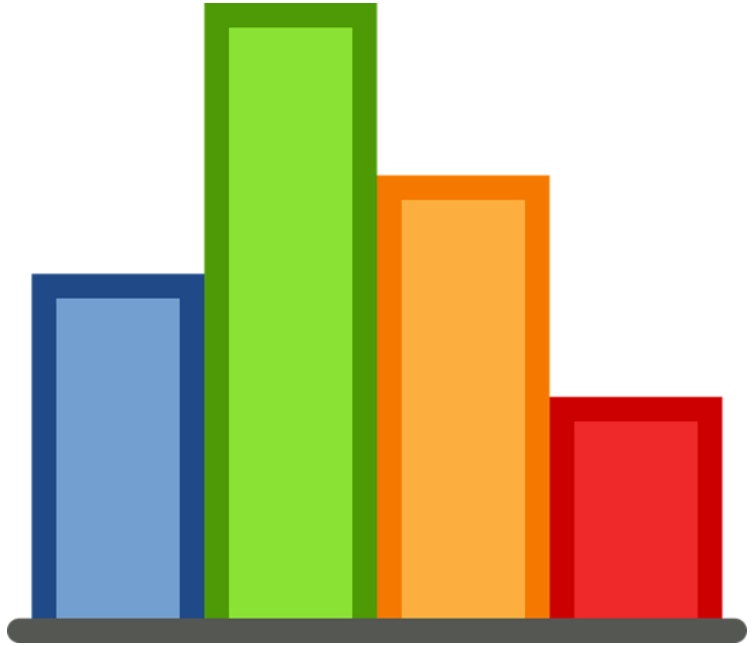


Statistique descriptive

Gérard Barmarin

Cours 5

VILLE DE WAVRE
IFOSUP
INSTITUT DE FORMATION SUPERIEURE

1 Définitions
2 Types de variables
3 Graphes
4 Indicateurs statistiques
5 Note Z, indicateur de symétrie & boîte à moustaches

Aujourd'hui:

6 Statistique à 2 variables, droite d'ajustement

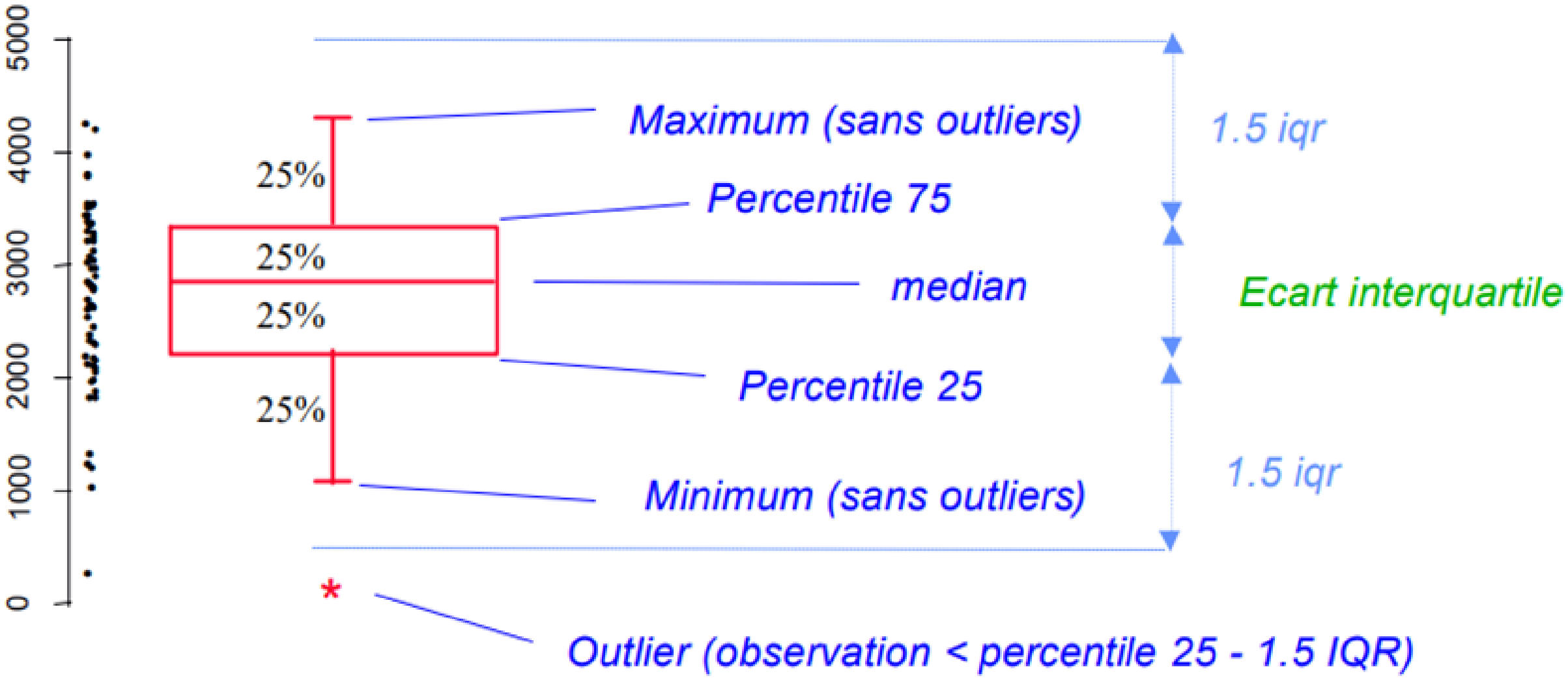
Mais avant…

Corrections exercices!

Exercices pour la semaine prochaine:

- Faire les boîtes à moustaches des exercices 3.6 et 3.7 et de l'exercice de la semaine passée basé sur la distribution de l'âge des employés d'une entreprise

Rappel:


5000
4000
3000
2000
1000
0
25%
25%
25%
25%
Maximum (sans outliers)
Percentile 75
median
Percentile 25
Minimum (sans outliers)
1.5 iqr
Ecart interquartile
1.5 iqr
*
Outlier (observation < percentile 25 - 1.5 IQR)

Le coefficient de Yule

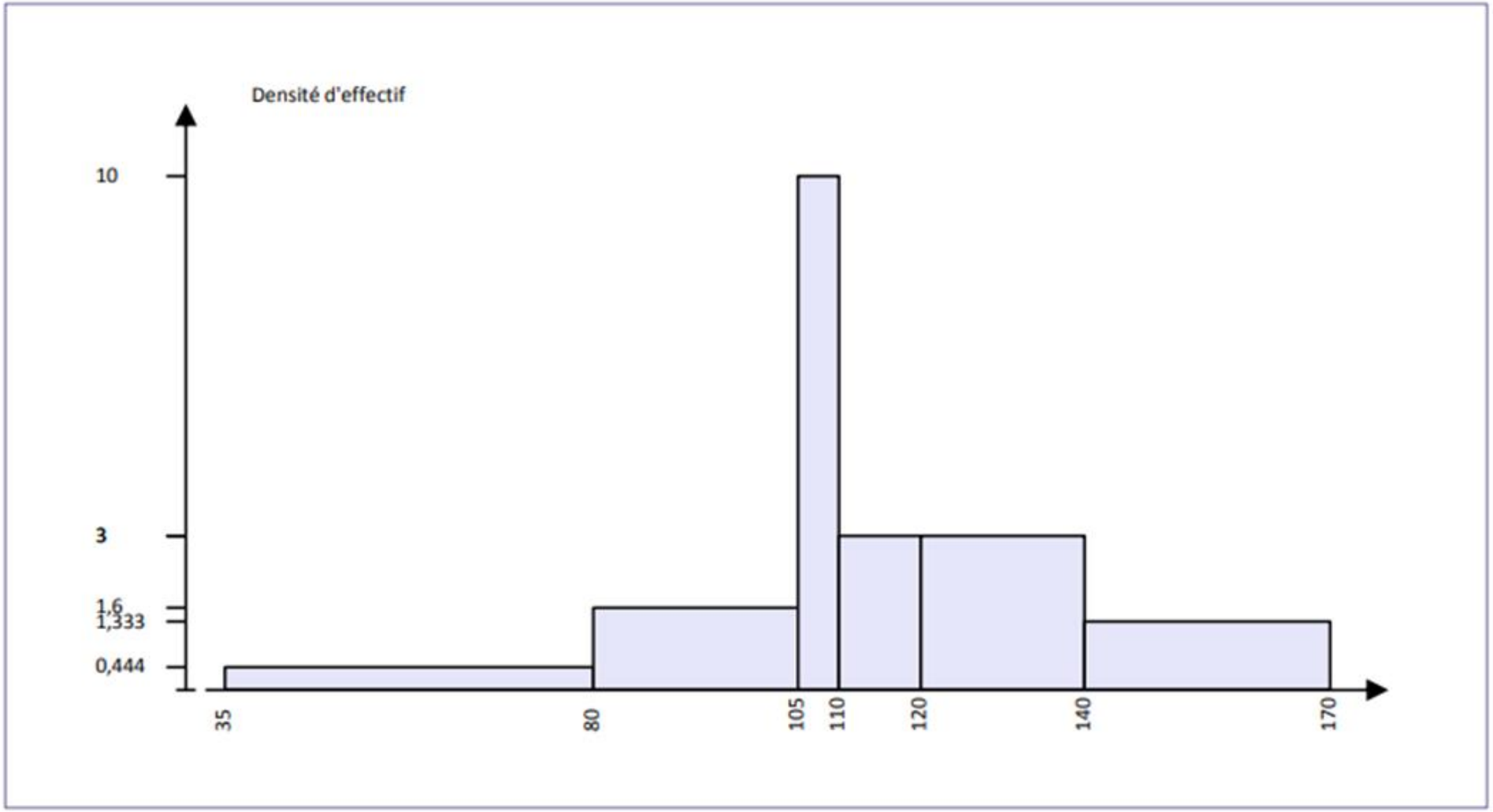
On peut donc résumer :

- **Si Sy = 0 alors la distribution est symétrique**
- **Si Sy < 0 alors la distribution est étalée à gauche (oblique à droite)**
- **Si Sy >0 alors la distribution est étalée à droite (oblique à gauche)**

$$S_Y = \frac{(Q3 - Med) - (Med - Q1)}{(Q3 - Med) + (Med - Q1)}$$

Ex 3.6

Voici l'histogramme représentant la distribution sur leur superficie (en ares) de 240 terrains agricoles vendus au premier trimestre 2013 à Cracovie.



