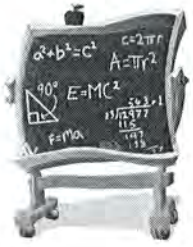
Chapitre 6 : Faire le point

Les statistiques



Nom : Catherine

Groupe : _____

Cours 1 : Rappel

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.

Mesure de tendance centrale	Distribution de données condensées	Distribution de données groupées en classes
Le **mode** est une mesure qui indique le centre de concentration d'une distribution.	Le mode est la valeur ou la modalité ayant l'effectif le plus élevé. la plus fréquente	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.
La **médiane** est une mesure qui indique le centre de position d'une distribution.	Dans une distribution **ordonnée:** • s'il y a un nombre **impair** de données, la médiane est la donnée du centre; • s'il y a un nombre pair de données, la médiane est la moyenne des deux données du centre.	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.
La **moyenne** est une mesure qui indique le centre d'équilibre d'une distribution.	$\text{Moyenne} = \frac{\text{somme des produits des valeurs par leur effectif}}{\text{nombre de données}}$	$\text{Moyenne} = \frac{\text{somme des produits des milieux des classes par leur effectif}}{\text{nombre de données}}$

Moyenne pondérée :

La moyenne d'un certain nombre de valeurs n'ayant pas toutes la même importance est appelée une **moyenne pondérée**.

Ex. : Un cours de géographie comporte trois étapes. En tenant compte de la note obtenue à chacune des étapes par un ou une élève et de l'importance relative de chacune des étapes, calcule la moyenne pondérée.

Étape	Note (%)	Pondération (%)
1	75	20
2	72	30
3	88	50

Moy = $75\% \cdot 20\% + 72\% \cdot 30\% + 88\% \cdot 50\%$

= $0,75 \cdot 0,2 + 0,72 \cdot 0,3 + 0,88 \cdot 0,5$

= $0,806$ ou $80,6\%$

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.

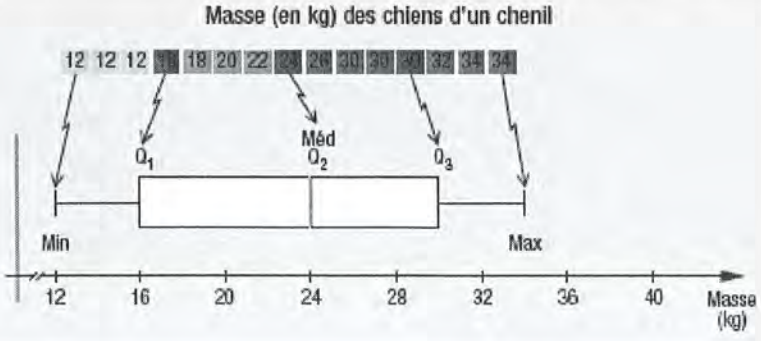
Mesure de dispersion	Distribution de données condensées	Distribution de données groupées en classes
L'**étendue** est une mesure qui indique jusqu'à quel point les données sont regroupées ou éloignées les unes des autres dans une distribution.	L'étendue est l'écart entre la donnée la plus élevée et la donnée la moins élevée. $E = Val\ max - Val\ min$	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.

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». On note généralement le premier quartile par «Q_1 », le deuxième quartile par «Q_2 » et le troisième quartile par «Q_3 ».

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. **Dans un diagramme de quartiles, chaque quart comprend le même nombre de données.**

Ex.: On a construit le diagramme de quartiles correspondant à la distribution ci-contre.



Voici, pour 3 distributions différentes, des exemples du mode, de la médiane, de la moyenne et de l'étendue

Ex.: 1) Voici une distribution ordonnée comportant 15 données:

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

Mode = 8

Médiane = 6 (8e donnée)

$$\text{Moyenne} = \frac{2+2+2+3+3+4+5+6+7+8+8+8+8+10+11}{15} = 5{,}8$$

Étendue = 11 − 2 = 9

2) Voici un tableau de distribution de données condensées:

Familles

Nombre d'enfants	1	2	3	4	5	Total
Effectif	6	16	12	10	9	53

Mode = 2 enfants, car l'effectif le plus élevé est 16.

Médiane = 3 enfants (27e donnée)

$$\text{Moyenne} = \frac{1\times 6+2\times 16+3\times 12+4\times 10+5\times 9}{53} = 3 \text{ enfants}$$

Étendue = 5 − 1 = 4 enfants

3) Voici un tableau de distribution de données groupées en classes:

Chenil

Taille des chiens (cm)	[20, 40[	[40, 60[	[60, 80[	[80, 100[	[100, 120[	Total
Effectif	18	19	13	20	10	80

Classe modale = [80, 100[cm

Mode ≈ 90 cm

Classe médiane = [60, 80[cm

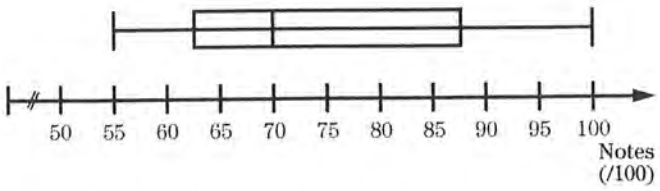
Médiane ≈ 70 cm

$$\text{Moyenne} \approx \frac{30\times 18+50\times 19+70\times 13+90\times 20+110\times 10}{80} = 66{,}25 \text{ cm}$$

Étendue = 120 − 20 = 100 cm

Cours 1 Exercices

1. Le diagramme de quartile suivant représente les résultats de 100 étudiants d'un cours de mathématiques de secondaire 4.



