Les statistiques

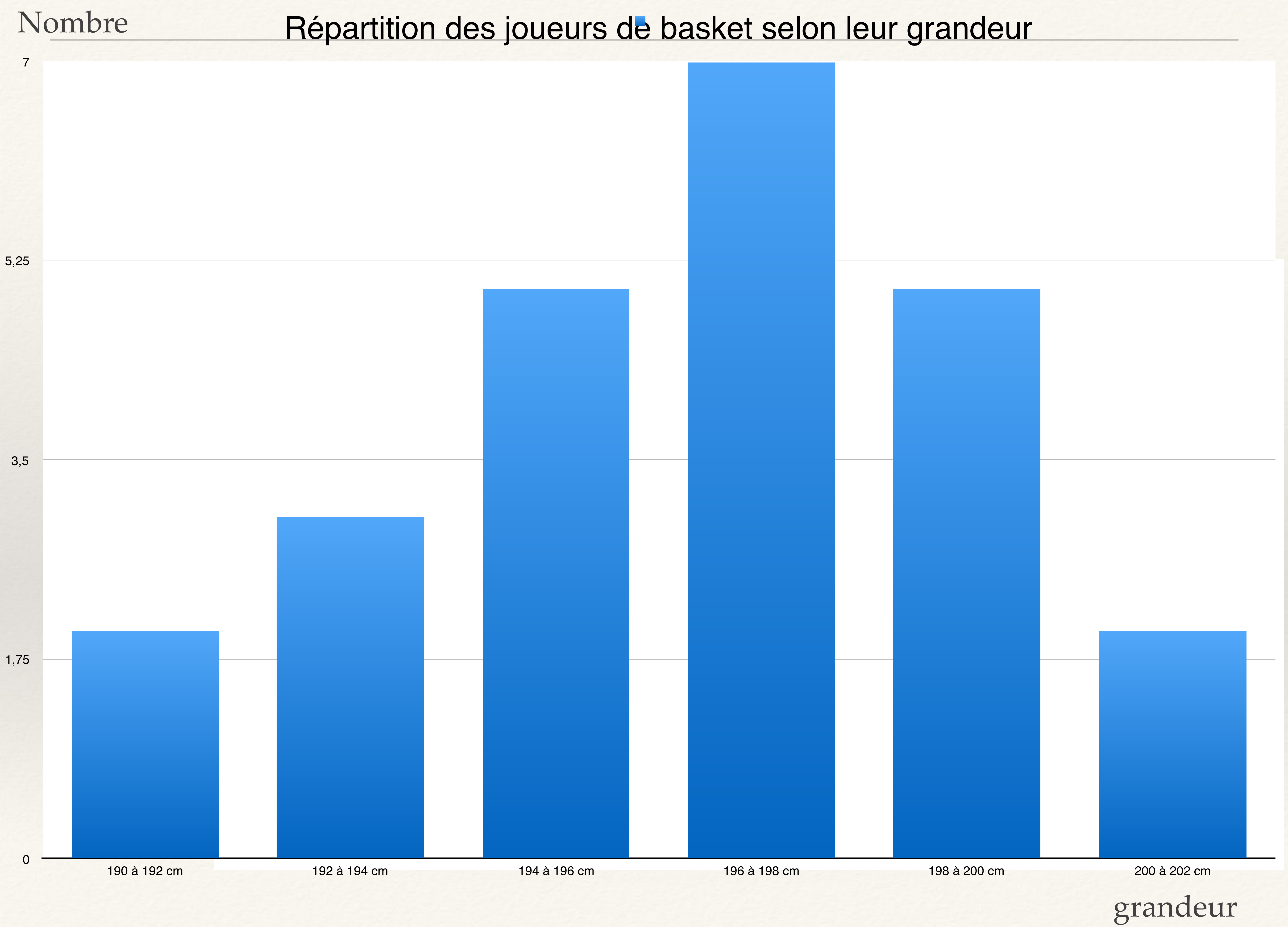
Chapitre 8

À quoi ça sert???

