Professeur	Bahloul Khalid (+212) 622-17-65-52
Chapitre	Statistiques (l'essentiel du cours + applications)
Niveaux	1ère et 2ème Bac International et Bac français

Série statistique à une variable

1- Définitions et vocabulaire

Population	l'ensemble de tous les individus ou éléments sur lesquels porte une étude statistique
Individu	un élément de base sur lequel porte l'étude statistique
Caractère	caractéristique d'une population que l'on étudie
Effectif	le nombre de fois qu'une valeur ou classe apparaît
Effectif total	somme de tous les effectifs de chaque valeur ou classe, c'est-à-dire le nombre total d'individus dans la population étudiée.
Effectif cumulé	la somme des effectifs d'une valeur et de toutes les valeurs qui la précèdent dans une série ordonnée
Moyenne	somme de toutes les valeurs d'un ensemble de données divisée par le nombre de valeurs
Médiane	mesure de tendance centrale qui représente la valeur du milieu dans un ensemble de données ordonnées
Fréquence	représente le nombre de fois que cette valeur apparaît dans un ensemble de données, divisé par l'effectif total de cet ensemble.
Quartile	une des trois valeurs qui divisent un ensemble de données triées en quatre parties égales, chaque partie contenant environ 25% des données

2- Calcul de la moyenne

Un élève obtient les notes suivantes

Matière	Note	Coefficient
Maths	12	4
Physique chimie	13	3
Français	15	2
Philosophie	15	1
SVT	16	4
Sport	11	1

$$\text{moyenne} = \frac{\sum x_i}{n}$$

Mais il ne faut pas oublier de multiplier chaque valeur par son effectif

$$M = \frac{(12\times4)+(13\times3)+(15\times2)+(15\times1)+(16\times4)+(11\times1)}{15} = 13.8$$

3- Calcul de la médiane

La médiane est une valeur d'un caractère qui correspond à un effectif cumulé de 50 % de l'effectif total

En pratique pour trouver la médiane d'une série statistique d'effectif global N :

- On ordonne les valeurs du caractère dans l'ordre croissant.
- Si N est pair, la médiane sera la moyenne des valeurs du terme de rang $\frac{N}{2}$ et du terme de rang $\frac{N}{2}+1$.
- Si N est impair, la médiane sera la valeur du terme de rang $\frac{N+1}{2}$.
- Lorsque l'effectif global est élevé, il est souvent utile de calculer les effectifs cumulés pour trouver cette valeur.

Application 1

Calculer la médiane de la série suivante

Matière	Note	Coefficient
Maths	12	4
Physique chimie	13	3
Français	15	2
Philosophie	15	1
SVT	16	4
Sport	11	1

N'oubliez jamais d'ordonner votre série

Méthode directe

Ecrire toutes les valeurs ordonnées et chercher la valeur centrale

11,12,12,12,12,13,13,13,15,15,15,16,16,16,16

Valeur centrale = médiane

Méthode générale

N=15 impair donc la médiane est la valeur de la série qui correspond au rang

$\frac{N+1}{2}$ = 8

On dresse le tableau des effectifs cumulé

	11	**12**	**13**	**15**	**16**
Effectif	1	4	3	3	4
Effectif cumulé	1	5	8	11	15

Le rang 8 est obtenu avec la valeur 13 de la série donc c'est la médiane

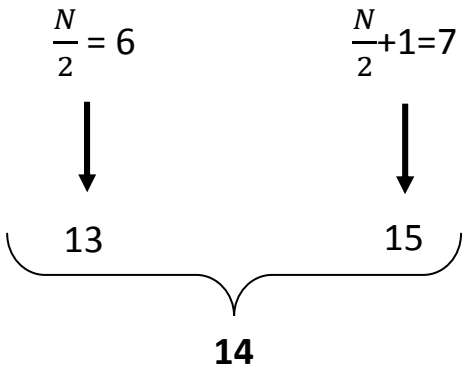
Application 2

Cas ou l'effectif total est pair

	11	**12**	**13**	**15**	**16**
Effectif	1	3	1	3	3
Effectif cumulé	2	5	6	9	12

N=12 impair donc la médiane est la moyenne des valeurs des rangs

$\frac{N}{2}$ et $\frac{N}{2}+1$



4- Les quartiles

- Le **premier quartile** Q1 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins un quart des données sont inférieures ou égales à Q1.
- Le **troisième quartile** Q3 d'une série statistique est la plus petite valeur des termes de la série pour laquelle au moins trois quarts des données sont inférieures ou égales à Q3.

L'**écart interquartile** est la différence entre le troisième et le premier quartile $Q_3 - Q_1$.