Répondez aux questions suivantes en vous basant sur le diagramme.

a) Quelle est la médiane 70 %

b) Que vaut Q1 ? 62,5 %

c) Que vaut Q2 ? 70 %

d) Que vaut Q3 ? 87,5 %

e) Est-ce qu'il y a des étudiants qui ont obtenu 100 % ? Oui

f) Quel a été le résultat le plus faible ? 55 %

g) Quel a été le résultat le plus élevé ? 100 %

h) Peut-on dire que 50 % des élèves de ce cours ont obtenu au moins 70 % ?

Justifiez.

Oui, car comme la médiane est de 70% et quelle sépare la distribution en 2 parties égales, il y a au moins 50% des élèves qui on 70% et plus

2. François fait de la course à pied. Deux fois par semaine, il court une distance de 10 km. Voici les 10 temps en minutes qu'il a compilés lors de ses dernières courses.

15	18	22	20	18
24	16	15	17	18

15, 15, 16, 17, 18, 18, 18, 20, 22, 24

À partir des résultats de François, répondez aux questions suivantes.

a)Détermine la médiane : 18

b) Détermine le mode : 18

c) Détermine l'étendue : 9

d) Détermine la moyenne : 18, 3

3. Élisabeth et Guillaume sont deux professeurs de mathématiques. Ils comparent les résultats de 10 de leurs étudiants au dernier examen. Les notes en % sont les suivantes : 62, 71, 72, 77, 78, 81, 82, 82, 93, 96

Classe d'Élisabeth = { 82, 93, 78, 62, 82, 96, 72, 77, 81, 71 }

Classe de Guillaume = { 91, 68, 66, 88, 76, 80, 72, 79, 81, 85 }

66, 68, 72, 76, 79, 80, 81, 85, 88, 91

a) Pour chaque classe, construisez le diagramme de quartile

Pour Elisabeth

$Q_1 = 72$ $Q_3 = 82$

$Q_2 = 79,5$ Val min = 62

Val max = 96

Pour Guillaume

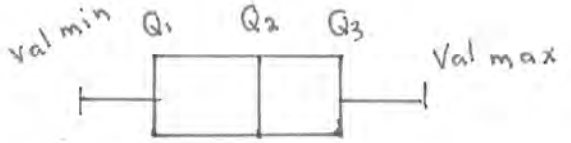
$Q_1 = 72$ $Q_3 = 85$

$Q_2 = 79,5$

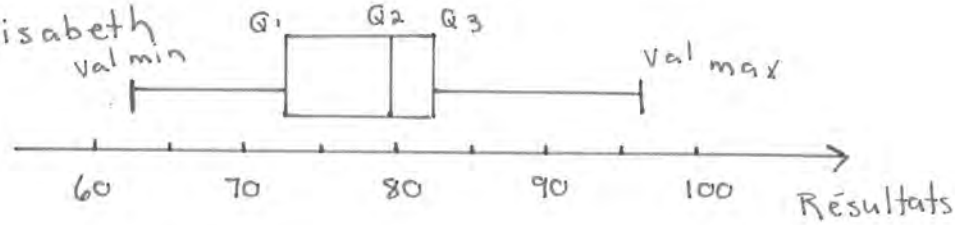
Val min = 66

Val max = 91

Guillaume



Elisabeth



4. On présente ci-dessous le relevé de notes de Julie.
Quelle est sa moyenne pondérée ?

Matière	Résultat %	Crédit
Français	95	6
Histoire	92	4
Mathématique	88	6
Anglais	88	4
Piano	86	2
Sciences	85	2

24 crédits

$\bar{x}$ pondérée = 90 %

$\bar{x}$ pondérée = $\frac{21,6}{24} \times 100$

5. La tableau ci-dessous présente l'âge des finissants et finissantes d'une université. Dans cette distribution :

Âge	Effectif
[20,25[	87
[25,30[	231
[30,35[	145
[35,40[	78
[40,45[	68
[45,50[	32

a) Quel est le nombre total de finissants(es) ?

641

b) Quelle est la classe modale ?

[25, 30[

c) Quelle est l'étendue ?

30

d) Quelle est la classe médiane ?

[30, 35[

e) Quelle est la moyenne d'âge des finissants(es)?