Exercice 3.6

Q_1 = 105 ares

Q_2 = 113,33 ares

Q_3 = 133,33 ares

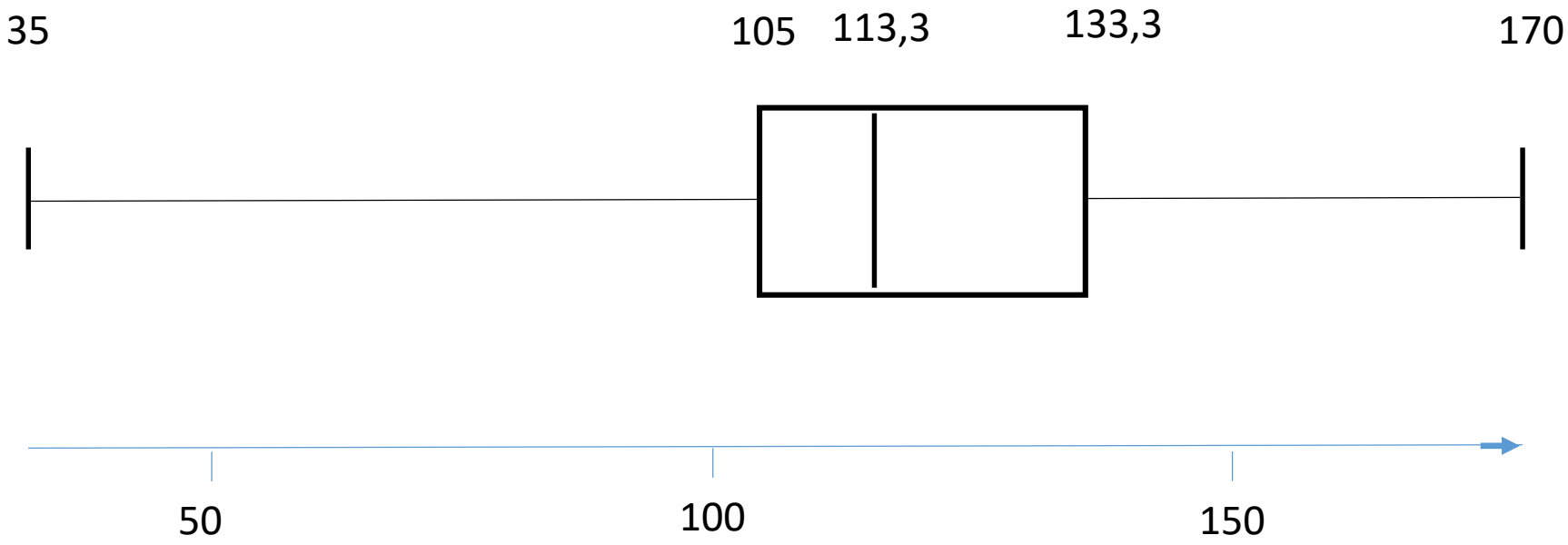
$Q_3 - Q_2$ = 20 ares

$Q_2 - Q_1$ = 8,33 ares

$Q_3 - Q_1$ = IQ = 28,33 ares

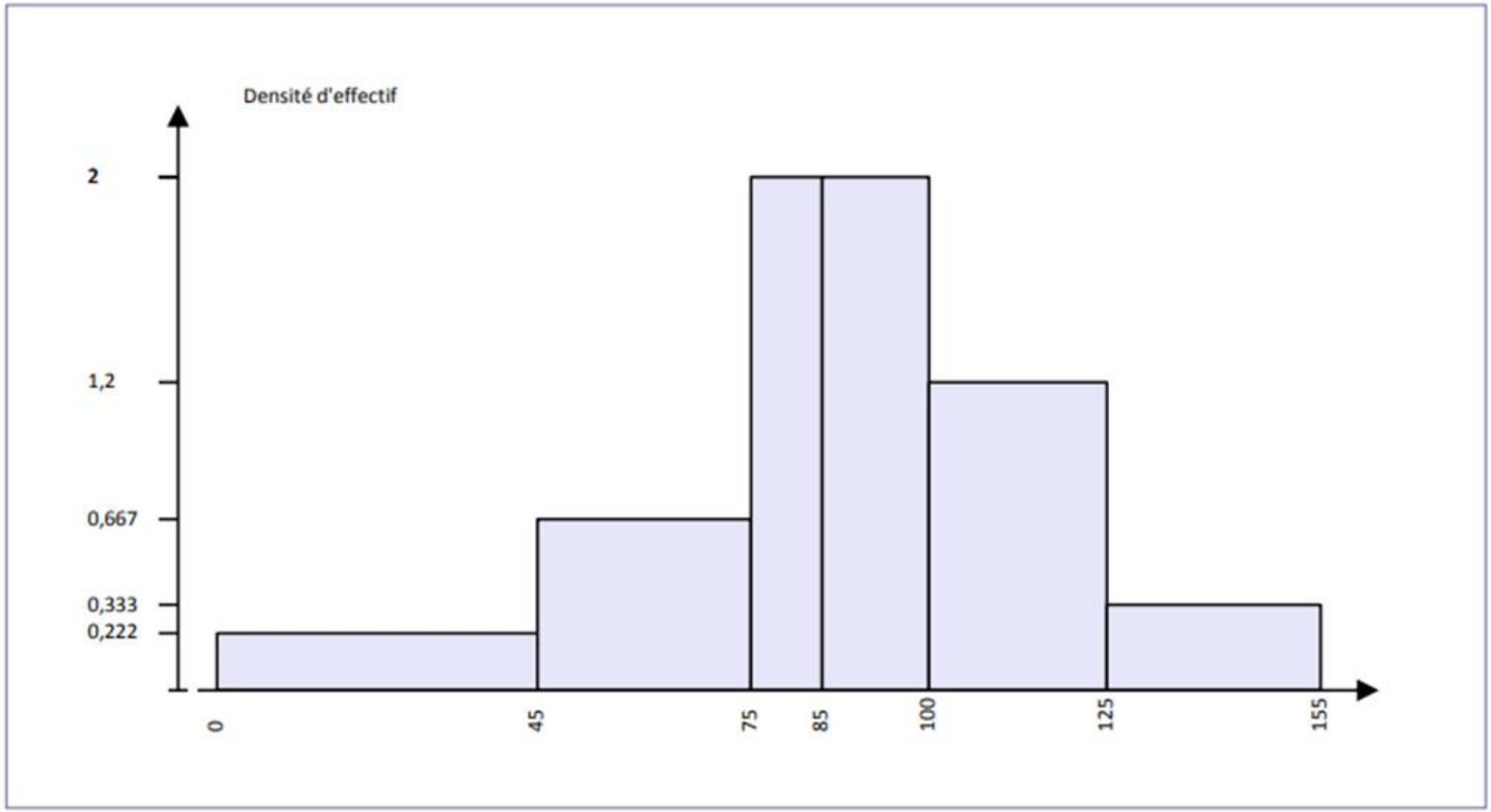
S_y = 0,41

1,5 x IQ = 42,5 ares



Ex 3.7

Voici l'histogramme représentant la distribution sur leur superficie (en ares) de 120 terrains agricoles vendus au premier trimestre 2014 à Cracovie.



Exercice 3.7

Q_1 = 75 ares

Q_2 = 90 ares

Q_3 = 108,33 ares

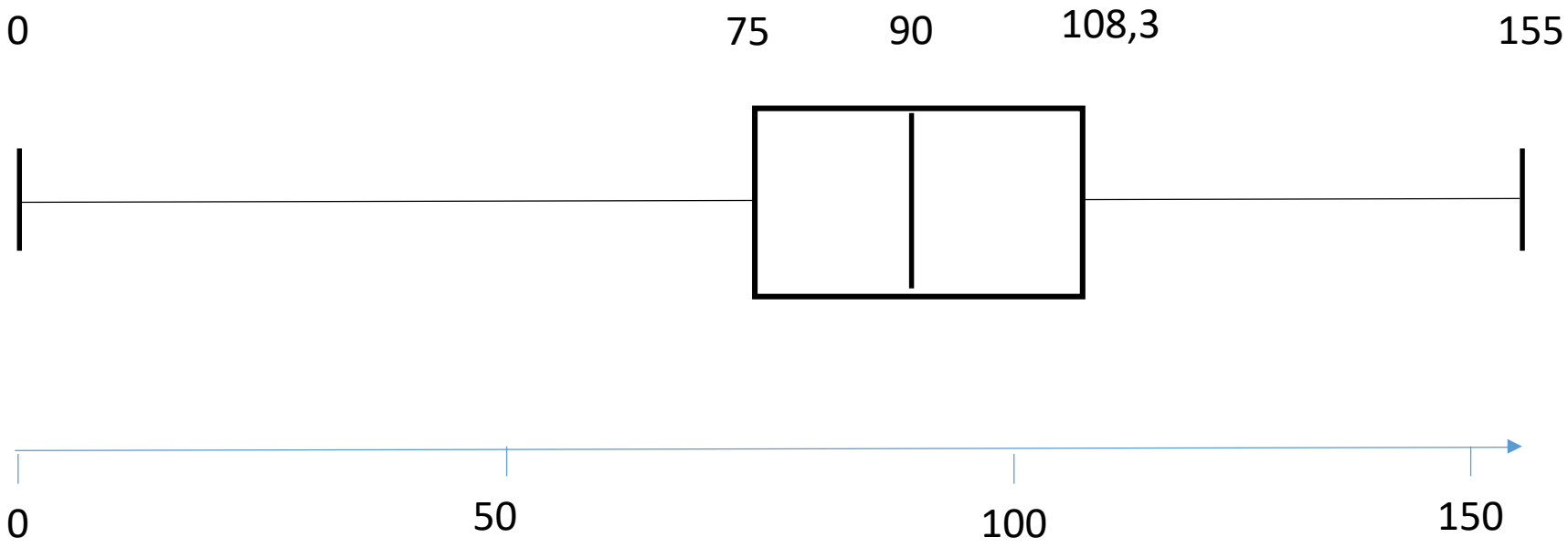
$Q_3 - Q_2$ = 18,33 ares

$Q_2 - Q_1$ = 15 ares

$Q_3 - Q_1$ = IQ = 33,33 ares

S_y = 0,1

1,5 x IQ = 50 ares



Exercice

Un magasin de chaussure de Wavre a entrepris une petite étude sur les pointures des chaussures de sport vendues dans le rayon masculin (enfants + adultes) du magasin pendant le mois de juillet.
Voici le tableau résumant la distribution :

Effectifs de la classe	[0,25[	1
Effectifs de la classe	[25,30[	4
Effectifs de la classe	[30,32[	3
Effectifs de la classe	[32,35[	7
Effectifs de la classe	[35,37[	8
Effectifs de la classe	[37,39[	10
Effectifs de la classe	[39,40[	9
Effectifs de la classe	[40,42[	8
Effectifs de la classe	[42,45[	7
Effectifs de la classe	[45,48[	3

Il n’y a aucune observation égale ou supérieure à borne supérieure de la dernière classe

1. Quels sont les individus observés ?
2. Combien y-a-t-il d'observations ?
3. Quelle est la variable ? Quel est le type de variable observée ? Quelles sont les valeurs observables pour cette variable ?
4. Transformer cette distribution en un tableau qui vous permette de calculer les indicateurs suivants :
5. Que vaut l'étendue ?
6. Que vaut la moyenne ?
7. Que vaut l'écart absolu moyen ?
8. Que vaut l'écart type ?
9. Représentez par un graphique approprié cette distribution
10. Comment s'appelle ce graphique ?
11. Quelle est la classe modale et la valeur modale de cette distribution ?
12. Faites le graphe des effectifs cumulés à gauche
13. Donnez une valeur approximative de la médiane (avec justification) et calculez ensuite sa valeur précise.
14. Cette valeur calculée précisément correspond-t-elle parfaitement à la valeur de la médiane de la distribution étudiée ? (Justifiez votre réponse)
15. Estimez la valeur du premier et du troisième quartile (pas de calculs, une estimation !) ainsi que de l'étendue interquartile. Vérifiez ensuite en déterminant les valeurs par calculs.
16. Calculez avec précision la proportion d'individus dont la pointure des chaussures achetées est inférieure à 35 ?
17. Calculez avec précision la proportion d'individus dont la pointure des chaussures achetées est comprise entre 42 et 35 ?
18. Que diriez-vous par rapport au groupe d'un individu dont la pointure vaut 46 ? (indice : utilisez l'écart type)

	Borne inférieure	Borne supérieure	Densité d'effectif	Amplitude de classe	Effectifs	Effectifs cumulés	Centre de classe	$N_i \times X_i$	$N_i \times X_i^2$	Ecart Absolu
[0,25[	0	25	0,040	25	1	0	12,5	12,5	156,3	24,9
[25,30[	25	30	0,800	5	4	1	27,5	110	3025,0	39,7
[30,32[	30	32	1,500	2	3	5	31	93	2883,0	19,3
[32,35[	32	35	2,333	3	7	8	33,5	234,5	7855,8	27,5
[35,37[	35	37	4,000	2	8	15	36	288	10368,0	11,4
[37,39[	37	39	5,000	2	10	23	38	380	14440,0	5,8
[39,40[	39	40	9,000	1	9	33	39,5	355,5	14042,3	18,7
[40,42[	40	42	4,000	2	8	42	41	328	13448,0	28,6
[42,45[	42	45	2,333	3	7	50	43,5	304,5	13245,8	42,5
[45,48[	45	48	1,000	3	3	57	46,5	139,5	6486,8	27,2
	48					60				

Individus: chaussure sport homme+enfant

Variable: pointure

Type: numérique continue (classes)

Effectif total: 60 (= nombre d'observations)