5- Variance et écart-type

La **variance** d'une série statistique est le nombre :

$$V = \frac{1}{N}\left(n_1\left(x_1 - \overline{x}\right)^2 + n_2\left(x_2 - \overline{x}\right)^2 + \ldots + n_p\left(x_p - \overline{x}\right)^2\right)$$

$$= \frac{1}{N}\sum_{k=1}^{p} n_k\left(x_k - \overline{x}\right)^2$$

L'**écart-type** est la racine carrée de la variance :

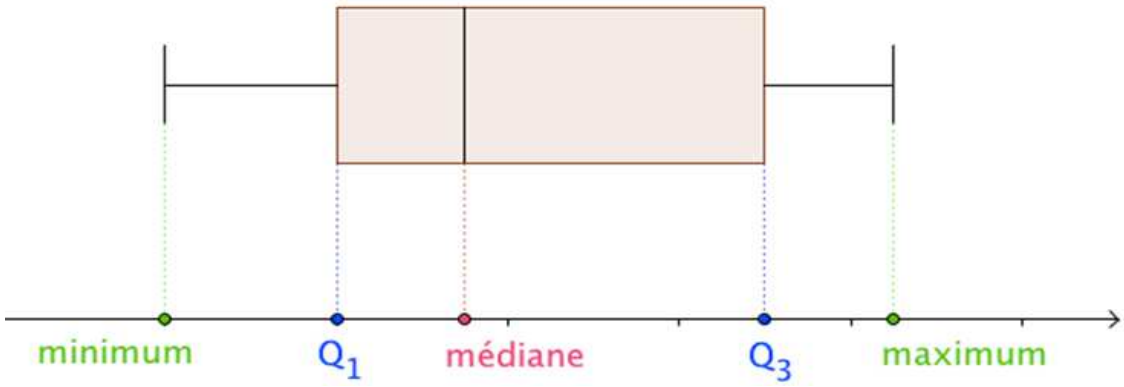
$$\sigma = \sqrt{V}$$

La **variance** d'une série statistique est égale à :

$$V = \frac{n_1 x_1^2 + n_2 x_2^2 + \ldots + n_p x_p^2}{N} - \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

$\overline{x}$: moyenne de l'échantillon

6- Diagramme en boite



Application 3

Une série statistique qui s'intéresse à la taille de 21 personnes , on trouve les valeurs suivantes

1,75	**1,68**	**1,76**	1,89	1,83	1,91	**1,78**
1,79	**1,74**	**1,67**	**1,74**	**1,70**	**1,80**	**1,75**
1,85	1,87	**1,73**	1,90	1,81	**1,79**	**1,74**

1- Quel est le caractère étudié
2- Ce caractère est-il quantitatif ou qualitatif ?
3- Quel est l'effectif et la fréquence de la valeur 1.74 ?

4- Que représente l'effectif cumulé croissant de la valeur 1.80 ?

Solution

a. Le caractère étudié est la taille des personnes.
b. Ce caractère est quantitatif.
c. L'effectif de la valeur 1,74 est égal à 3.
d. Sa fréquence est égale à $\frac{3}{21}=\frac{1}{7}$.
e. L'effectif cumulé croissant de 1,80 est égal au nombre de personnes dont la taille est inférieure ou égale à 1,80 , soit 14 (en bleu).

Application 4

Cette série statistique porte sur la taille d'un échantillon de population. Pour simplifier la lecture du tableau, on a effectué un regroupement en classes :

Taille	1,65 à 1,69	1,70 à 1,74	1,75 à 1,79	1,80 à 1,84	1,85 à 1,89	1,90 à 1,94	Total
Effectif	16	38	59	25	8	5	151

(Rappel : pour le calcul de la moyenne, on prendra pour valeur le centre de chaque classe).
Déterminer pour cette série :

Moyenne :	Étendue :	1er quartile :	Médiane :	3ème quartile :

Solution

Cette série statistique porte sur la taille d'un échantillon de population. Pour simplifier la lecture du tableau, on a effectué un regroupement en classes :

Taille	1,65 à 1,69	1,70 à 1,74	1,75 à 1,79	1,80 à 1,84	1,85 à 1,89	1,90 à 1,94	Total
Effectif	16	38	59	25	8	5	151
ECC	**16**	**54**	**113**	**138**	**146**	**151**	
Centre de classe	**1,67**	**1,72**	**1,77**	**1,82**	**1,87**	**1,92**	

Moyenne :

$$\bar{x}=\frac{1,67\times16+1,72\times38+1,77\times59+1,82\times25+1,87\times8+1,92\times5}{16+38+59+25+8+5}=\frac{266,57}{151}\approx1,77$$

Etendue : 1,94 – 1,65 = 0,29

Médiane : l'effectif n = 151 est impair : $\frac{n+1}{2}=\frac{152}{2}=76$

→ le 76ème rang est dans la classe $[1,75;1,79]$ donc la Médiane est dans la classe $[1,75;1,79]$.