$\bar{x}$ = 31,76 ans

6. Remplissez le tableau suivant :

Distribution	Mode	Médiane	Moyenne	Étendue
3,6,7,7,8,8,8,12,14,15,17,18, 21,23,28,30,30	8	14	15	27
5,5,5,5,5,5,6,6,6,7,8,9,9,9, 9,9,9,10	5 et 9	6,5	≈ 7,06	5
12,14,15,23,24,25,33,34, 35,44,44,44,44,44	44	33,5	≈ 31,07	32
6,7,12,14,16,18,20,22,25, 27,29,34,37	Aucun	20	≈ 20,54	31

Cours 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 caractère (un seul sujet).

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 d'un ou des deux côtés d'une colonne, appelée tige. Dans un tel diagramme :

- Chaque ligne est associée à une classe
- Chaque données est décomposée en deux parties se trouvant sur une même ligne : la partie constituée des ses premiers chiffres formant la tige et la partie constituée des ses derniers chiffres formant une feuille.
- Un titre et une légende doit obligatoirement accompagner chaque diagramme.

Exercices : Pour chacune des distributions, tracez le diagramme à tige et à feuille :

1. Les données de la distribution ci-dessous correspondent aux pulsations cardiaques de 31 personnes à la suite d'un effort physique.
   70, 73, 73, 76, 78, 81, 82, 85, 85, 87, 88, 88, 89, 90, 92, 92, 92, 96, 97, 99, 101, 101, 101, 104, 106, 106, 107, 112, 114, 115, 118.

Pulsations cardiaques
(nbre de battements /min)

7	0 3 3 6 8
8	1 2 5 5 7 8 8 9
9	0 2 2 2 6 7 9
10	1 1 1 4 6 6 7
11	2 4 5 8

légende 7|0
signifie 70

2. Les données des deux distributions ci-dessous correspondent à la taille (en mm) de 58 saumons.

Mâles : 115, 148, 154, 180, 190, 195, 218, 225, 232, 232, 258, 259, 273, 273, 306, 311, 315, 342, 343, 369, 385, 390, 435, 444, 445, 446, 450, 462, 473, 538, 555, 566, 567, 570.

Femelles : 105, 112, 136, 147, 152, 180, 181, 258, 259, 271, 271, 276, 332, 341, 362, 363, 370, 399, 407, 419, 489, 500, 525, 531,.

Taille de 58 saumons (mm)

Mâle		Femelle
95 90 80 54 48 15	1	05 12 36 47 52 80 81
73 73 59 58 32 32 25 18	2	58 59 71 71 76
90 85 69 43 42 15 11 06	3	32 41 62 63 69 70 99
73 62 50 46 45 44 35	4	07 19 89
70 67 66 55 38	5	00 25 31

légende: 15|1|05
signifie 115 mm Mâle
105 mm Femelle

3. Voici un diagramme à tige et à feuille représentant le nombre de dvd que les familles de ton quartier possèdent à la maison.

Nombre de Dvd des familles de ton quartier

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

Légende : 1|0 signifie 10

Déterminez le mode, la moyenne et la médiane de cette distribution.

Mode = 12

Moyenne = (1 + 2 + 5 + 6 + 8 + 10 + 11 + 12 + 12 + 18 + 30 + 33 + 37 + 39 + 40 + 41 + 42 + 43) ÷ 18

$\bar{x} = 21,\overline{6}$

Médiane = Med = 15

Mesure de dispersion : Écart moyen

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.

Voici comment calculer l'écart moyen :

$$EM = \frac{\text{Somme des écarts à la moyenne}}{\text{Nombre total de données}}$$

Exemple : Voici une distribution comportant 8 données :
1, 4, 5, 6, 8, 8, 9, 11. La moyenne de cette distribution est de 6,5.
Complète le tableau suivant :

Donnée x_i	Moyenne $\overline{x}$	Écart à la moyenne
1	6,5	$\lvert 1 - 6,5 \rvert = 5,5$
4	6,5	2,5
5	6,5	1,5
6	6,5	0,5
8	6,5	1,5
8	6,5	1,5
9	6,5	2,5
11	6,5	4,5

Quel est l'écart entre la température d'hier qui était de 2 °C et celle d'aujourd'hui qui est de -12 °C ? 14 °C Ta réponse est -elle positive ou négative ? positive

L'écart entre deux données est toujours positive, c'est pourquoi on utilise le symbole $|\,|$.
↳ valeur absolue

Maintenant, calculez l'écart moyen :

$$EM = \frac{5,5 + 2,5 + 1,5 + 0,5 + 1,5 + 1,5 + 2,5 + 4,5}{8}$$

$$EM = 2,5$$

Exercice : Voici une distribution comportant les résultats des 9 patineurs lors de la dernière compétition : 7,6 ; 8,4 ; 5 ; 9,1 ; 6,8 ; 5,6 ; 7,2 ; 7; 9,4. Notez que la note maximale est de 10 points pour ce type d'épreuve. Détermine l'écart moyen.