- Parce que nous avons besoin de renseignements pour surveiller les changements importants et les répercussions que ces changements ont sur chacun d'entre nous.
- Les statistiques fiables nous aident à faire face aux nombreux défis que notre monde doit relever. Nos gouvernements, nos entreprises et nos collectivités ont absolument besoin des renseignements statistiques pour prendre des décisions éclairées.

On les utilise…

- Pour comprendre la situation de votre pays, évaluer le rendement des gouvernements
- Syndicats utilisent les statistiques sur la moyenne des salaires versés
- Environnement: gaz à effet de serre
- Enseignants suivent les statistiques pour dégager les nouvelles tendances
- Vérifier les rendements à la bourse…


Nombre
Répartition des joueurs de basket selon leur grandeur
7
5,25
3,5
1,75
0
190 à 192 cm
192 à 194 cm
194 à 196 cm
196 à 198 cm
198 à 200 cm
200 à 202 cm
grandeur

L'__histogramme____ permet de ___représenter____ des données groupées en ___classes__.

TABLEAUX

Le ___tableau___ de données condensées est ___utilisé___ pour représenter une distribution qui comporte un grand nombre de données ayant tendance à de ___répéter__.

L'___effectif___ de chaque modalité ou valeur correspond au nombre de fois qu'elle apparaît dans la _distribution__.

Tableaux de données condensées

Nombre d'employés	Effectif
3	4
4	7
5	5
6	3
7	4
8	2
TOTAL	25

Le __tableau__ de données groupées en __classes__ est utilisé pour représenter une distribution qui comporte un grand __nombre__ de données ayant tendance à ne pas se répéter.

Toutes les __classes__ du tableau ont la même __amplitude__.

L'effectif de chaque classe correspond au __nombre__ de __données__ que cette classe comprend.

Participants à un sondage

Âge	Effectif
[15,20[	17
[20,25[	25
[25,30[	34
[30,35[	45
[35,40[	41
[40,45[	32
TOTAL	194

Mesures de tendance centrale

MESURE	DÉFINITION	TYPES DE DISTRIBUTIONS	
		DONNÉES CONDENSÉES	DONNÉES GROUPÉES EN CLASSE
MODE	Valeur de la variable la plus fréquente	Modalité dont l'effectif est le plus élevé	Milieu de la classe modale dont l'effectif est le plus élevé
MÉDIANE	Le centre de position d'une distribution	Donnée du centre si impair, moyenne des deux si pair	Milieu de la classe médiane
MOYENNE	Centre d'équilibre d'une distribution	somme des produits divisée par effectif	somme des produits par leurs effectifs

EXEMPLE

Dans un processus d'embauche pour un poste dans la fonction publique fédérale, on veut garder ceux qui ont obtenu les meilleurs résultats, en %, parmi les suivants:

62 64 65 65 66 68 69 72 75 78 **79 79 81 82 82 84 85 86 87 89**

Dans le but de diminuer significativement la taille du groupe, on décide d'éliminer la moitié des candidats. Pour établir la note minimale de sélection, on utilise alors la médiane qui sera $78,5$ (donnée entre 78 et 79).

Mode: donnée la plus populaire: 65, 79 et 82

Moyenne: 62+64+65+65+66+68+69+72+75+78+79+79+81+82+82+84+85+86+87+89 = 1518 divisé par 20 = 75,9

MESURES DE DISPERSION

L'__étendue___ est l'__écart___ de l'échantillon Il suffit de déterminer la valeur maximale_ et la valeur __minimale___.

$$E = x_{\max} - x_{\min}$$

DIAGRAMME DE QUARTILES

Une mesure de _quartile_ permet d'établir la position d'une donnée par rapport aux autres _________ de la série.

Les _quartiles_ (Q_1, Q_2, Q_3) sont des mesures de _répartition__ qui divisent une série de «n» (nombre de données) données ordonnée en quatre ensembles renfermant le même nombre de donnée.

Q2: ___médiane________

Q1: premier quartile____

Q3: _troisième quartile__

Le _diagramme_ de __quartiles__ permet d'analyser la dispersion de l'_étendue__ des données d'une distribution.