Valeurs observables: 0 à 48

Etendue: 48 – 0 = 48

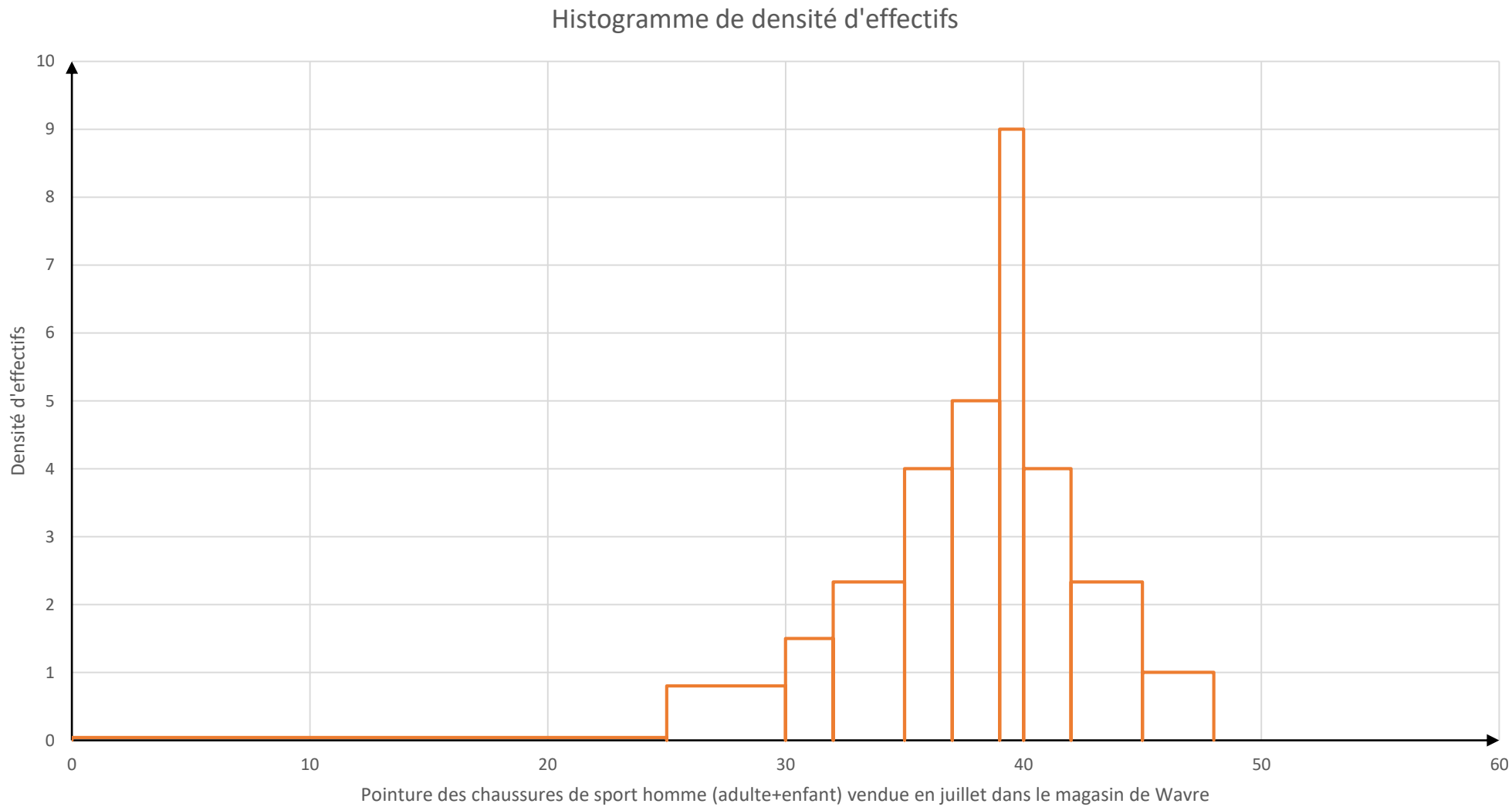
Moyenne: 37,4

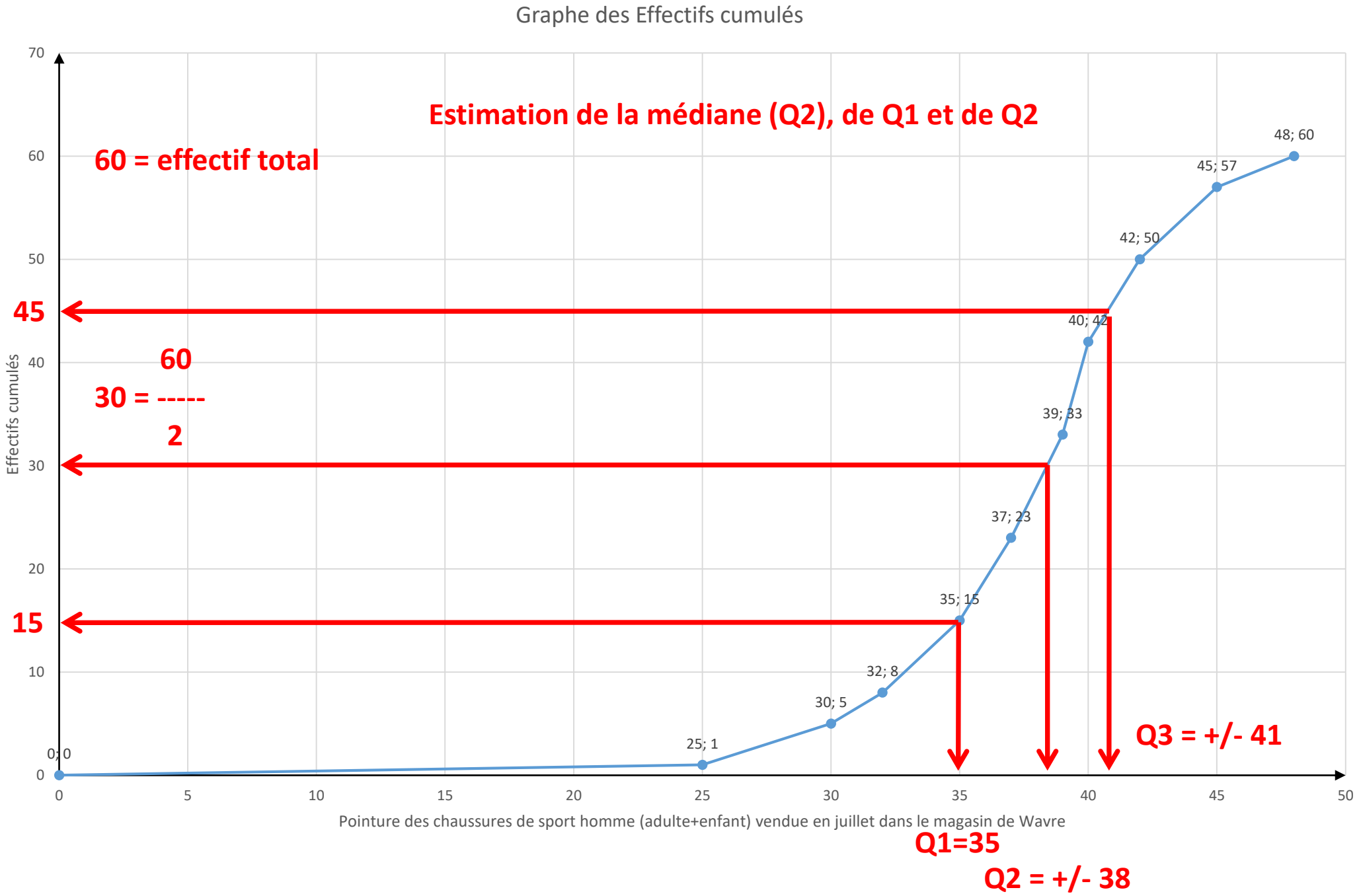
Ecart Type ou σ : 5,65

Ecart absolu moyen: 4,1

Classe modale: [39,40[

Valeur modale: 39,5




Graphe des Effectifs cumulés
Estimation de la médiane (Q2), de Q1 et de Q2
60 = effectif total
45
60
30 = -----
2
15
Q3 = +/- 41
Q1=35
Q2 = +/- 38
Effectifs cumulés
Pointure des chaussures de sport homme (adulte+enfant) vendue en juillet dans le magasin de Wavre
0; 0
25; 1
30; 5
32; 8
35; 15
37; 23
39; 33
40; 42
42; 50
45; 57
48; 60
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
35
40
45
50

Calcul précis de Q1, de la médiane (Q2), de Q3 et de IQ

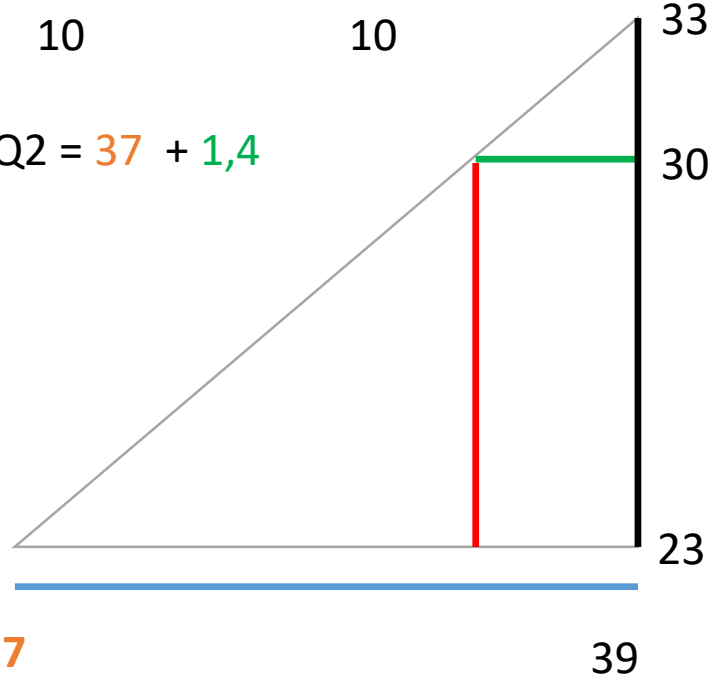
Q1: effectif cumulé = 15 qui apparait justement directement dans la liste des effectifs cumulés et correspond pile poil à la borne supérieur de la classe [32,35[, donc Q1 = 35

Q2: effectif cumulé = 30 donc dans la classe de pointure [37,39[dont les effectifs commencent à 23 et montent à 33

Q3: effectif cumulé = 45 donc dans la classe de pointure [40,42[dont les effectifs commencent à 33 et montent à 42

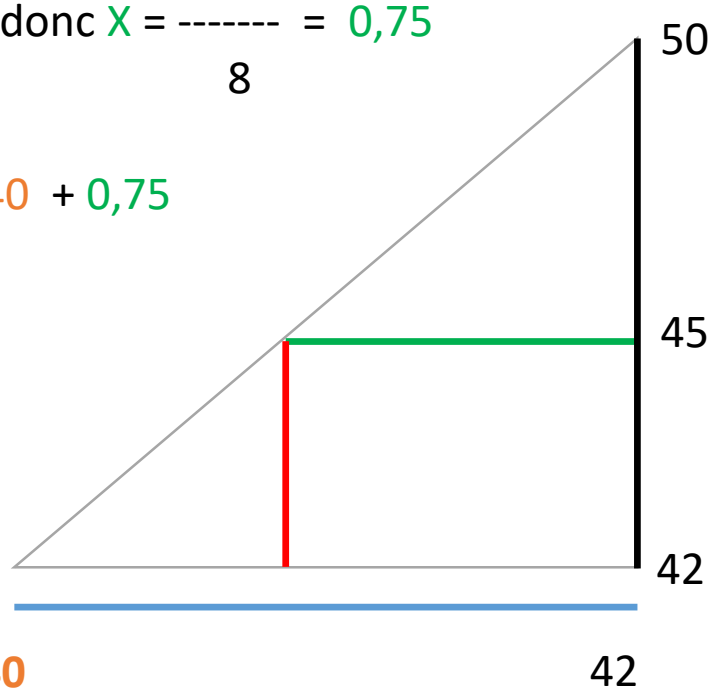
$$\frac{X}{7} = \frac{2}{10} \quad \text{donc } X = \frac{2 \times 7}{10} = 1{,}4$$

Donc Q2 = 37 + 1,4
= 38,4



$$\frac{X}{3} = \frac{2}{8} \quad \text{donc } X = \frac{2 \times 3}{8} = 0{,}75$$

Donc Q3 = 40 + 0,75
= 40,75



IQ = 40,75 – 35 =5,75

Calculez avec précision la proportion d'individus dont la pointure des chaussures achetées est inférieure à 35 ?

35 correspond pile poil à la borne supérieur de la classe [32,35[, il suffit donc de regarder les effectifs cumulés à cette borne: 15
Il y a donc 15/60 observations < 35 soit 25%

Calculez avec précision la proportion d'individus dont la pointure des chaussures achetées est comprise entre 42 et 35 ?

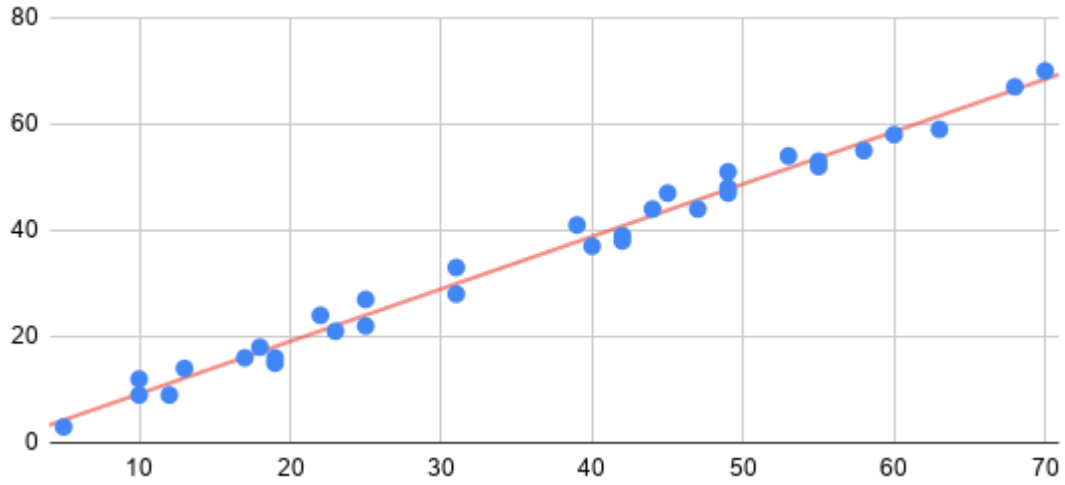
On va calculer le nombre d'observations < 42 (qui par hasard correspond à une borne dont on connaît l'effectif cumulé) et puis on fera la différence entre ce nombre et le nombre d'observations < 35 (cfr ci-dessus) pour obtenir le nombre d'observations comprises entre 35 et 42; il suffira alors de diviser par 60 pour obtenir la proportion:

50 - 15 = 35 et 35/60 = 58,3 %

Que diriez-vous par rapport au groupe d'un individu dont la pointure vaut 46 ? (indice : utilisez l'écart type)

46 est situé à (46 – 37,4)/ 5,65 = 1,5 écart type de la moyenne
donc il n'y a que 22% ($1/(1,5)^2$ = 0,44 soit 44% des gens à plus de 1,5 écart-type de la moyenne de part et d'autre donc 22% des chaussures achetées étaient au dessus de 46 (46 = 37,4 +1,5x5,65) et 22% en dessous de 37,4 – 1,5x5,65 = 28,9)
Conclusion: la proportion des chaussures vendues avec une pointure plus grande que 46 n'est donc que de 22%

Statistique à deux variables



Souvent, nous sommes amenés à observer et à étudier en même temps deux caractères des éléments d'une série statistique et à nous demander si ces caractères sont « liés » et comment ils le sont.

Par exemple, on peut s'intéresser simultanément

- au revenu moyen et à la durée de vie moyenne des individus d'une population ;
- à la consommation de tabac et à la fréquence d'un type de cancer dans plusieurs pays ;
- à l'âge des jeunes de 10 à 18 ans et à l'argent de poche par mois dont ils disposent ;
- à l'année et à la population mondiale
- etc.