1) Trouvons la moyenne:

$$\overline{x} = \frac{7,6 + 8,4 + 5 + 9,1 + 6,8 + 5,6 + 7,2 + 7 + 9,4}{9}$$

$\overline{x} = 7,34$

2)

Donnée :	Ecart à la moyenne
7,6	0,26
8,4	1,06
5	2,34
9,1	1,76
6,8	0,54
5,6	1,74
7,2	0,14
7	0,34
9,4	2,06

3)

$$EM = (0,26 + 1,06 + 2,34 + 1,76 + 0,54 + 1,74 + 0,14 + 0,34 + 2,06) \div 9$$

$EM = 1,14$

****Important** Plus l'écart moyen est petit, plus les données de la distribution sont** concentrées **autour de la moyenne. Dans le cas des résultats à un examen, cela signifie qu'il y a moins de données éloignées par exemple.**

Exercices : Volume : ai-je bien compris p. 55 # 1a) et p. 57
p.63 # 1 ab, # 2a, #5, #6

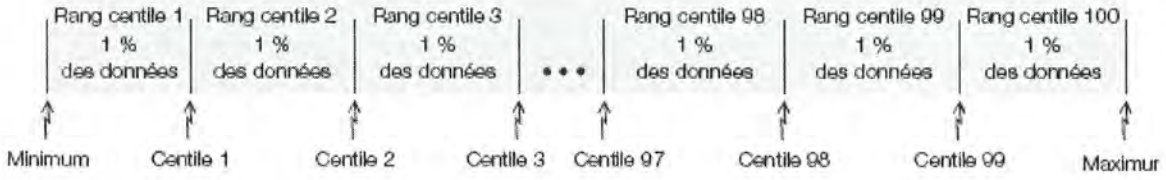
Mini-test #1 au prochain cours

Cours 3 :
Les mesures de position

Rang centile :

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 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-ensemble 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.

$$\text{Rang centile d'une donnée} = \left(\frac{\text{Nombre de donnée inf à cette donnée} + \frac{1}{2}\text{Nbre de donnée égale}}{\text{Nbre total de données}} \right) \times 100$$

****IMPORTANT**** Si le résultat n'est pas un nombre entier, on prend toujours l'unité **supérieure.**

Exercices :

Voici une distribution comportant 158 données :

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

1. Détermine le rang centile de 21 :

$\frac{1}{2} \cdot 3$

R/100 de 21 = $\left(\frac{65 + 1,5}{158} \right) \times 100 = 42,09$

Rép 43e R/100

2. Détermine le rang centile de 52 :

R/100 de 52 = $\left(\frac{113 + 0,5}{158} \right) \times 100 = 71,84$

Donc 72e R/100

Comment déterminer une donnée si l'on connaît son rang centile ?

Maintenant que tu sais comment calculer le rang centile d'une donnée apprenons à retrouver dans la distribution la donnée dont tu connais le rang centile.

- ✓ Étape 1 : Détermine 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 prend l'unité INFÉRIEURE.**

$$\frac{\text{Rang centile}}{100} \times \text{Nbre total de données}$$

- ✓ Chercher dans la liste des données **ordonnées** celle qui occupe la position trouvée.

Exercices :

Ex.: Voici une distribution comportant 158 données:

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

1. Détermine la donnée ayant 75 pour rang centile :

$\text{Position} = \frac{75}{100} \times 158$

$= 118,5$

Donc on recherche la 118e donnée

Rép : 57

2. Détermine la donnée ayant 71 pour rang centile :

$\text{Position} = \frac{71}{100} \times 158 = 112,18$

Donc on recherche la 112e donnée

Rép : 50

Exercices : Volume : ai-je bien compris p. 59
p. 66 # 13b), 14 et 16

(voici les réponses pour certains numéros de votre devoir d'aujourd'hui :
p. 59 a) 1) 61e rang centile 2) 58e rang centile 3) 52e rang centile
p. 66 # 13 b) 48e rang centile
#14 a) 1) 22e rang centile 2) 62e rang centile 3) 87e rang centile
16 a) 65e rang centile
Pour les autres numéros, consultez votre corrigé habituel)
Mini-test #2 au prochain cours

Cours 4 :
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.

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

Exemple : On considère la mesure du pied droit et la taille de chacun des joueurs d'une équipe de basket-ball. Ces deux mesures sont inscrites dans le tableau suivant.

Joueur	Mesure du pied (cm)	Taille (cm)	Joueur	Mesure du pied (cm)	Taille (cm)	Joueur	Mesure du pied (cm)	Taille (cm)
1	27,5	178	5	28,5	181	9	27,5	179
2	26,5	179	6	28,0	180	10	26,0	172
3	25,0	172	7	29,5	185	11	24,5	170
4	31,0	186	8	28,0	183	12	29,0	181

Dans cet exemple, l'unité statistique correspond à chacun des joueurs de l'équipe et les caractères étudiés sont la mesure du pied droit (en cm) et la taille du joueur (en cm).

Y a-t-il vraiment un lien entre la taille du joueur de basket-ball et la mesure de son pied droit ? Pour répondre à cette question, étudions la corrélation.

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.
- Une corrélation est dite positive ou négative selon le sens de variation des variables.

> **Positive :** lorsque les valeurs des variables varient dans le même sens. Les deux augmentent ou les deux diminuent.

> **Négative :** lorsque les valeurs des variables varient dans le sens contraire. Une augmente et l'autre diminue.