Exemple: Déterminer les quartiles de cette distribution: 60,32,87,98,56,75,35,68,86,90,75,59,61,84,64,48

- 1. Placer en ordre croissant: 32,35,48,56,59,60,61,64,68,75,75,84,86,87,90,98
- 2. Déterminer la médiane: Cette distribution est constituée de 16 données. Par conséquent, la médiane sera la moyenne entre la 8e et la 9e donnée.
- (64+68)/2=66
- 3. Déterminer la valeur du premier quartile: La première moitié est composée des données suivantes : 32,35,48,56,59,60,61,64, la médiane est donc: 57,5
- 4. Déterminer la valeur du troisième quartile: La deuxième moitié est composée des données suivantes :68,75,75,84,86,87,90,98, la médiane est donc 85.
- 5. Tracer le diagramme:

8.1 DIAGRAMME À TIGE ET À FEUILLES

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.**

Exemple

Femmes		Hommes
6-5-2-2	0	2-3-4-5
4-3-3-1	1	0-0-1-6-8
9-9-8-4-3	2	2-2-5-7-8
9-8-7-6-5	3	1-8-8-8-9

La colonne du centre représente la tige et la donnée à droite représente les données des femmes (unités).

Femmes: 2 - 2 - 5 - 6 - 11 - 13 - 13 - 14 - 23 - 24 - 28 - 29 - 29 - 35 - 36 - 37 -38 - 39

Hommes: 2 -3 -4 - 5 - 10 - 10 - 11 - 16 - 18 - 22 - 22 - 25 - 27 -28 - 31 - 38 - 38 - 38 - 38 - 39

DIAGRAMME À TIGE ET À FEUILLES

Un _diagramme__ à tige et à __feuilles____, permet de mieux visualiser la répartition des _données__. Les nombres situés à _gauche___ de la ligne verticale s'appellent _tiges__ et correspondent aux chiffres des _dizaines_, tandis que les __données____ situés à _droite_ de la ligne verticale s'appellent feuilles et correspondent aux chiffres des ___unités__ dans les _données___.

Chaque __rangée___ est associée à une classe et chaque donnée est ___séparée____ en deux parties se trouvant sur une même ligne.

8.2 ÉCART MOYEN ET RANG CENTILE

L'écart _moyen_ est une mesure de _dispersion_ qui indique de combien en moyenne les _données_ d'une _série_ s'écartent de la moyenne des données de la série.

$$\textit{Écartmoyen} = \frac{somme \quad des \quad écarts \quad à \quad la \quad moyenne}{moyenne}$$

1. CALCULER LA MOYENNE
2. CALCULER LES ÉCARTS ENTRE LA MOYENNE ET CHAQUE DONNÉE (VALEUR ABSOLUE)

3. ADDITIONNER TOUS LES ÉCARTS

4. DIVISER CETTE SOMME PAR LE NOMBRE DE DONNÉES

L'ÉCART TYPE

La _ VARIANCE _ ou l'écart type est une mesure de DISPERSION qui indique la moyenne des DONNÉES des écarts à la _ MOYENNE

1. _ CALCULER LA MOYENNE _____
2. _ CALCULER LES ÉCARTS ENTRE LA MOYENNE ET CHAQUE DONNÉE (VALEUR ABSOLUE) ET METTRE AU CARRÉ
3. _ ADDITIONNER TOUS LES ÉCARTS
4. DIVISER CETTE SOMME PAR LE NOMBRE DE DONNÉES ET EXTRAIRE LA RACINE

Ex 1: Trouve l'écart moyen et l'écart type dans la liste suivante

4 – 8 – 12 – 13 – 18

Donnée	Moyenne de la distribution	Écart à la moyenne	Écart à la moyenne au carré
4	11	7	49
8	11	3	9
12	11	1	1
13	11	2	4
18	11	7	49
Total = 55	55	20	112

$$\text{Moyenne} = \frac{4 + 8 + 12 + 13 + 18}{5} = 11$$

Ex 1: Suite

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

$$\text{Écart moyen} = \frac{20}{5}$$

$$\text{Écart moyen} = 4$$

$$\text{Écart type} = \sqrt{\frac{\text{Somme des carrés des écarts à la moyenne}}{\text{Nombre total de données}}}$$

$$\text{Écart type} = \sqrt{\frac{112}{5}}$$

$$\text{Écart type} = \sqrt{22{,}4}$$

$$\text{Écart type} = 4{,}73$$

Effectue les opérations suivantes et regarde les réponses par la suite.