On veut savoir si le comportement d'une variable est influencé par la valeur de l'autre:

Revenu moyen	⟺	Durée de vie moyenne
Tabagisme	⟺	cancer
Age	⟺	Argent de poche
Année	⟺	Population

Exemple :

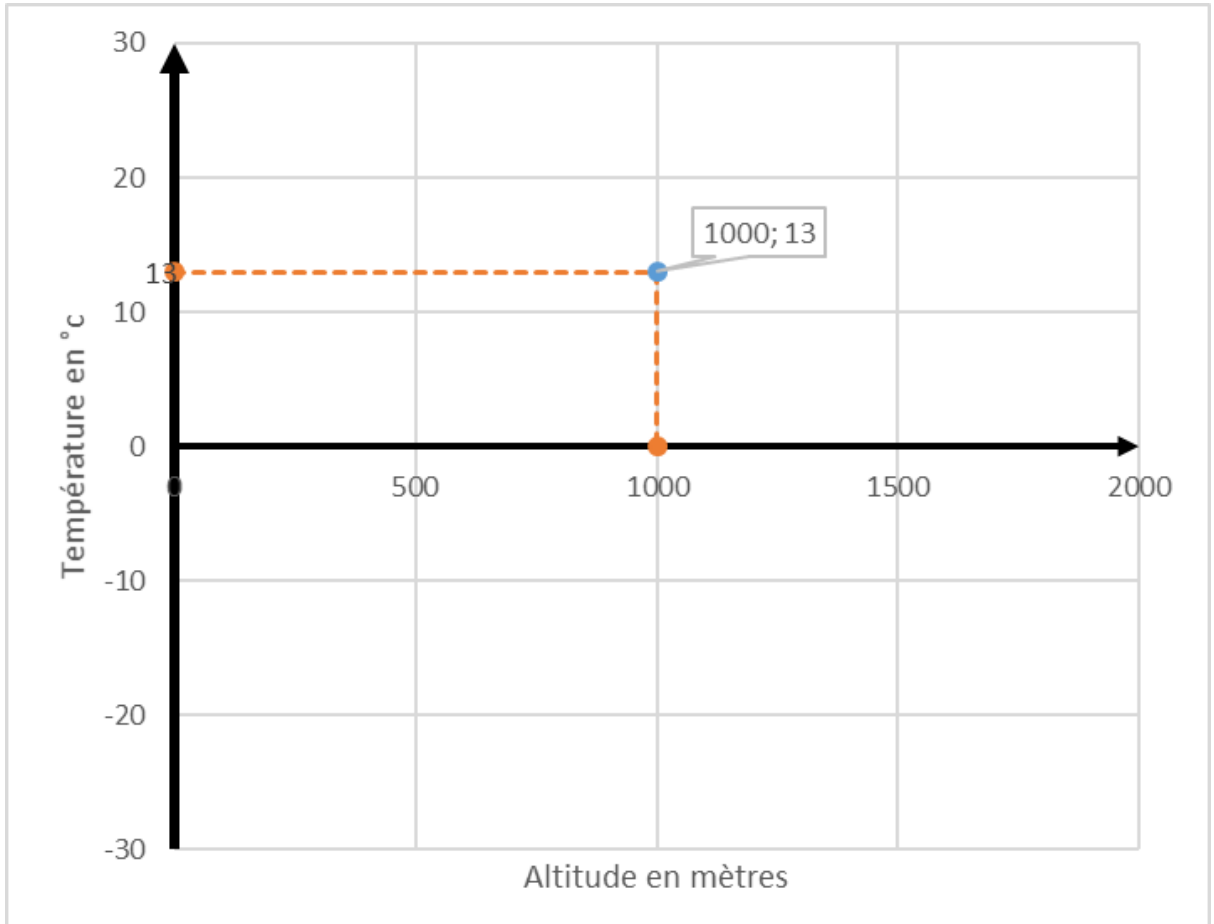
Dans un avion, en cours d'ascension nous notons la température extérieure en degrés Celsius avec l'altitude correspondante en mètres.

Nous obtenons le tableau suivant :

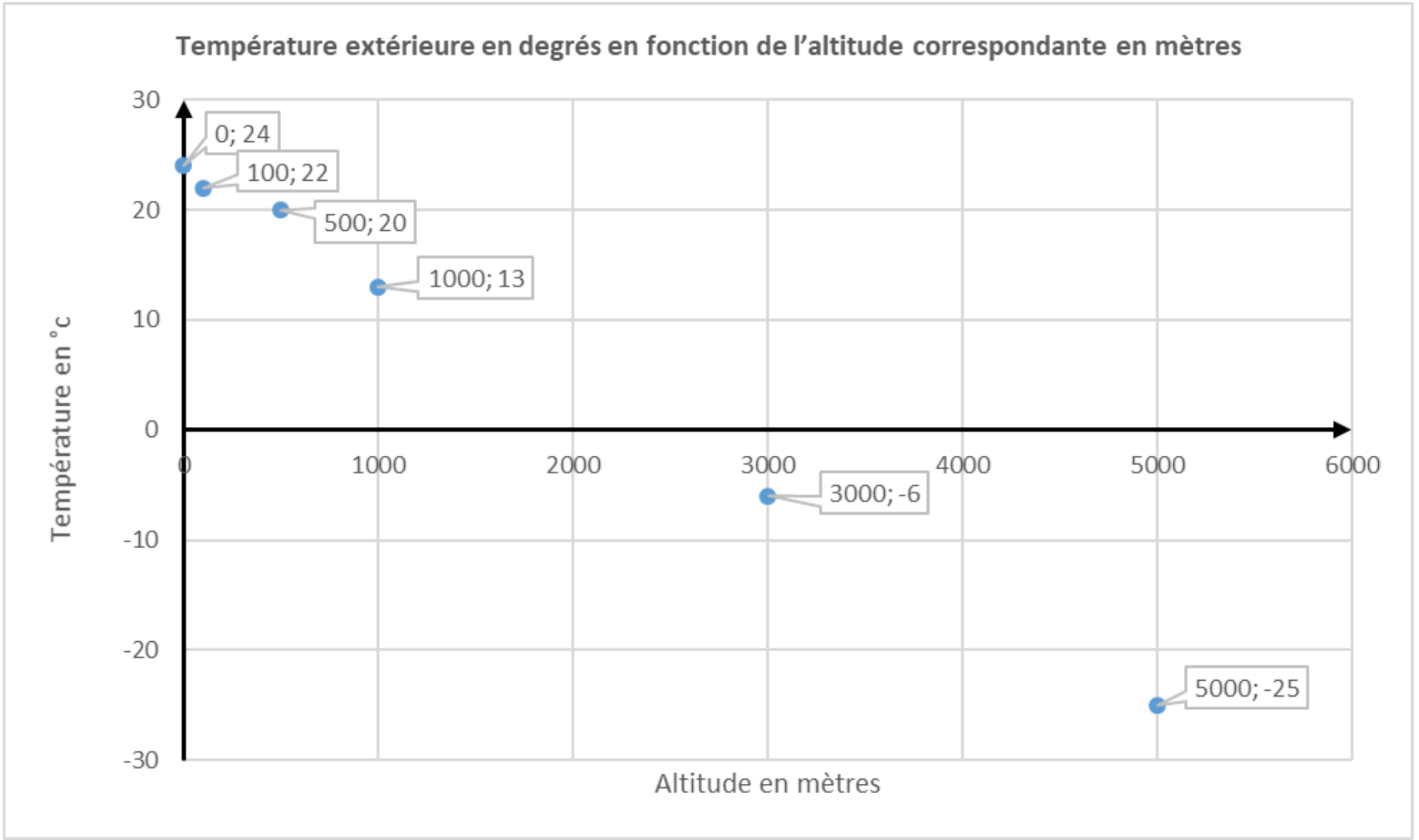
Altitude (xi)	0	100	500	1000	3000	5000
Températures (yi)	24°	22°	20°	13°	– 6°	– 25°

Nous représenterons ces données dans un repère du plan par des points Mi(xi ;yi) afin de constituer ce que nous appelons un nuage de points ou diagramme de dispersion (scatter Diagram) ou encore graphe XY.

Altitude (xi)	0	100	500	1000	3000	5000
Températures (yi)	24°	22°	20°	13°	– 6°	– 25°

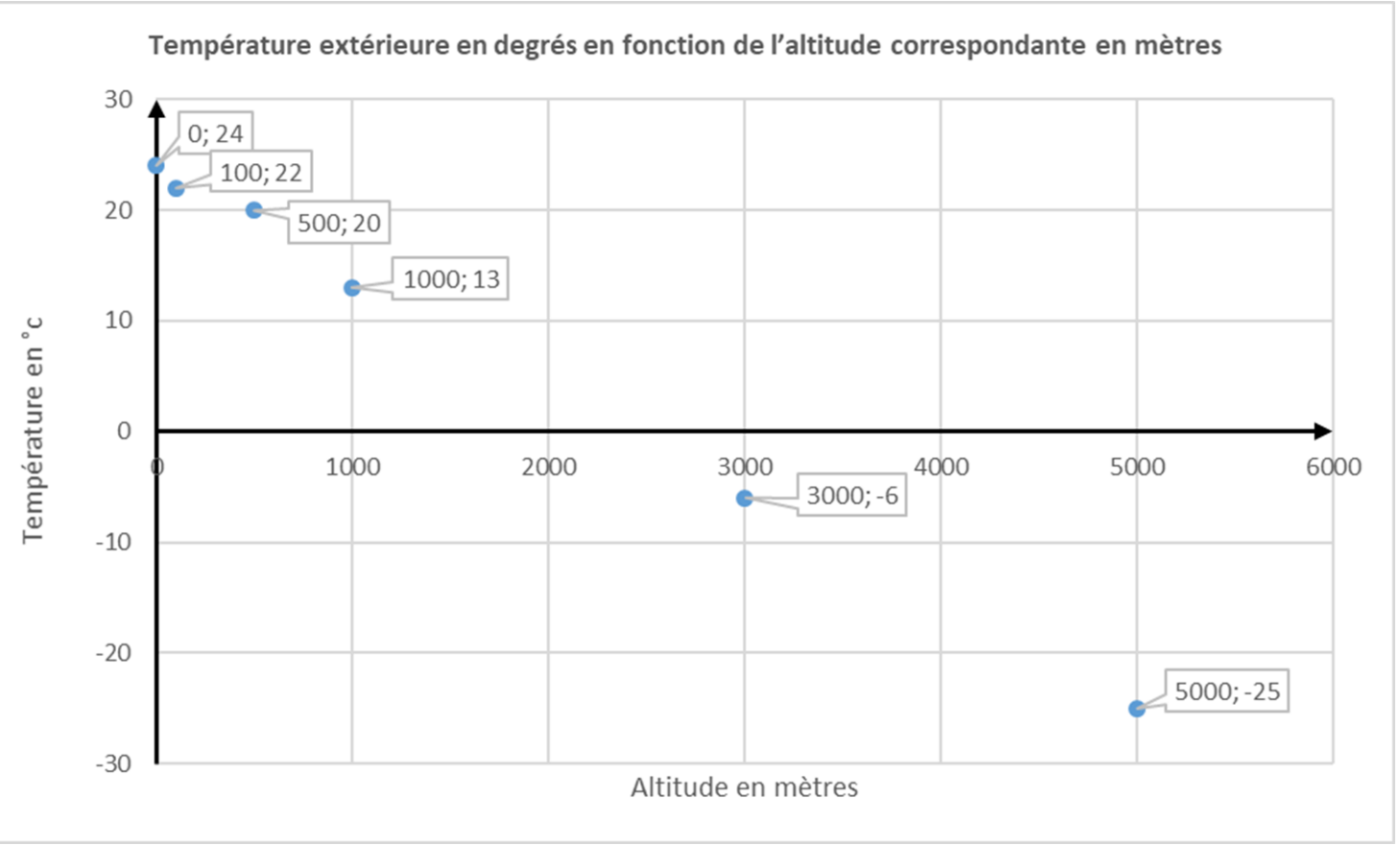


Au final, on obtient:



Utilisation d’Excel pour faire un nuage de points

(voir feuille Excel)

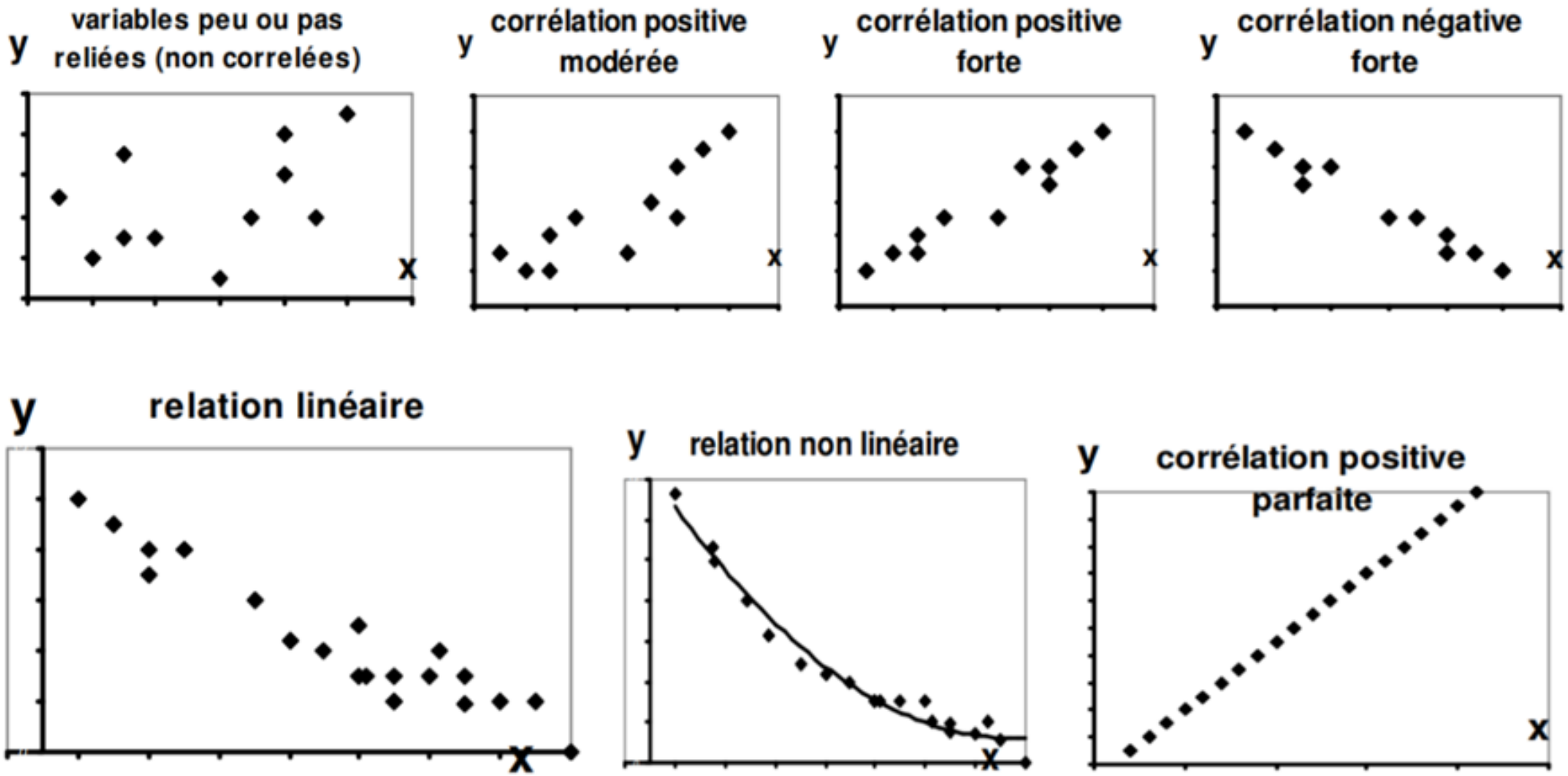


Les points de notre nuage semblent bien s'aligner le long d'une droite descendante.