- Une corrélation est dite **nulle, faible, forte ou parfaite** selon l'intensité du lien entre les variables.

Les modes de représentation d'une distribution à deux caractères :

1. Le tableau à double entrée

Le tableau à double entrée est un outil permettant d'organiser les données d'une distribution à deux caractères. Chaque couple de la distribution est représenté dans le tableau à double entrée (par un trait verticale ou par un crochet) afin d'en faire le décompte.

Complète le tableau à double entrée ci-dessous à l'aide des données de la distribution précédente.

Mesure du pied (cm) / Taille (cm)	[24,26[	[26,28[	[28,30[	[30,32[	Total
[170, 175[	II	I			3
[175, 180[		III			3
[180, 185[			IIII		4
[185, 190[			I	I	2
Total	2	4	5	1	12

Ce tableau à double entrée montre que la corrélation entre la taille du joueur et la mesure de son pied est linéaire positive

Certaines conclusions peuvent être tirées à partir du tableau. Voici quelques exemples :

- 75 % des joueurs de basket-ball mesurent plus de 175 cm.
- 1 joueurs sur 6 a un pied droit mesurant moins de 26 cm.
- Etc.

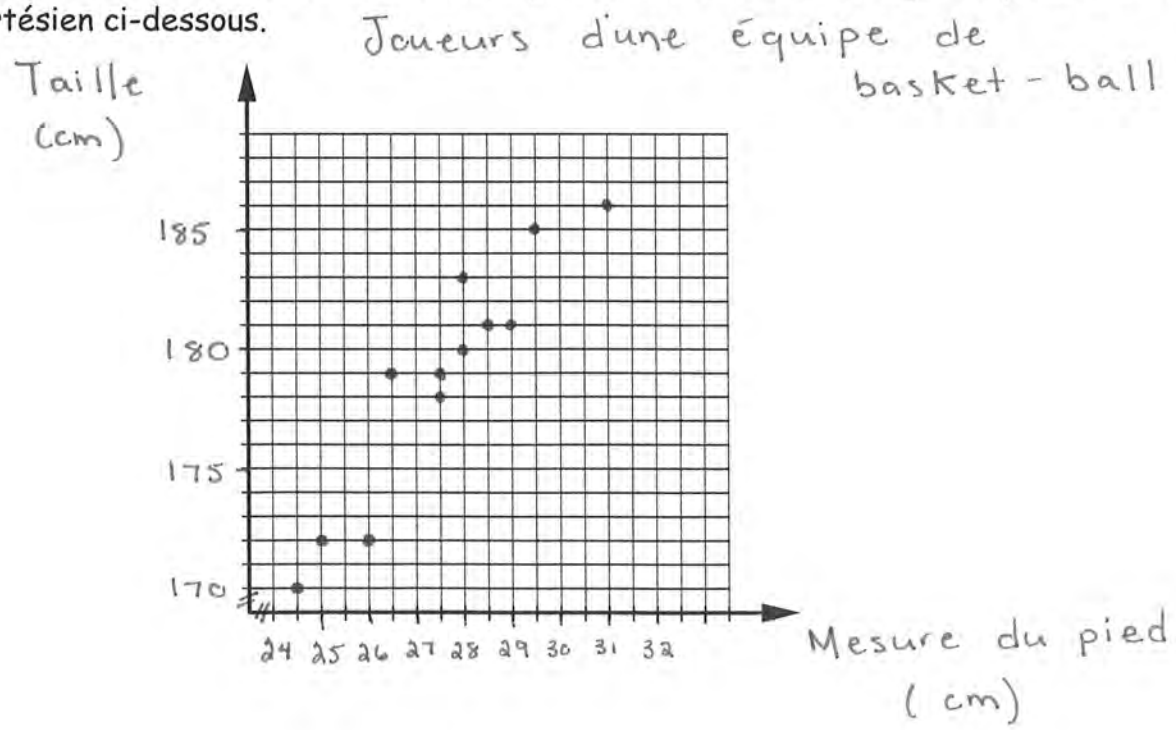
2. Le nuage de points :

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 (c'est-à-dire du lien) qui peut exister entre les deux variables.

Dans un nuage de points :

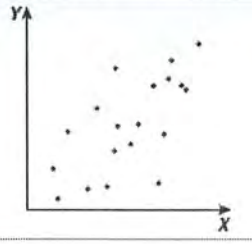
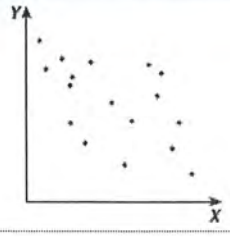
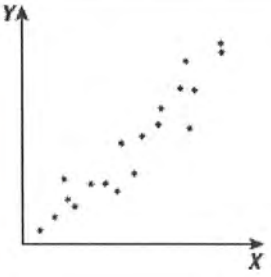
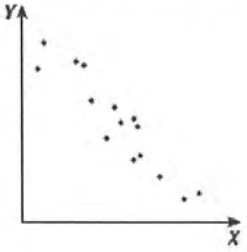
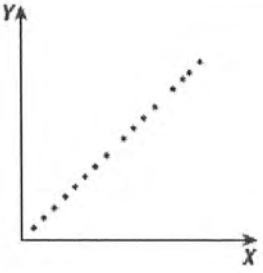
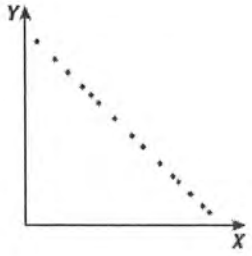
- L'une des variables est associée à l'axe des abscisses et l'autre variable à l'axe des ordonnées.
- Chacun des couples de la distribution est représenté par un point.
- La corrélation est dite linéaire lorsque les points tendent à former une droite oblique.