1er quartile : $n\times25\%=151\times\frac{25}{100}=37,75$ →la 38ème valeur est dans la classe $[1,70;1,74]$.

3ème quartile : $n\times75\%=151\times\frac{75}{100}=113,25$ →la 114ème valeur est dans la classe $[1,80;1,84]$.

Moyenne : **1,77**	Étendue : **0,29**	1er quartile : dans la classe $[1,70;1,74]$	Médiane : dans la classe $[1,75;1,79]$	3ème quartile : dans la classe $[1,80;1,84]$

Echantillonnage

1- Définitions

Echantillon est un sous-ensemble de **n** individus extraits de la population pour lesquels on a mesuré (observé) un **caractère** quantitatif ou qualitatif (*Taille, poids, salaire, sexe, profession d'un groupe donné d'individus…)*

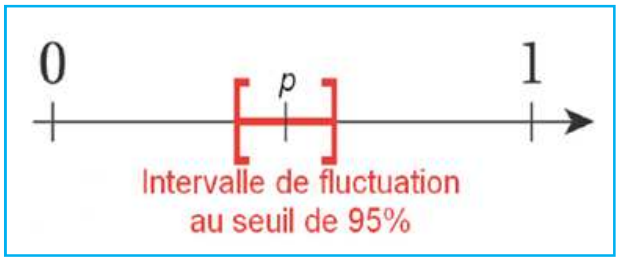
Taille d'un échantillon est le nombre d'individu de l'échantillon ou le nombre de fois qu'on répète une expérience

Fréquence théorique est la proportion d'apparition d'une valeur prise par un caractère étudié sur toute une population (probabilité)

Fréquence observée est la proportion d'apparition d'une valeur prise par un caractère étudié sur un échantillon de taille **n**

Intervalle de fluctuation est un intervalle de fréquence centré autour de *la valeur théorique* qui s'il *contient la fréquence observée* alors celle-ci est considérée comme *représentative de la population*

L'intervalle de fluctuation au seuil de 95% d'un échantillon de taille *n est* l'intervalle centré autour de la proportion théorique *p tel que la fréquence observée f se* trouve dans l'intervalle avec une probabilité égale à 0,95.



2- Principe

En réalisant l'expérience un certain nombre de fois (échantillon), on mesure la fréquence d'apparition (**fréquence observée**) d'une valeur d'un caractère.

Si cette fréquence et la valeur théorique sont trop "éloignées" (dépassent un seuil fixé qui est **l'intervalle de fluctuation**) alors on peut rejeter la valeur théorique, Dans le cas inverse, on considère que l'échantillon est représentatif de cette population.

3- Propriété

Pour ***0,2 < p < 0,8 et n > 25****, l'intervalle de fluctuation au seuil de 95% de f est l'intervalle*

$$\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$$

Cela signifie qu'on a une probabilité de 0,95 pour que la fréquence observée se trouve dans cet intervalle

Application 5

22% articles produits par une entreprise sont défectueuses. La proportion théorique *p est donc égale à 22%.*

On prélève un échantillon de taille 200 parmi cette production et on compte le nombre d'articles défectueux parmi cet échantillon.

Ce nombre est égal à 41.

L'échantillon prélevé est – il représentatif de cette population ?

Solution

la fréquence observée *f est égale à* 41/200 soit 0,205

Pour un échantillon de taille 200, l'intervalle de fluctuation de la fréquence *p* au seuil de 95 %, est un intervalle de centre 0,22 tel que les fréquences observées pour cette taille d'échantillons sont dans cette intervalle avec une probabilité de 0.95

Voyant maintenant si notre échantillon est représentatif de cette population d'articles en calculant l'intervalle de fluctuation

$$\left[0,22-\frac{1}{\sqrt{200}};0,22+\frac{1}{\sqrt{200}}\right] \qquad [0,15\ ;\ 0,29]$$

Notre fréquence observée est de 0.205 et elle est bien dans cet intervalle

Série statistique à deux variables

Etude simultanée de deux variables X et Y définies sur une même population P : mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables pour pouvoir prévoir la variation de l'une en fonction de l'autre

1- Variables liées

- **Variables liées** : les variations de l'une dépendent des variations de l'autre.
- **Variables indépendantes** : les deux variables varient indépendamment l'une de l'autre.