Trouve l'écart moyen et l'écart type.

Résultat	1	2	3	4	5
Effectif	2	1	5	12	10

Réponses

a) Écart moyen = 0,813
Écart type = 1,106

Résultat	2	4	6	8	10
Effectif	2	1	5	12	10

b) Écart moyen = 1,627
Écart type = 2,212

Résultat	0	1	2	3	4
Effectif	10	8	6	4	2

c) Écart moyen = 1,067
Écart type = 1,247

Résultat	0	2	4	6	8
Effectif	10	10	4	3	3

d) Écart moyen = 2,133
Écart type = 2,59

CENTILES ET RANG CENTILE

Les ____centiles____ ou _rang centile_ partagent la série de données en _100_ tranches renfermant chacune 1% des __données__. De gauche à droite, on _attribue___ aux données situées dans la 1ère tranche, le 1er rang _centile___, aux données situées dans la 2e tranche, le 2e rang centile...

La formule suivante permet de déterminer le rang centile d'une donnée dans une série:

$$Rang \quad centile \quad de \quad x = \frac{Nombre \quad \inf. \quad de \quad données \quad à \quad x \quad + \frac{1}{2} \quad Nombre \quad de \quad données \quad égales \quad à \quad x}{Nombre \quad total \quad de \quad données} \times 100$$

On arrondit le rang centile obtenu à l'entier supérieur le plus proche.

Exemple:

- Calculer le rang centile de 45 dans cette distribution:
- 3, 10, 12, 34, 37, 42, 45, 45, 45, 48, 51, 65, 79, 80, 90

$$Rang \quad centile \quad de \quad x = \frac{Nombre \quad \inf. \quad de \quad données \quad à \quad x \quad + \frac{1}{2} \quad Nombre \quad de \quad données \quad égales \quad à \quad x}{Nombre \quad total \quad de \quad données} \times 100$$

Il y a 6 données inférieures à 45

Il y a 3 données égales à 45

$$Rangcentile \quad 45 \quad = \frac{6+\frac{1}{2}(3)}{15} \times 100$$

$$Rang \quad centile \quad 45 \quad = \frac{7,5}{15} \times 100 = 50$$

Le rang centile de 45 est 50.

Exemple: Voici les résultats sur 100, de 150 personnes à un test de maths: $\underbrace{45, 54, 56, 57,...69}_{60 données}$, 70, 70, 71, 71, 73, 74, 74, 76, 76, 78, $\underbrace{79,..., 84, 84, 86, 89, 92, 97}_{67 données}$.

Détermine le rang centile de 74:

$$= \frac{70 + 1/2(2)}{150} \times 100 = 47,333 = 48$$

Détermine le rang centile de 86:

Si on connait le rang centile et on cherche la donnée:

$$\frac{Rangcentile}{100} \times Nombre \quad de \quad données$$

Détermine la donnée dont le rang centile est 3:

$$\frac{3}{100} \times 150 = 4,5e \quad donnée$$

Arrondir à la donnée inférieure

La 4e donnée est 57

On utilise le __ rang centile __ **lorsqu'une série contient un très grand** __ nombre __ **de données.**

8.3 CORRÉLATION, TABLEAU DE DISTRIBUTION À DOUBLE ENTRÉE ET NUAGE DE POINTS

Toute étude _statistique_ faisant intervenir deux variables X et Y permet d'__associer__ à chaque individu d'un __échantillon______ le couple (x, y).

On appelle _distribution_ à deux _variables_ l'ensemble des couples obtenus.

Un _tableau___ à double entrée est un tableau donnant la répartition des individus d'un échantillon selon les modalités conjointes de deux _variables__.

Un tableau qui met en __relation___ deux variables permet de faire des _comparaisons_ et d'établir, lorsqu'il y a lieu, un lien entre les _deux_ _variables___.