Exemple : À partir de la distribution précédente, trace le nuage de points dans le plan cartésien ci-dessous.

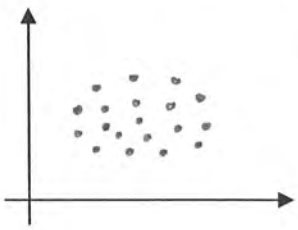


Comment qualifierais-tu la corrélation linéaire que tu viens de représenter ? Pour t'aider à répondre à cette question, base-toi sur le tableau ci-dessous :

Rép: positive et forte

Sens / Intensité	Positif (les valeurs des deux variables varient dans le même sens)	Négatif (les valeurs des deux variables varient dans le sens contraire)
Faible		
Forte		
Parfaite		

Une corrélation peut être nulle également. En voici un exemple :



Exercices : ai-je-bien compris p.69 et p.71 #1

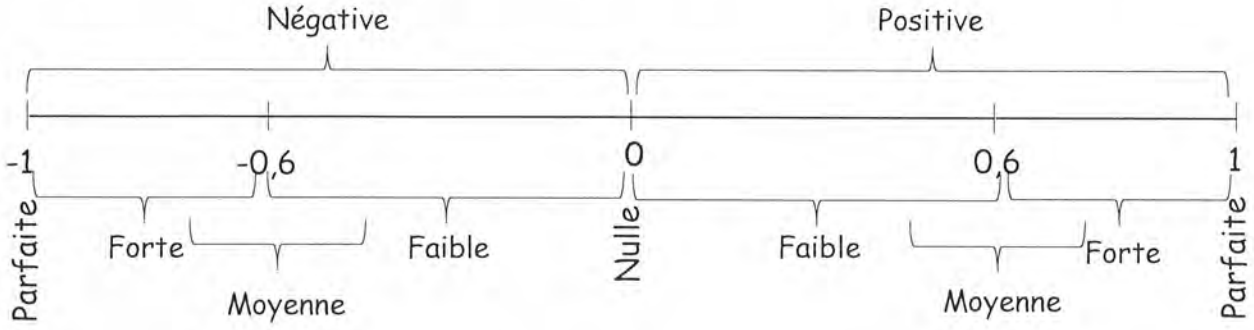
p. 76 # 4, 5, 9, 10

Cours 5

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 coefficient de corrélation et on le désigne par la lettre *r*.

Le schéma suivant associe les valeurs attribuées au coefficient de corrélation linéaire et des nuages de points obtenus à partir de distributions à deux variables.



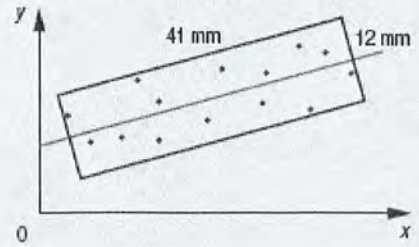
Il existe plusieurs méthodes pour **approximer le coefficient de corrélation** linéaire d'une distribution à deux variables. L'une d'elles est une méthode d'**estimation graphique** faisant intervenir **un rectangle dans un nuage de points.**

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 :

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

Ex.: Distribution à deux variables

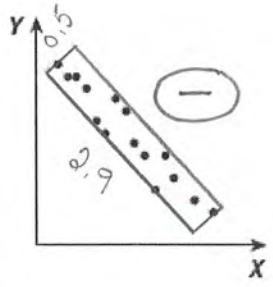


$r \approx 1 - \frac{12}{41} \approx 0{,}71$

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

Exercices : Pour chacun des nuages de points suivants, estime le coefficient de corrélation à l'aide de la méthode du rectangle et qualifie la corrélation.