Nous cherchons à déterminer et à tracer la droite d'équation y = ax+b (c'est l'équation générale d'une droite) passant le plus près possible des points.

Si cela est possible, nous dirons que nous avons réalisé un ajustement affine (on dit parfois abusivement linéaire) du nuage de points et donc que localement nous avons trouvé une relation simple affine (linéaire) de la forme y = ax + b entre les deux variables x et y de notre tableau.

Les points ne s'alignent pas toujours aussi bien le long d'une droite…



Première méthode : méthode de Mayer ou méthode des moyennes discontinues

Pour réaliser un ajustement affin, nous avons une première méthode qui donne « la droite de Mayer ».

Une droite étant déterminée quand on connaît deux de ses points, on peut chercher à déduire à partir des données deux points qui permettront d'ajuster une droite à ces données.

Pour cela, on partage le nuage de points en deux sous-nuages de même importance.
Par exemple, on met dans le premier groupe la première moitié de l'effectif (après avoir ordonné les valeurs de x), et dans le deuxième groupe la seconde moitié. Pour un nombre impair de données, on inclut un point (du milieu) dans les deux groupes.

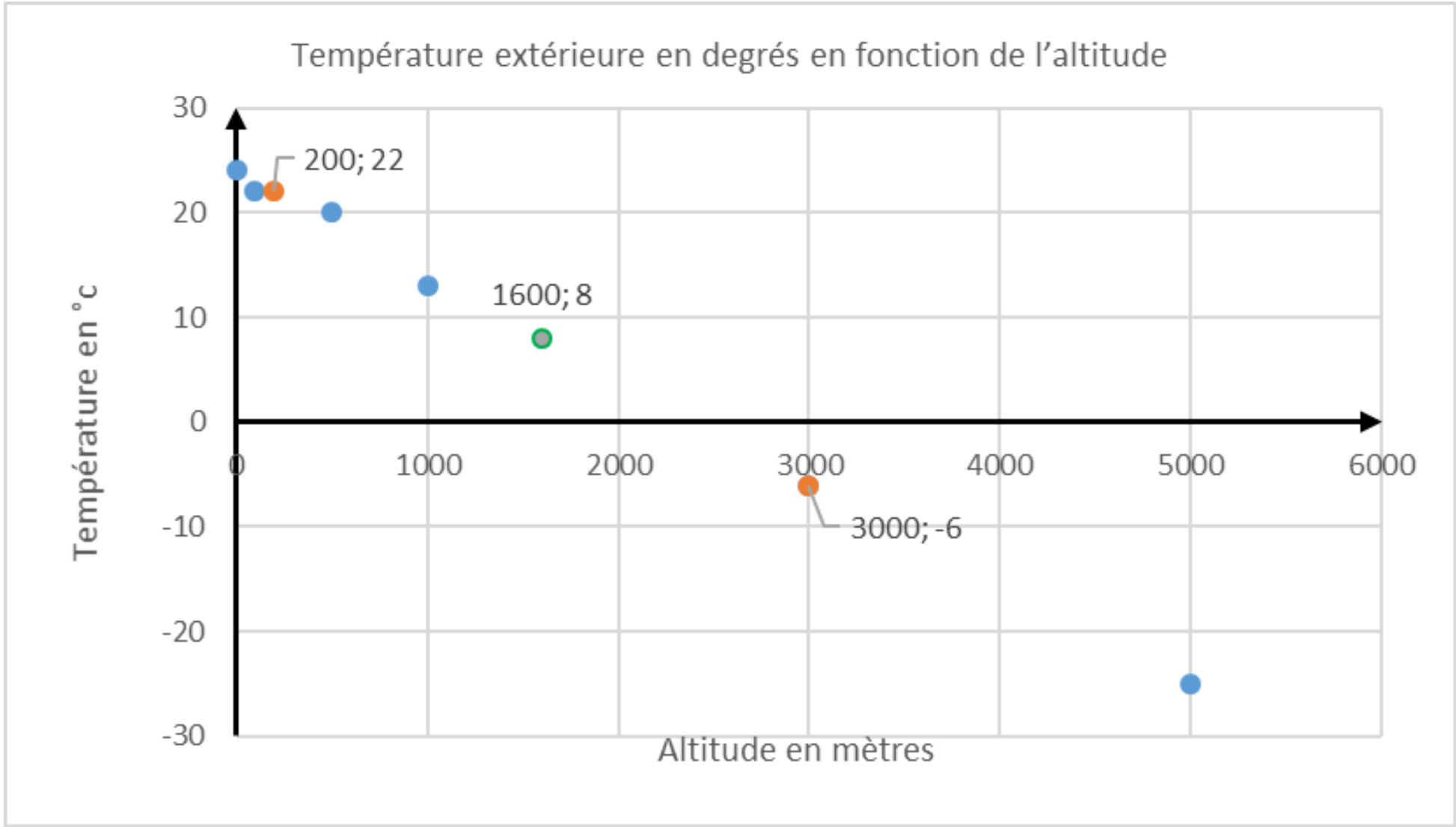
Ensuite pour chaque groupe, on détermine les points moyens de ces deux sous-nuages G1 et G2.
La droite qui passe par les deux points moyens G1 et G2 est choisie comme droite d'ajustement (Droite de Mayer).

X	0	100	500
y	24°	22°	20°

X	1000	3000	5000
y	13°	-6°	-25°

$G_1(200 ; 22)$

$G_2(3000 ; -6°)$


Température extérieure en degrés en fonction de l'altitude
Température en °c
Altitude en mètres
30
20
10
0
-10
-20
-30
0
1000
2000
3000
4000
5000
6000
200; 22
1600; 8
3000; -6

La droite passant par G_1 et G_2 (Droite de de Mayer) a une équation de la forme $y = ax + b$.

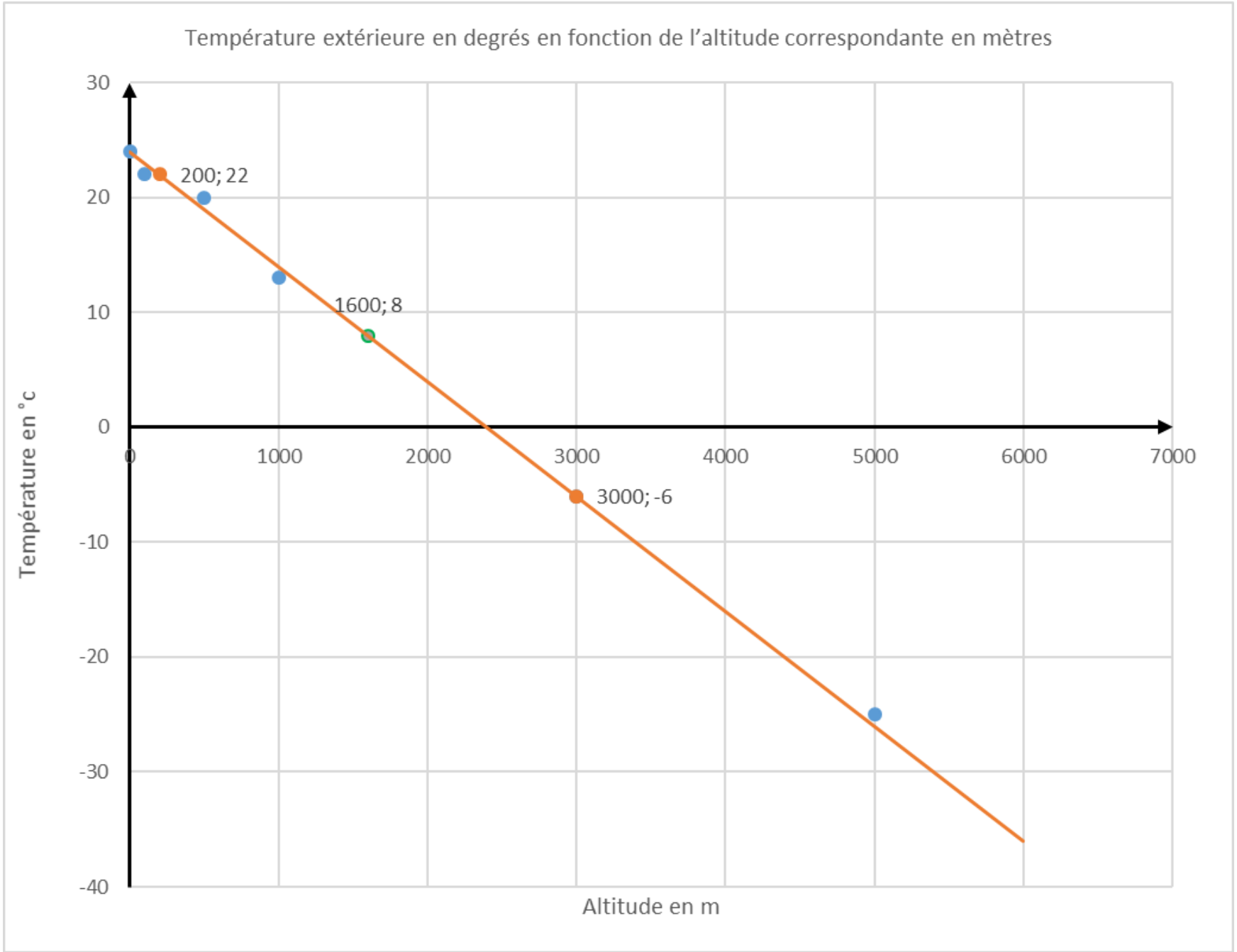
Où a vaut : $\dfrac{(Y_{G2} - Y_{G1})}{(X_{G2} - X_{G1})}$

Soit dans notre exemple : $a = (-6 - 22)/(3000 - 200) = -0{,}01$.

Pour trouver b, nous utilisons un des points, par exemple $G_1(200 ; 22)$ et on introduit les valeurs correspondantes dans l'équation :

$22 = -0{,}01(200)+b$ et donc $22 = -2 + b$ c'est-à-dire $b = 24$.

La droite **(G_1G_2) a donc pour équation $y = -0{,}01x + 24$**.


