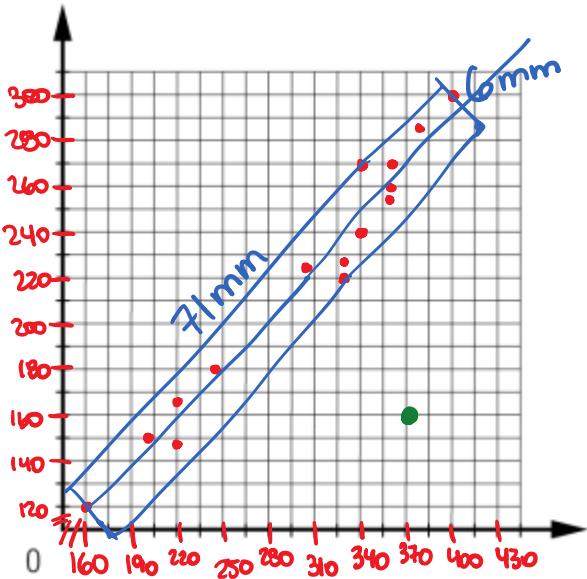
Exemple 2 : Qualifiez le sens et l'intensité de la corrélation linéaire entre les variables des nuages de points ci-dessous.



Exemple 3 : Lors d'une expérience, on a laissé tomber un ballon de différentes hauteurs et on a noté la hauteur du premier rebond. Chacun des couples suivants indique respectivement la hauteur initiale (en cm) du ballon et la hauteur (en cm) du premier rebond.

(360, 254), (320, 228), (340, 255), (240, 180), (300, 225), (380, 285), (200, 150), (220, 148), (240, 240), (360, 270), (400, 300), (360, 260), (220, 165), (320, 220), (160, 120)

a) Tracez un nuage de points représentant cette situation.



b) Qualifiez la corrélation entre la hauteur initiale du ballon et la hauteur du premier rebond.

$$r \approx +\left(1 - \frac{6}{71}\right)$$

$$\approx +0,92$$

positive et forte

Remarques :

Lorsqu'on a recours à une droite de régression pour estimer la valeur d'une variable à partir d'une autre, il faut toujours s'interroger quant à la fiabilité de la valeur calculée. Généralement, plus la corrélation linéaire est forte, plus il y a de chances que la prédiction soit fiable.

Lorsqu'on n'a pas accès aux technologies pour déterminer l'équation de la droite de régression, il est plus simple de déterminer celle de la droite de Mayer. Cependant, il est préférable d'avoir recours à l'équation de la droite médiane-médiane si la distribution présente des points aberrants, puisque la droite de Mayer est très sensible aux données extrêmes.

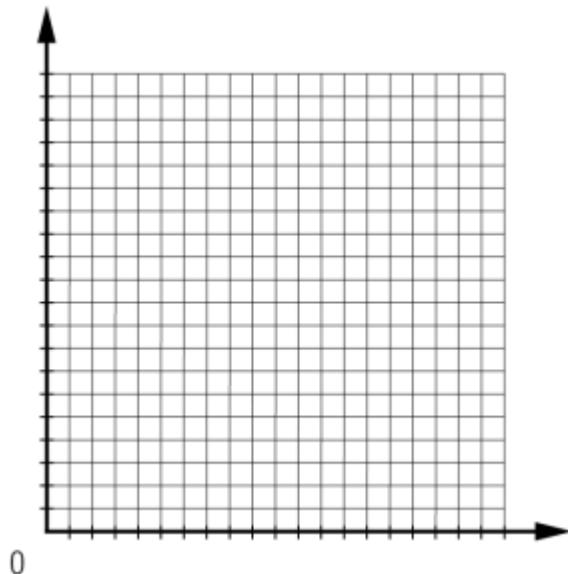
Si $r \approx \pm 0,75$ ou plus

Exemple

Le tableau ci-dessous présente la relation entre l'âge de différents anacondas élevés en captivité et leur longueur.

Âge de l'anaconda	8	5	4	4	3	3	7	9	6
Longueur (m)	9	6	5	8	6	5	2	12	6

- a) Un biologiste désire prédire l'âge d'un anaconda mesurant 7 m. Il ne dispose d'aucune autre information sur ce serpent. Dans ce cas précis, est-il préférable qu'il utilise la méthode de la droite de Mayer ou celle de la droite médiane-médiane ?



- b) Prédis l'âge de l'anaconda de 7 m en utilisant la méthode privilégiée en a.

E) Tableau à double entrée

Un tableau à double entrée permet de représenter une distribution à deux variables, et de qualifier le type et le sens de la corrélation qui peut exister entre les deux variables.

Dans un tableau à double entrée :

- l'une des variables est associée à la première colonne et l'autre variable, à la première ligne ;
- les données peuvent être regroupées en classes ;
- la ligne et la colonne des totaux indiquent les effectifs des deux variables ;
- la case inférieure droite comporte l'effectif total de la distribution ;
- la corrélation est dite linéaire lorsque la majorité des effectifs suivent l'une des deux diagonales.

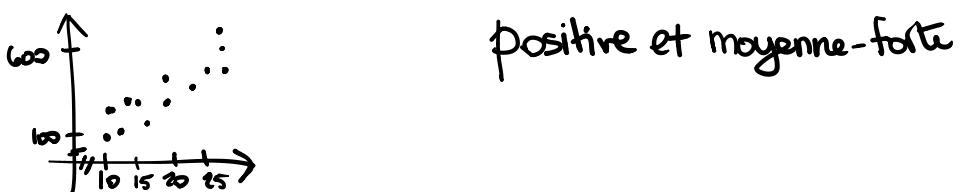
Exemple

On essaie d'établir le lien entre l'âge et le nombre de branches de 20 arbres de la même essence. Le tableau à double entrée ci-dessous représente les données recueillies.

Arbres de la même essence

Âge (années) χ	Nombre de branches y						Total
		[100, 200[[200, 300[[300, 400[[400, 500[[500, 600[
[10, 15[2	1	0	0	0	3	3
[15, 20[0	3	0	0	0	3	3
[20, 25[0	0	4	2	0	6	6
[25, 30[0	0	1	5	0	6	6
[30, 35[0	0	0	0	2	2	2
Total	2	4	5	7	2	20	

Qualifiez le sens et l'intensité de la corrélation linéaire entre les deux variables de cette distribution.



Si les points se regroupent autour de la diagonale décroissante, alors la corrélation est positive.

Exercice

Qualifiez le sens et l'intensité de la corrélation linéaire entre les deux variables de cette distribution.

$x \backslash y$	10	11	12	13	14	15	Total
1	1	1	3	0	0	0	5
2	1	1	2	0	0	0	4
3	3	1	1	1	2	0	8
4	0	0	0	1	2	0	3
5	0	0	1	1	0	2	4
Total	5	3	7	3	4	2	24

positif et faible.