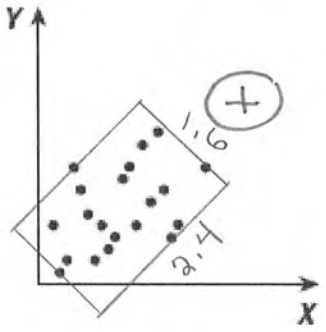
a)



$r \approx -\left(1 - \frac{0,5}{2,9}\right)$

$r \approx -0,83$ Donc négative et forte

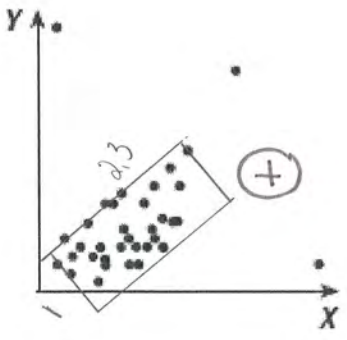
b)



$r \approx +\left(1 - \frac{1,6}{2,4}\right)$

$r \approx +0,\overline{3}$ Donc positive et faible

c)



$r \approx +\left(1 - \frac{1}{2,3}\right)$

$r \approx +0,57$ Donc positive et faible

2. Voici 4 coefficients de corrélation linéaire : -0,82 ; -0,43 ; 0,22 et 0,63. Classe en ordre croissant ces 4 coefficients selon l'intensité de leur corrélation linéaire.

0,22 ; -0,43 ; 0,63 ; -0,82

3. Vrai ou faux : Une distribution ayant un coefficient de corrélation linéaire de -0,85 est dite plus faible qu'une distribution dont le coefficient de corrélation linéaire est de 0,75. Justifie ta réponse

Faux plus un coefficient s'approche de -1 ou de 1 plus il est fort

4. Explique la différence entre deux distributions statistiques dont une a un coefficient de corrélation linéaire de -0,9 et l'autre de 0,9.

Celui de -0,9 a des variables statistiques qui varient dans le sens contraire alors que celui de 0,9 les variables varient dans le même sens.

Interprétation d'un coefficient de corrélation :

Plusieurs facteurs peuvent intervenir dans l'interprétation de la corrélation entre deux variables. C'est pourquoi il faut être vigilant lorsque vient le temps de faire des prédictions et de tirer des conclusions.

Interprétation	Exemple
• Le lien entre deux variables peut être un rapport de cause à effet, c'est-à-dire que l'une des variables agit directement sur l'autre. Dans de telles situations, la corrélation est parfaite et la relation entre les deux variables se définit par une règle.	La corrélation entre l'altitude et la température est parfaite puisque la température varie directement en fonction de l'altitude.
• La corrélation entre deux variables peut être importante sans que les deux variables soient directement liées entre elles. Elles peuvent dépendre toutes deux d'une troisième variable qui, en variant, engendre des variations pour les deux premières.	En été, il peut sembler y avoir une forte corrélation entre le nombre de cornets de crème glacée vendus et le nombre de climatiseurs vendus dans une ville, alors qu'en fait, ces deux variables dépendent plutôt d'une troisième, qui est la température.

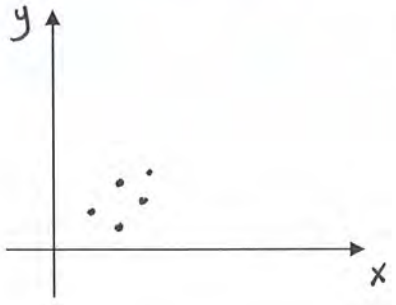
Interprétation	Exemple
• Considérer une corrélation comme étant linéaire alors qu'un autre modèle serait plus approprié.	La croissance de la population d'une métropole peut être étudiée selon une corrélation linéaire. Toutefois, l'utilisation d'un modèle exponentiel serait plus appropriée.
• Il arrive parfois qu'il y ait une corrélation entre deux variables seulement sur un intervalle donné.	Sur l'intervalle [5, 10] ans, la corrélation entre l'âge et la taille d'une personne est linéaire. Toutefois, avant et après cet intervalle, le modèle linéaire n'est pas le mieux adapté.
• Une distribution à deux variables peut comporter des données aberrantes, en raison notamment d'erreurs de manipulation ou de mesure.	Le degré de précision de l'instrument utilisé lors de la collecte des données laisse à désirer.

Une forte corrélation indique l'existence d'un lien statistique. Cependant, elle n'explique pas la raison et la nature du lien. Par la suite, on essaie de caractériser qualitativement et quantitativement ce lien et d'établir des prédictions tout en étant conscient des limitations de ces prédictions. Il faut exercer son jugement !!!

Les sources de biais :

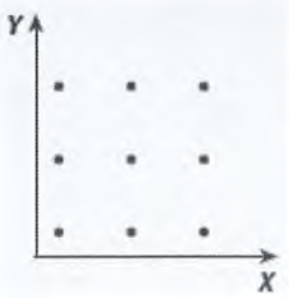
Certaines sources de biais peuvent mener à de mauvaises interprétations de fausses prédictions ou à des conclusions erronées. Il faut donc tenter de les éviter. En voici quelques-unes.

1)Échantillon non représentatif



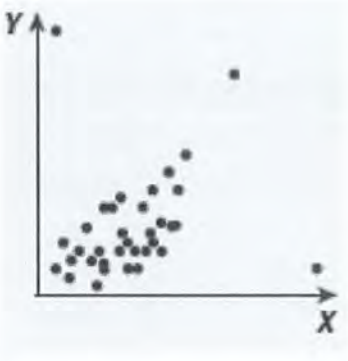
Un échantillon de trop petite taille peut laisser croire que la corrélation est plus forte qu'elle ne l'est en réalité.