Température extérieure en degrés en fonction de l'altitude correspondante en mètres
Température en °c
Altitude en m
30
20
10
0
-10
-20
-30
-40
0
1000
2000
3000
4000
5000
6000
7000
200; 22
1600; 8
3000; -6

Ressource vidéo :

https://www.youtube.com/watch?v=vNpBIxdreIg (4’58’’) Nuage de points et relation entre les variables

https://www.youtube.com/watch?v=Nn6uckb3RvE 4’39’’ Nuage de points et Point moyen

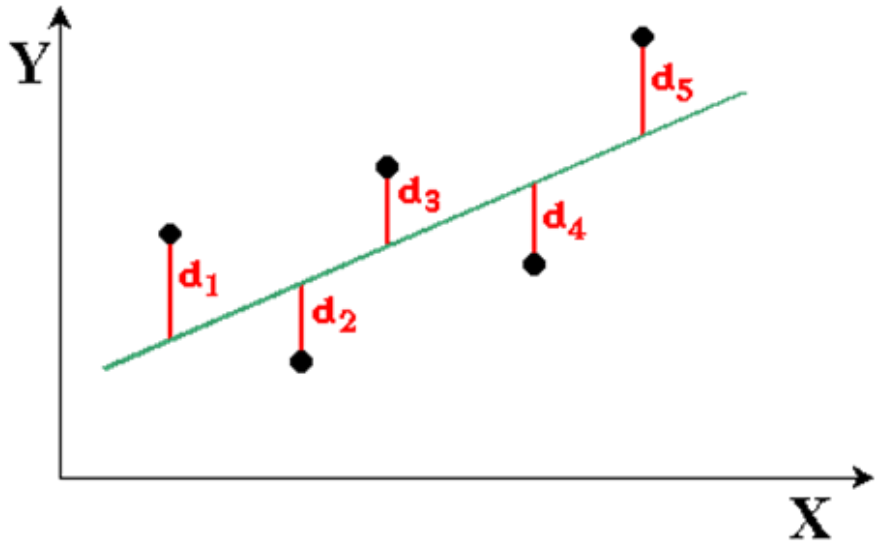
https://www.youtube.com/watch?v=KjMNSQGJFqU 9’31’’ Droite d’ajustement, Droite de Mayer

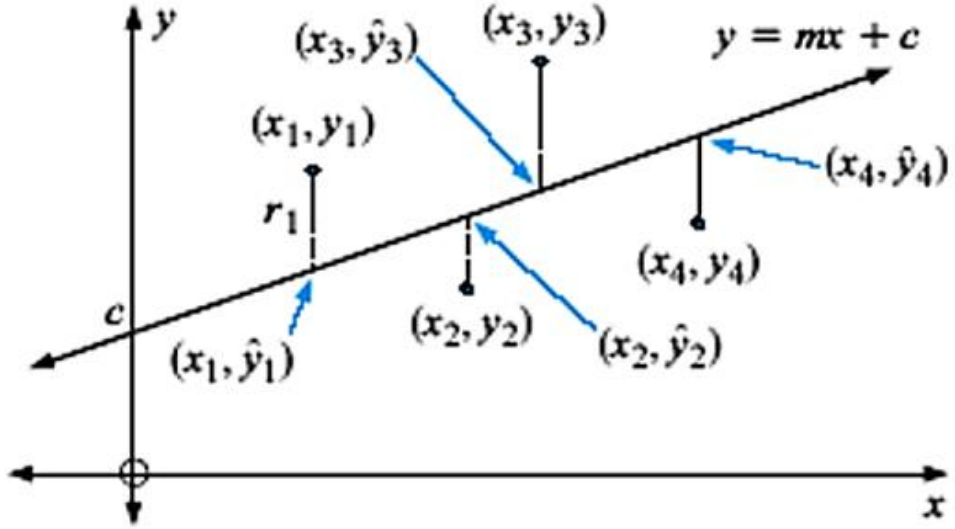
N’hésitez pas à aller voir les ressources vidéo pour revoir la matière ou si vous ne comprenez pas du premier coup!

Deuxième méthode : méthode des moindres carrés

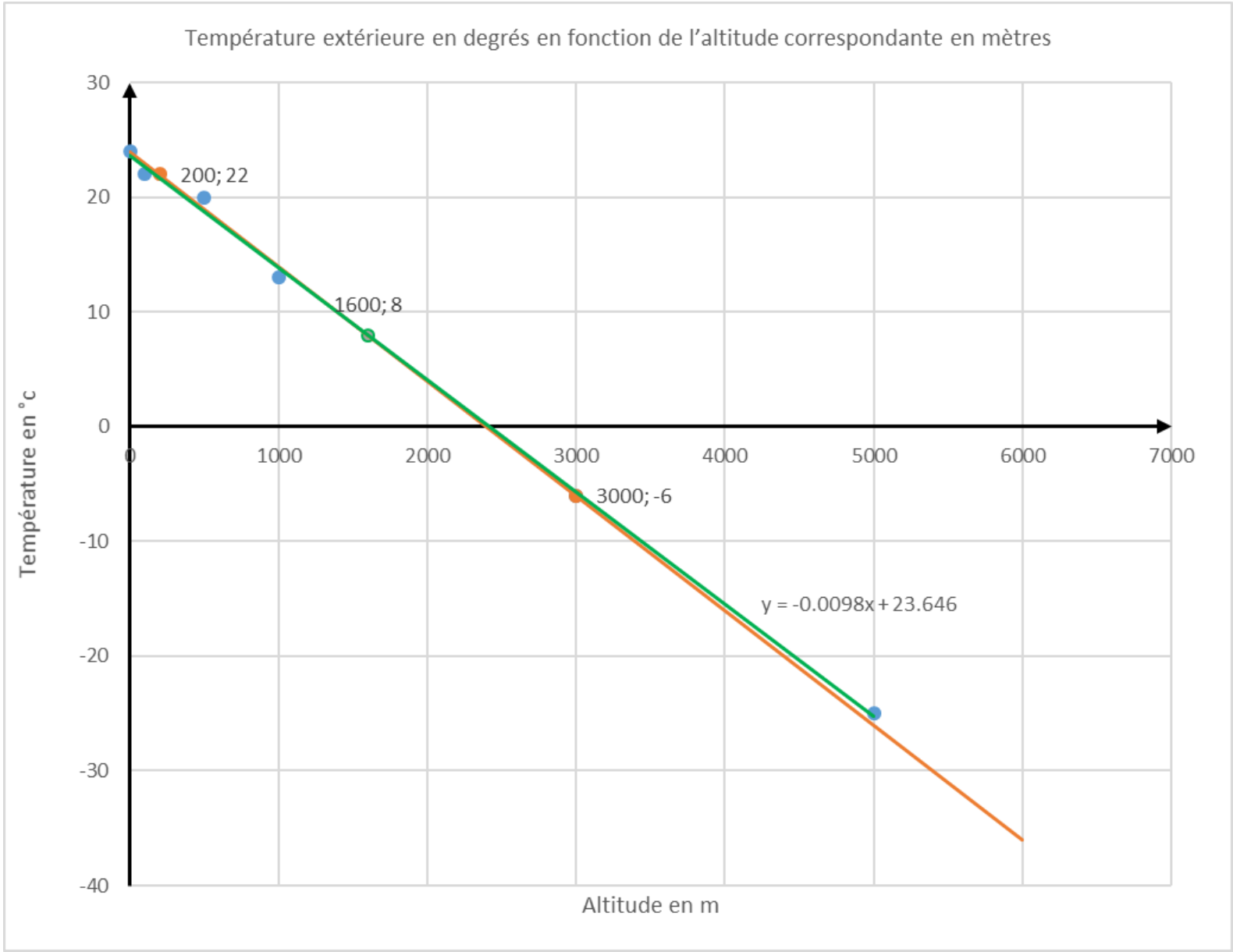
la méthode dite « des moindres carrés » s'est imposée face à la méthode de Mayer.

Pour mesurer la qualité de la droite d'ajustement d'équation $y = ax + b$, on considère, pour chaque valeur x_i, la différence entre la valeur observée y_i, et la valeur correspondant à la droite calculée par la formule $\hat{y}_i = ax_i + b$. On prend donc en compte l'écart (en rouge) entre les points (en noir) et la droite d'ajustement (en vert) dans la figure ci-dessous.





Voilà ce que cela donne: en vert: Droite de la méthode des moindres carrés, en orange: droite de Mayer



On constate que :

- Les deux droites (Mayer en orange et moindres carrés en bleu) sont dans notre exemple très proches l'une de l'autre.

- Elles passent toutes les deux aussi par le point moyen G(1600 ; 8°).

Comment calculer les coefficients a et b de l'équation de la droite ?

Les coefficients *a* et *b* peuvent être calculés à partir des formules suivantes:

a (= Pente de la droite):

$$a = \frac{(X_1 - \overline{X}).(Y_1 - \overline{Y}) + (X_2 - \overline{X}).(Y_2 - \overline{Y}) + \ldots + (X_n - \overline{X}).(Y_n - \overline{Y})}{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}$$

ou

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

B (= Ordonnée à l'origine)

$$b = \overline{Y} - a.\overline{X}$$

Rappels:

$$\overline{X} = \frac{1}{n} \sum X$$

$$\overline{Y} = \frac{1}{n} \sum Y$$

En résumé: méthode des moindres carrés

Droite ax + b

Calcul de a:

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

Calcul de b:

$$b = \overline{Y} - a.\overline{X}$$

Avec

$$\overline{X} = \frac{1}{n} \sum X$$

Moyenne des X

$$\overline{Y} = \frac{1}{n} \sum Y$$

Moyenne des Y

Utilisation d'Excel pour calculer l'équation de la droite
Et récupérer les valeurs de a et b

(voir feuille Excel)

Coefficient de corrélation (de Pearson)

Même si on arrive (presque) toujours à déterminer une droite d'ajustement pour un nuage de points (x_i , y_i), les variables x et y peuvent être corrélées à des degrés très différents, et la droite trouvée ne s'ajuste pas nécessairement bien aux données.

Le signe de la pente a donne le sens de la corrélation, mais pas sa qualité.

$a > 0$	corrélation positive
$a < 0$	corrélation négative
$a = 0$	pas de corrélation

Coefficient de corrélation (de Pearson ou de Bravais-Pearson)

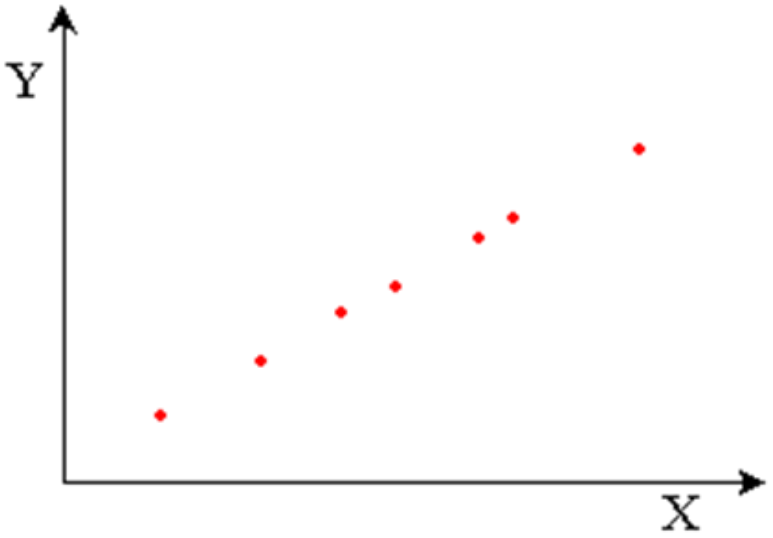
La qualité de la corrélation peut être mesurée par le **coefficient de corrélation** ***r***.

$$r = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \times \sqrt{\sum (Y - \overline{Y})^2}}$$

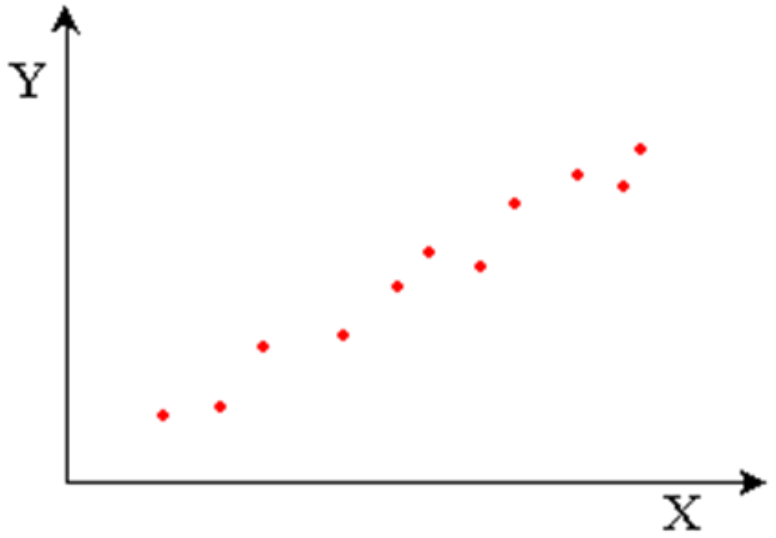
Le coefficient de corrélation est compris entre -1 et +1.

$r = +1$	corrélation positive parfaite	$r > 0$	corrélation positive
$r = -1$	corrélation négative parfaite	$r < 0$	corrélation négative
$r = 0$	absence totale de corrélation	$r = 0$	pas de corrélation

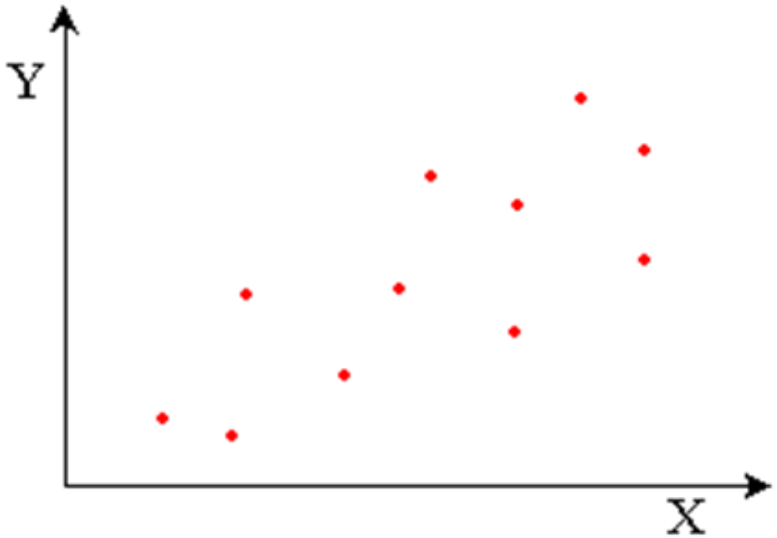
Plus il s'éloigne de zéro, meilleure est la corrélation.
Ce nombre décrit la validité de la droite d'ajustement et mesure le degré de dépendance linéaire entre les variables x et y