NUAGE DE POINTS

Lorsque que les variables x et y sont ___reliées___, on peut représenter chaque ___donnée___ par un point dans le plan ___cartésien___. L'ensemble des ___points___ correspondant aux données s'appelle un ___nuage___ de ___points___. On appelle ___donnée___ ___aberrante___, toute donnée qui s'éloigne passablement des autres ___données___ du nuage. La représentation du nuage permet de ___vérifier___ s'il existe un lien entre les deux ___variables___. Lorsque les variables sont quantitatives, un tel lien est appelé ___corrélation___.

CORRÉLATION LINÉAIRE

On dit qu'il y a _corrélation_ entre des variables quantitatives x et y pour indiquer une _dépendance_ statistique entre les deux variables.

La correlation est dite __linéaire__ lorsque les _points_ du nuage ont tendance à se rapprocher d'une __droite_.

La correlation est dite __non linéaire_____ lorsque les points ont tendance à se rapprocher d'une courbe.

La correlation est dite __positive____ lorsque les variables varient dans le même sens.

La correlation est dite __négative____ lorsque les variables varient en _sens__ __contraire_.

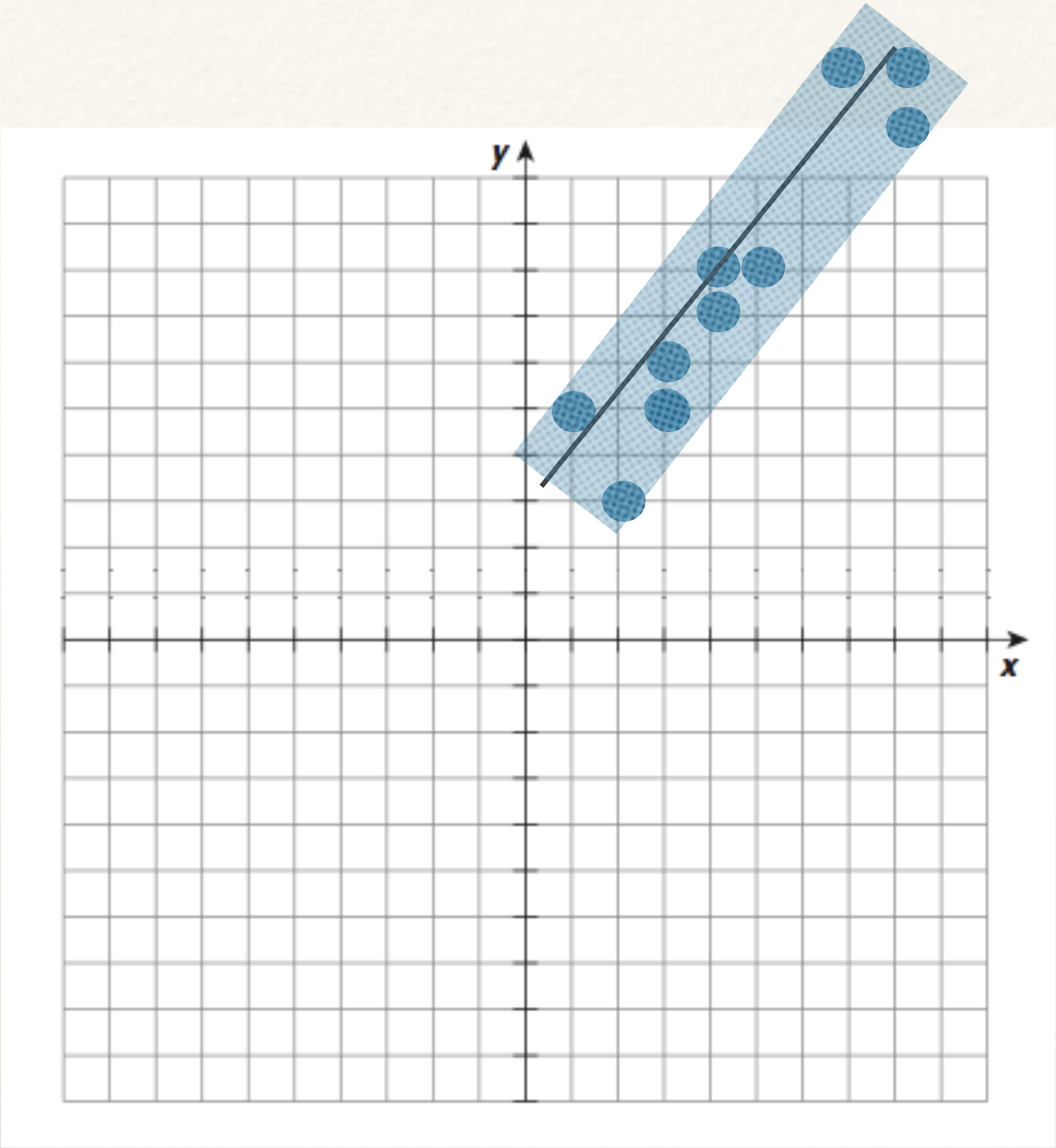
8.4 INTERPRÉTATION QUANTITATIVE DE LA CORRÉLATION ET COEFFICIENT DE CORRÉLATION LINÉAIRE