2) Les corrélations non linéaires



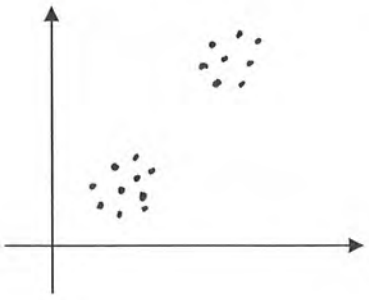
Avant de conclure qu'il n'existe pas de corrélation entre les variables, il faut observer le nuage de points. Le modèle linéaire n'est pas approprié s'il existe une corrélation non linéaire.

3) La présence de données aberrantes (éloignées)



La présence de données éloignées peut fausser l'interprétation des résultats. En présence de ces données, vérifier ce qu'elles représentent dans le contexte et s'il s'agit d'anomalies, les exclure de l'analyse.

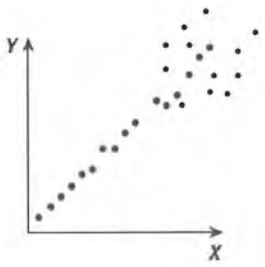
4) La présence de deux groupes distincts



La présence de 2 groupes de données peut laisser croire que la corrélation est plus forte.

Séparément, la corrélation est presque nulle.

5) La présence d'une corrélation par intervalles



Parfois on peut observer des corrélations différentes dans des intervalles différents. Il est alors préférable d'en faire la distinction. La corrélation est forte pour les petites valeurs de x alors qu'elle est faible pour les grandes valeurs de x.

Exercices : Volume : ai-je bien compris p.81 p.86 #1, #2 et #3
p.101 #5, 7, p.102 #8

Cours 6

Période d'exercices

Cours 7

Droite de régression

Dans un nuage de points mettant en relation deux variables statistiques, la droite qui représente le mieux l'ensemble des points est appelée la **droite de régression.**
Il existe différentes méthodes pour déterminer l'équation d'une droite de régression.

Méthode de la droite de Mayer :

Voici comment déterminer l'équation d'une droite de régression à l'aide de la 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, si possible égaux.
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.

Exercices : Détermine la droite de régression pour chacune des distributions suivantes :

1.

x	y
6	23
7	26
10	39
13	44
14	48
15	55
18	50
19	65
23	68
25	72

1° Trouvons P_1 :

$x_1 = \frac{6+7+10+13+14}{5}$ $\quad y_1 = \frac{23+26+39+44+48}{5}$

$x_1 = 10$ $\quad y_1 = 36$ $\quad P_1(10, 36)$

2° Trouvons P_2 :

$x_2 = \frac{15+18+19+23+25}{5}$ $\quad y_2 = \frac{55+50+65+68+72}{5}$

$x_2 = 20$ $\quad y_2 = 62$ $\quad P_2(20, 62)$

3° Trouvons a:

$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{62-36}{20-10}$

$a = 2.6$

4° Trouvons b: Avec P_1 ou P_2

$y = 2.6x + b$

$62 = 2.6 \cdot 20 + b$

$62 - 52 = b$

$10 = b$

Donc $y_r = 2.6x + 10$

2.

x	2	4	5	6	6	7	8	9	11	14	18
y	5	7	7	6	9	10	14	13	14	15	16

1° Trouvons P_1 :

$$x_1 = \frac{2+4+5+6+6+7}{6} = 5$$

$$y_1 = \frac{5+7+7+6+9+10}{6} = 7,\overline{3}$$

2° Trouvons P_2 :

$$x_1 = \frac{7+8+9+11+14+18}{6} = 11,1\overline{6}$$

$$y_1 = \frac{10+14+13+14+15+16}{6} = 13,\overline{6}$$

3° Trouvons a :

$$a = \frac{13,\overline{6} - 7,\overline{3}}{11,1\overline{6} - 5} = 1,03$$

4° Trouvons b :

$y = 1,03x + b$

$7,\overline{3} = 1,03 \cdot 5 + b$

$7,\overline{3} - 5,15 = b$

$2,18 = b$

Rép : $y_r = 1,03x + 2,18$

Notez bien : 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 peut être fiable.

Exercice :

#1 Dans l'exercice précédant, au numéro 1, à l'aide de la droite de régression que tu as trouvé, estime : $y = 2,6x + 10$

a) la valeur de y lorsque x sera égale à 32

$y = 2,6 \cdot 32 + 10$

$y = 93,2$

} extrapolation

b) la valeur de x lorsque y sera égale à 52

$52 - 10 = 2,6x + 10 - 10$

$\frac{42}{2,6} = \frac{2,6x}{2,6}$

$16,15 = x$

} interpolation

Exercices : Volume : ai-je bien compris p. 81 et p.91
p. 86 #3
p. 97 # 1, 2 et 3

Mini-test #3 au prochain cours

Cours 8 et 9 :

Document d'exercices préparatoires

Voici des numéros supplémentaires dans le volume :

p. 100 # 2a), #4, #10 a) c), #16 et # 25

Cours 10 : Examen