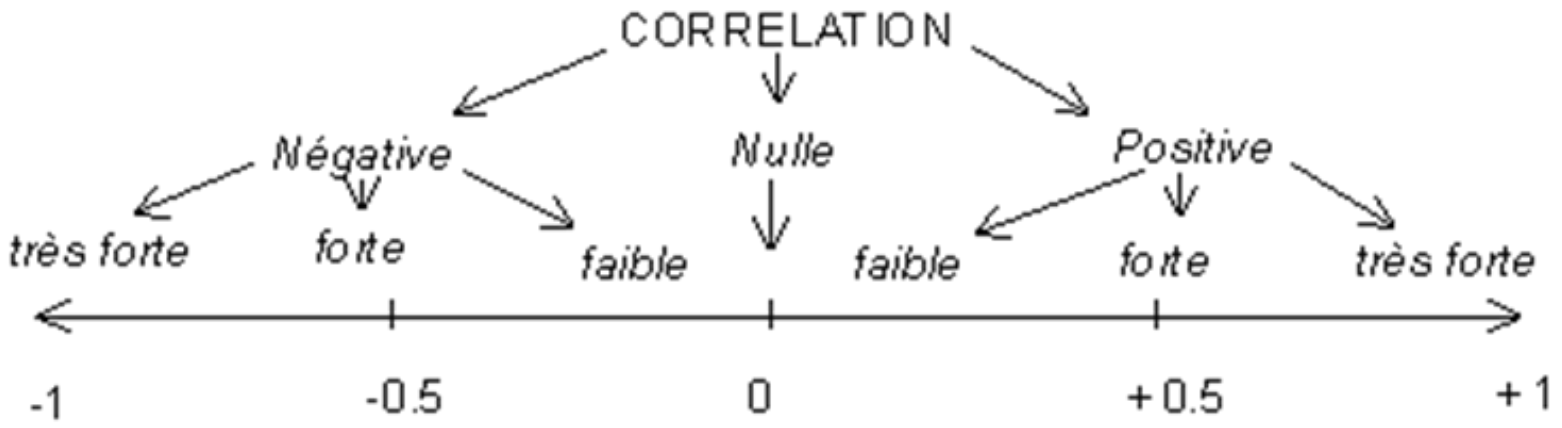
Corrélation parfaite



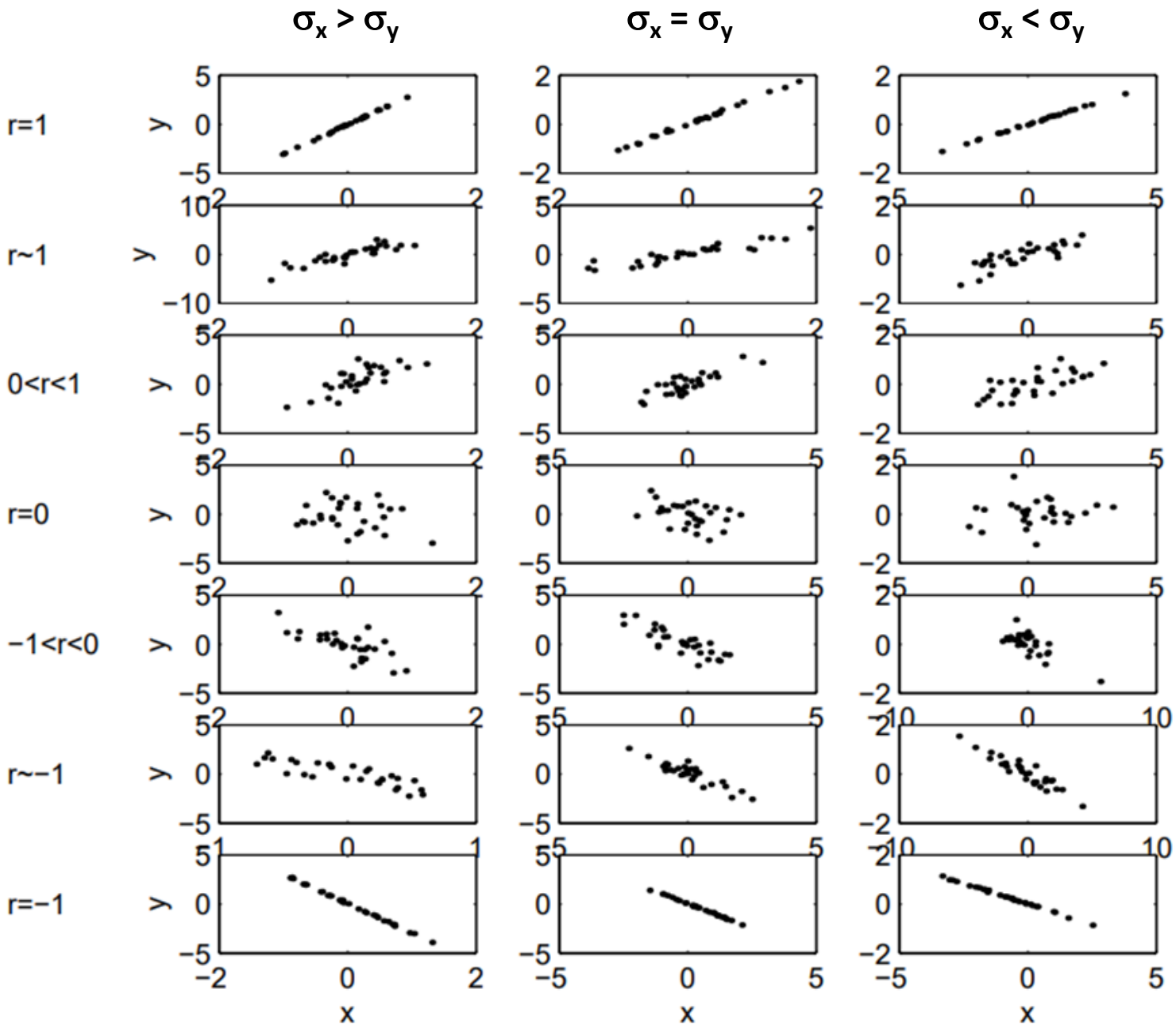
Corrélation forte



Corrélation faible



Liens entre les nuages de points, le cœfficient de corrélation linéaire r et les écarts types σ_x et σ_y des variables

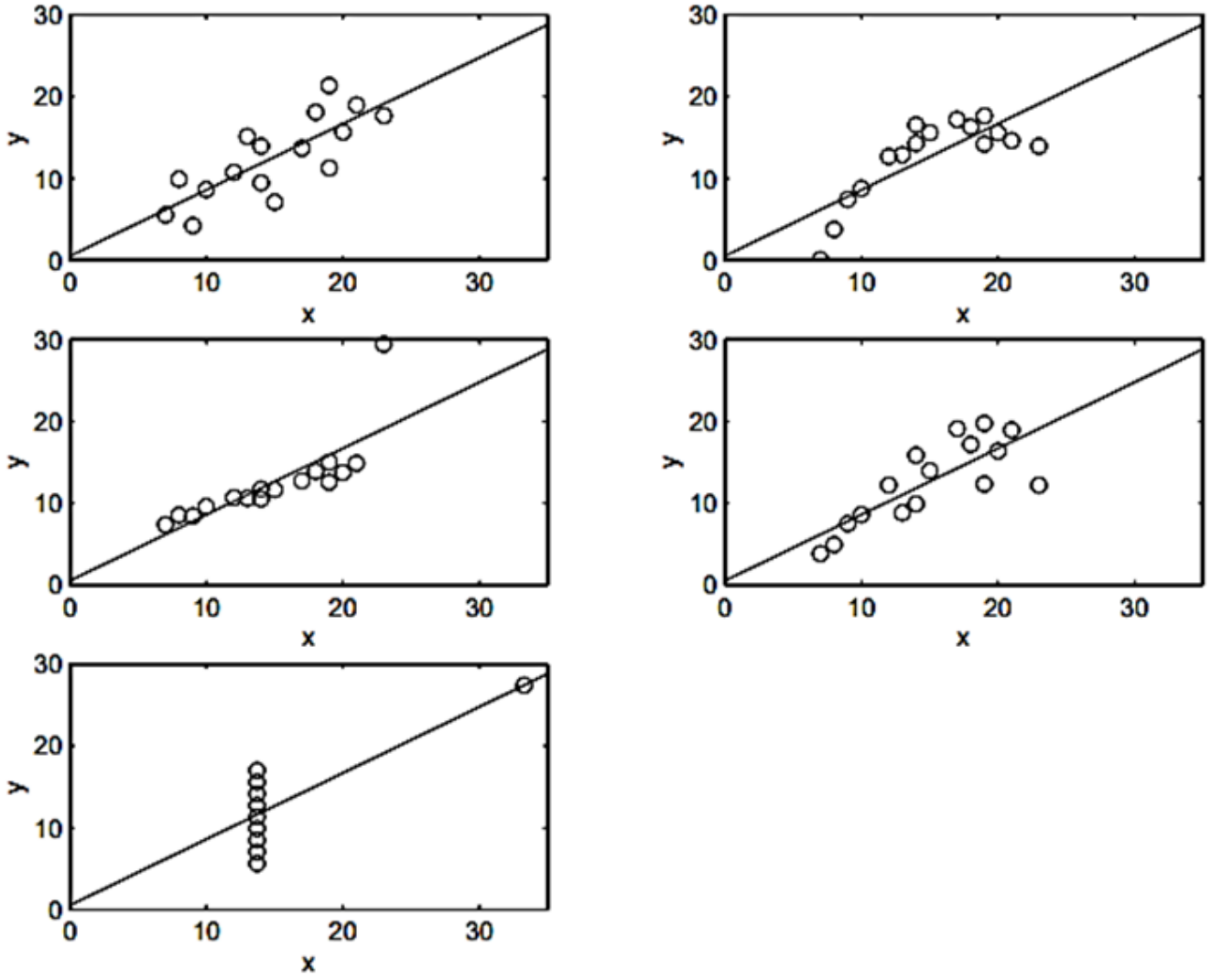


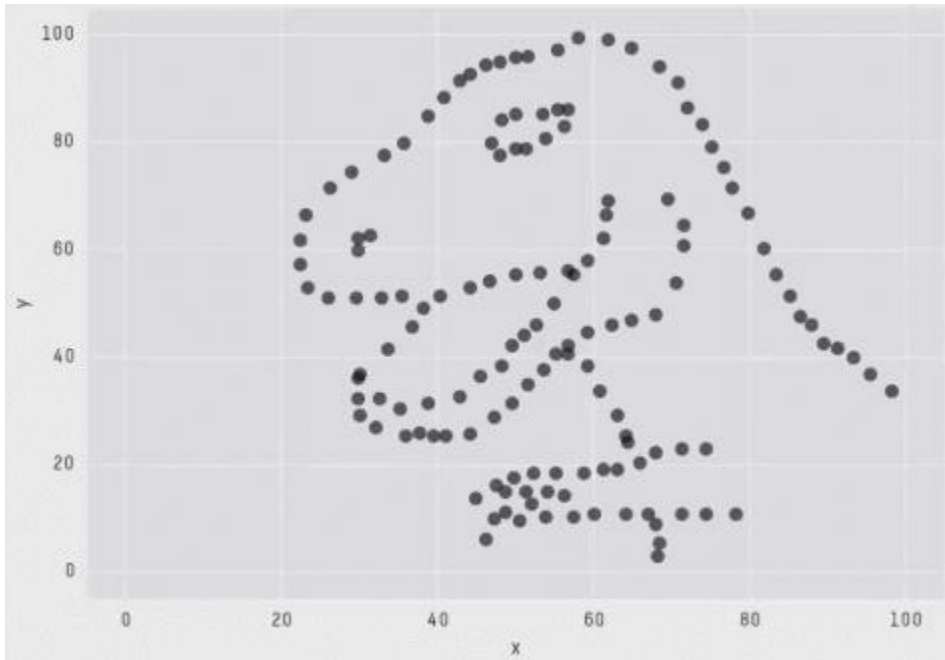
Attention toujours dessiner le nuage de points

Le coefficient de corrélation linéaire ne mesure que la possibilité d'une relation affine entre les variables.

Pour les 5 graphiques de la figure suivante, la moyenne en x, la moyenne en y, S_x et S_y et r ont la même valeur et par conséquent la droite de régression linéaire a la même équation.

En regardant l'allure des nuages de points on comprend pourtant immédiatement que la relation entre les variables n'est pas du tout la même.





X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Remarques:

Le coefficient de corrélation linéaire nous donne des informations sur l'existence d'une relation affine ou linéaire (sous forme d'une droite) entre les deux grandeurs considérées.

Un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux grandeurs. Il peut exister une relation non linéaire entres elles.

Il ne faut pas confondre corrélation et relation causale.
Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

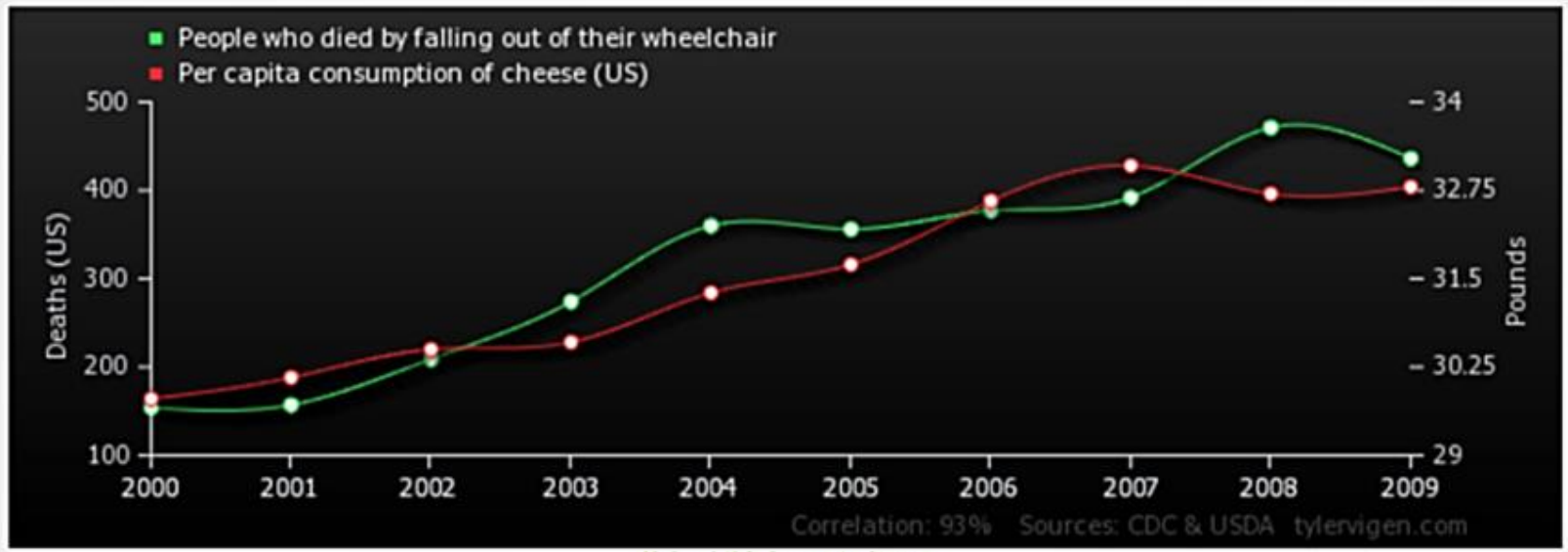
L'existence d'une corrélation, aussi bonne soit elle,
n'est jamais la preuve d'une relation de cause à effet.