COEFFICIENT DE CORRÉLATION LINÉAIRE		SIGNIFICATION
NÉGATIF	POSITIF	
PRÈS DE 0	PRÈS DE 0	Corrélation nulle
PRÈS DE -0,5	PRÈS DE 0,5	Corrélation faible
PRÈS DE -0,75	PRÈS DE 0,75	Corrélation moyenne
PRÈS DE -0,87	PRÈS DE 0,87	Corrélation forte
ÉGAL À -1	ÉGAL À 1	Corrélation parfaite

ESTIMATION DU COEFFICIENT LINÉAIRE GRAPHIQUEMENT

1. Tracer une droite qui représente le plus possible les points
2. Construire un rectangle (+< possible) excluant données aberrantes
3. Calculer d(largeur) et D(longueur) du rectangle et appliquer la formule:

$$r \sim \pm\left(1 - \frac{d}{D}\right)$$



d = 1,1

D = 5,3

$$r \sim \pm\left(1 - \frac{1,1}{5,3}\right)$$

$$r \sim 0,8$$

Corrélation forte et positive

ESTIME LE COEFFICIENT DE CORRÉLATION LINÉAIRE:

(3,5) (4,7) (7,12) (2,4) (1,5) (8,12) (3,6) (4,8) (5,8) (8,11)

8.5 DROITE DE RÉGRESSION OU LA DROITE DE MAYER

Pour déterminer l'équation de la droite de Mayer lorsqu'on se retrouve avec un nuage de points, on veut en fait trouver une ___ équation ___ **qui serait le plus possible** _ représentative __ **du résultat de l'étude statistique. Voilà les étapes:**

1. Placer les données en ordre croissant (selon les valeurs en x)
2. Séparer les données en deux groupes égaux (pour les groupes impairs,utiliser la donnée du centre dans les deux groupes)
3. Faire la moyenne du premier groupe (moyenne en x et moyenne en y), ce sera le premier point utilisé pour trouver l'équation.
4. Faire la moyenne du deuxième groupe (moyenne en x et moyenne en y), ce sera le deuxième point utilisé pour trouver l'équation.

Utiliser ces deux points pour trouver l'équation de la droite.

Exemple: Voici une étude statistique effectuée quant au nombre d'inscriptions à différentes activités sportives en fonction du nombre d'enfants.

NB D'ENFANTS	2	6	5	4	1	5	3
NB D'INSCRIPTIONS	6	9	7	6	5	8	4

Placer les données en ordre croissant (selon les valeurs en x)

NB D'ENFANTS	1	2	3	4	5	5	6
NB D'INSCRIPTIONS	5	6	4	6	7	8	9

Séparer les données en deux groupes égaux (pour les groupes impairs,utiliser la donnée du centre dans les deux groupes)

Faire la moyenne du premier groupe (moyenne en x et moyenne en y), ce sera le premier point utilisé pour trouver l'équation.

Faire la moyenne du deuxième groupe (moyenne en x et moyenne en y), ce sera le deuxième point utilisé pour trouver l'équation.

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

$$y_1 = \frac{6+7+8+9}{4} = 7{,}5$$

Trouver l'équation (2,5 ; 8,75) et (5, 7,5)