L'existence d'une corrélation, aussi bonne soit elle, n'est jamais la preuve d'une relation de cause à effet.

Nombre de personnes handicapées décédées d'une chute de leur fauteuil

corrélé avec

La consommation de fromage par habitant



Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
People who died by falling out of their wheelchair Deaths (US) (CDC)	154	157	209	274	360	356	377	392	471	436
Per capita consumption of cheese (US) Pounds (USDA)	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8

Correlation: 0.931497

Crédits : Tyler Vigen – http://www.tylervigen.com/spurious-correlations

Covariance

La covariance est notion élargie de celle de variance.
Pour rappel, la variance vaut :

$$\frac{\sum n_i \left(x_i - \bar{x}\right)^2}{\sum n_i} = \frac{\sum n_i \, x_i^2}{\sum n_i} - \bar{x}^2$$

La covariance est la moyenne du produit des écarts à la moyenne.
La covariance est donc définie par

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right) . \left(Y_i - \bar{Y}\right)$$

ou

$$Cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i \;.\; Y_i\right) - \left(\bar{X} \;.\; \bar{Y}\right)$$

- La deuxième forme est généralement plus facile à calculer.

- On constate que la covariance d'une variable avec elle-même correspond à la variance de cette variable.

A quoi ça sert?

La covariance peut être positive ou négative.
Une covariance positive (respectivement négative) indique une relation entre les données croissantes (respectivement décroissantes), i.e. que les valeurs élevées d'une série correspondent, dans l'ensemble, à des valeurs élevées (respectivement faibles) de l'autre.

La notion de covariance nous permet aussi de réécrire les calculs des paramètres de la droite de régression linéaire et du coefficient de corrélation sous la forme :

a = Cov(X,Y) / $(\sigma_X)^2$

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right).\left(Y_i - \overline{Y}\right)$$

b = moyenne(Y) - a. moyenne(X)

$$b = \overline{Y} - a.\overline{X}$$

Elle nous permet aussi de réécrire la formule du coefficient de corrélation linéaire de Pearson :

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X . \sigma_Y}$$

Exemple de calcul du coefficient de corrélation de Bravais-Pearson:

Calcul de la corrélation linéaire entre taille des pieds et intelligence de 10 enfants d'âge scolaire

Tableau :

enfant (i)	Xi	Yi	(Xi -mX)	(Yi -mY)	(Xi-mX)(Yi-mY)
A	31	50	-3	-26	78
B	31	55	-3	-21	63
C	32	52	-2	-24	48
D	33	56	-1	-20	20
E	33	63	-1	-13	13
F	34	65	0	-11	0
G	35	69	1	-7	-7
H	36	90	2	14	28
I	37	110	3	34	102
J	38	150	4	74	296
moyenne	34	76	0	0	64.1
écart-type	2.4	32			

La covariance de X et Y étant égal à 64.1, on obtient le coefficient de corrélation de X et de Y en divisant la covariance par le produit de l'écart-type de X et de l'écart-type de Y : r(X,Y) = 64.1 / (2.4 * 32) = +0.83

Qu’est-ce que l’on peut dire à partir de là?

Qu’est-ce que l’on peut dire à partir de là?

+0,83 : nous sommes en présence d'une corrélation positive forte, qui semble indiquer qu'il existe une relation linéaire (de type Y=aX+b) reliant le quotient intellectuel des enfants et la taille de leurs pieds.

Toutefois, le coefficient de corrélation ne nous indique pas:

- si la relation observée est significative (fruit du hasard ou non)
- si elle correspond à une relation de cause à effet entre les deux facteurs X et Y étudiés.

De plus, l'importance de la corrélation linéaire ne préjuge pas de l'existence d'un meilleur ajustement, qui serait quant à lui de type non-linéaire.

Exercices

Exercice 5.1

Nous voulons étudier l'évolution de la population d'une commune.

Un relevé a été fait donnant le tableau ci-dessous
Calculer les coordonnées du point moyen G.
Représenter ce nuage de points.
Avec Excel ou votre calculatrice, déterminer les coefficients a et b
de la droite d'ajustement par la méthode des moindres carrés.
Tracer la sur le graphique.
Vérifier que G appartient à cette droite.
Quelle prévision pour 2020 cette droite permet-elle de faire ?

Années	1980	1990	2000	2002	2010
Population y	2030	2500	3000	3200	3400

Exercice 5.2

Le PDG d'une entreprise fait analyser la production d'un produit sur 10 ans.
Nous avons le tableau ci-dessous

Représenter graphiquement ces données.

Pourquoi un ajustement affine est - il possible ?

Placer G le point moyen.

Tracer la droite (D1) passant par G et le dernier point (10 ; 68). ON considère qu'elle réalise un ajustement linéaire valable du nuage. Donner l'équation de (D1).
Utiliser votre calculette pour déterminer a et b les coefficients de la droite (D2) d'ajustement affine par la méthode des moindres carrés. Tracer (D2).
Faire une prévision pour 15 ans en utilisant (D1) et (D2). Quelle est l'erreur en % commise en prenant (D1) à la place de (D2).

Années x	1	2	3	4	5	6	7	8	9	10
Production y	49	48	50	50	56	57	62	65	65	68

Exercice 5.3

Nous avons le tableau suivant :

A	B	C	
1	x_i	y_i	ax_i+b
2	20	50	?
3	30	68	?
4	50	108	?
5	70	150	?
6	80	175	?
7	100	220	?
8	120	250	?

Entrer ces données dans une feuille de calcul Excel.

En utilisant les commandes :

=droitereg(B2 :B11 ;A2 :A11) et
=ordonnee.origine(B2 :B11 ;A2 :A11)

déterminer a te b les coefficients de la droite (D) d'ajustement par la méthode des moindres carrés.

Calculer alors $ax_i + b$

Faire un graphique dans la feuille pour illustrer ceci.
(En sélectionnant la colonne x_i et $ax_i + b$, nous pouvons tracer (D))

Exercice 5.4

Un couple de restaurateur étudie une formule Brunch-Culture. Ils ont recensé le nombre de personnes intéressées en fonction du prix fixé.
Soit x_i le prix en euros et y_i le nombre de personnes correspondant à ce prix.

x_i	y_i
18	47
20	45
23	42
25	40
28	36
30	30
33	25
35	22
38	18
40	15

1a Représenter graphiquement ces données.

1-b Peut-on émettre l'hypothèse d'une relation simple entre x et y. Si oui, quelle genre de formule proposez-vous ?

2 Déterminez les coordonnées du point moyen G du nuage représentés précédemment.

3 On choisit de faire un ajustement affine par la droite (D) de coefficient directeur -1,5 passant par G. Donner l'équation réduite de cette droite (D) puis tracer la. Lire sur le graphique à partir de quel prix, personne ne viendra utiliser la formule proposée. Vérifier par le calcul.

4 Quelle prévision donne (D) si on choisit x = 25€. Quel est en % l'erreur commise avec la réalité ?

Correction Exercice 5.5

Un hypermarché dispose de 20 caisses. On s'intéresse au temps moyen d'attente en fonction du nombre de caisses ouvertes un jour de semaine. Le tableau ci-dessous donne x le nombre de caisses ouvertes et y le temps moyen d'attente correspondant :

Xi	3	4	5	6	8	10	12
Yi	16	12	9.6	7.9	6	4.7	4

1) Calculez la moyenne du nombre de caisses ouvertes et du temps d'attente moyen**.**

2) Montrer que la droite de régression des moindres carrés passe par le point moyen.

3). Y a-t-il corrélation entre le nombre de caisses et le temps d'attente ?

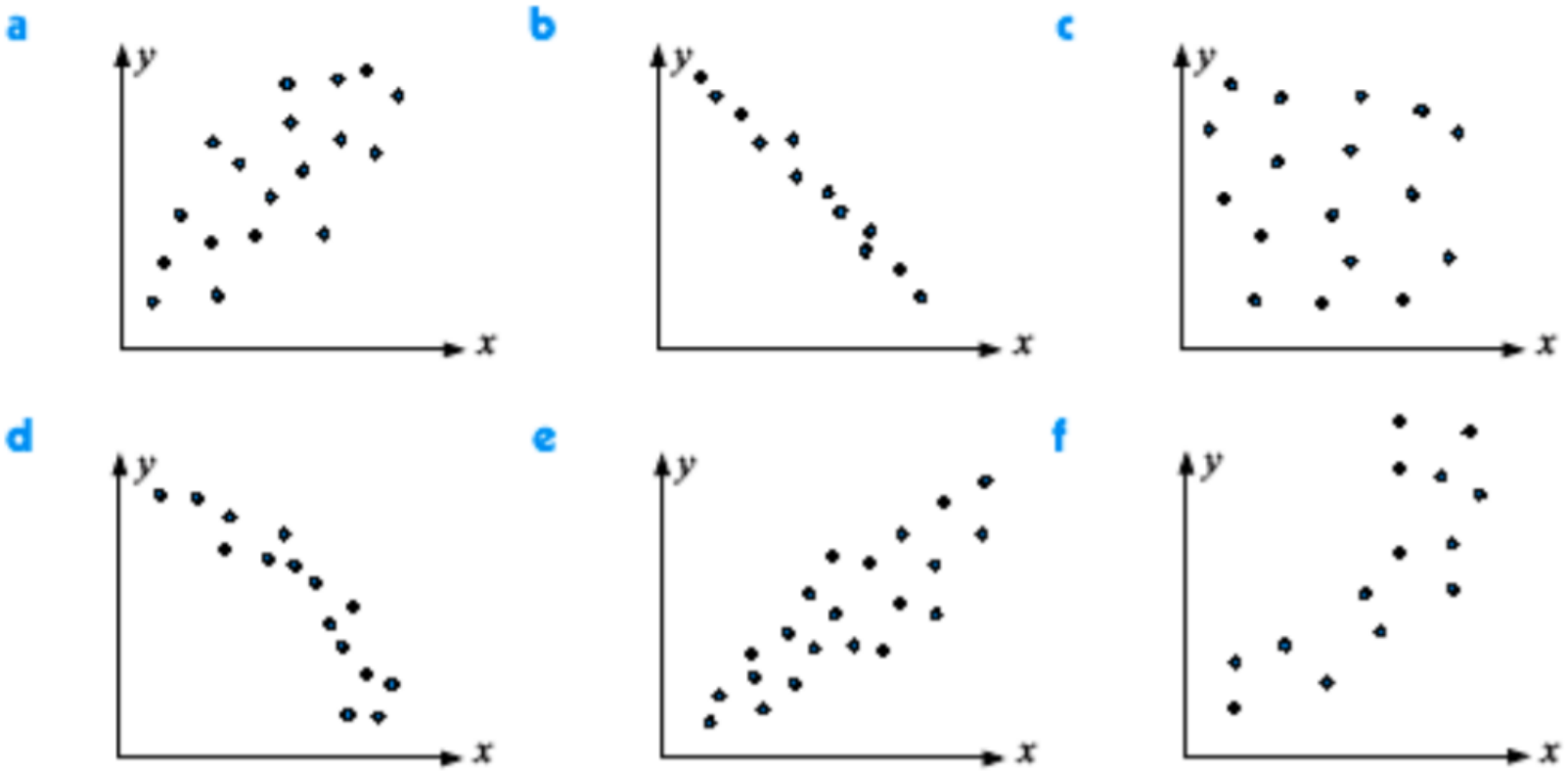
4) Selon vous, y a-t-il causalité entre le nombre de caisses et le temps d'attente ?

5) Faites une estimation du temps d'attente si on ouvre qu'une caisse, 7 caisses, les 20 caisses.

Exercice 5.5

Dites pour chacun des nuages de points

- s'il y a un lien positif, un lien négatif, ou pas de lien entre les variables
- si la relation est linéaire ou autre
- le degré du lien (zéro. faible. moven. fort)



Exercice 5.6

Voici le tableau présentant la population de New York (en milliers d'habitants) entre 1800 et 1910.

Année	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910
Milliers hab.	75	110	130	230	400	700	1100	1400	1900	2300	3200	4500

Calculer la droite d'ajustement par les moindres carrés.

Calculer le coefficient de corrélation linéaire entre x et y

Y-a-t-il une relation linéaire entre ces deux variables ?

Faites le graphe en nuage et faites-y apparaître la droite de régression linéaire.

Quelle conclusion en tirez-vous ?