2- Nuage de points

On considère deux séries statistiques X et Y observées sur une même population de "n" individus.

Soient x_1, x_2, x_3,……x_n les valeurs prises par X

Soient y_1, y_2, y_3,……y_n les valeurs prises par Y

Les couples $(x_1 ; y_1), (x_2 ; y_2), \dots, (x_n ; y_n)$ forment une série statistique à deux variables.

Le nuage de points est l'espace formé par les couples (x_i,y_j)

3- Point moyen

C'est le point de coordonnées G $(\bar{x}, \bar{y})$ avec

$\bar{x}$ *et* $\bar{y}$ *les valeurs moyennes des variables X et Y*

Application 6

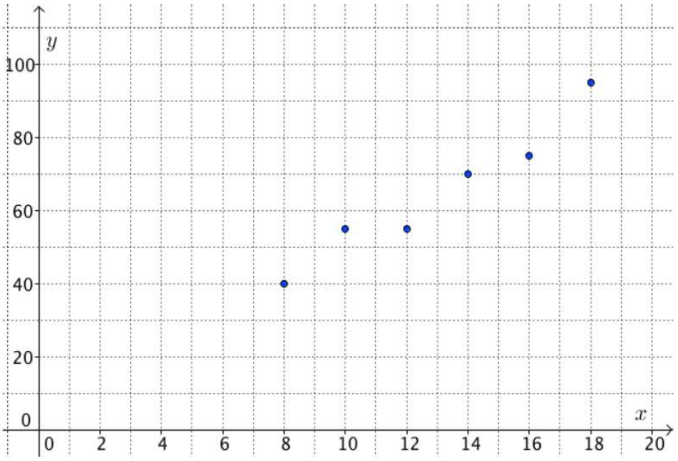
Le tableau suivant présente l'évolution du budget publicitaire et du chiffre d'affaire d'une société au cours des 6 dernières années :

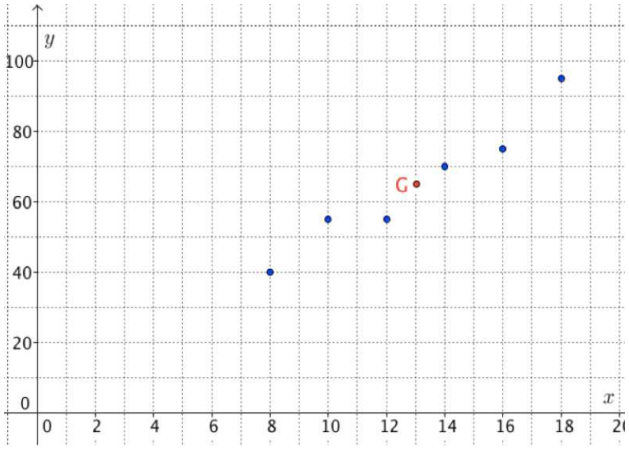
Budget publicitaire en milliers d'euros x_i	8	10	12	14	16	18
Chiffre d'affaire en milliers d'euros y_i	40	55	55	70	75	95

a) Dans un repère, représenter le nuage de points $(x_i ; y_i)$.

b) Déterminer les coordonnées du point moyen G du nuage de points.

Solution





b) $\bar{x}$ = (8 + 10 + 12 + 14 + 16 + 18) : 6 = 13

$\bar{y}$ = (40 + 55 + 55 + 70 + 75 + 95) : 6 = 65

4- Ajustement affine

Lorsque les points d'un nuage sont sensiblement alignés, on peut construire une droite, appelé **droite d'ajustement (ou droite de régression)**, passant « au plus près » de ces points.

Méthode de Mayer

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage et adopter cette droite comme ajustement affine.

NB: pour obtenir deux points moyens G_1 et G_2 on choisit deux groupes de 50 % des points de chaque série

Application 7

Reprenez l'exemple précédent:

a) Calculer les coordonnées de G_1 et G_2.

b) On prend (G_1G_2) comme droite d'ajustement. Tracer cette droite.

2) À l'aide du graphique :

a) Estimer le chiffre d'affaire à prévoir pour un budget publicitaire de 22 000 €.

b) Estimer le budget publicitaire qu'il faudrait prévoir pour obtenir un chiffre d'affaire de 100 000 €.

c) La méthode utilisée dans les questions 2a et 2b consiste-t-elle en une interpolation ou une extrapolation ?

Solution

1) a) $\overline{x_1} = (8 + 10 + 12) : 3 = 10$

$\overline{y_1} = (40 + 55 + 55) : 3 = 50.$

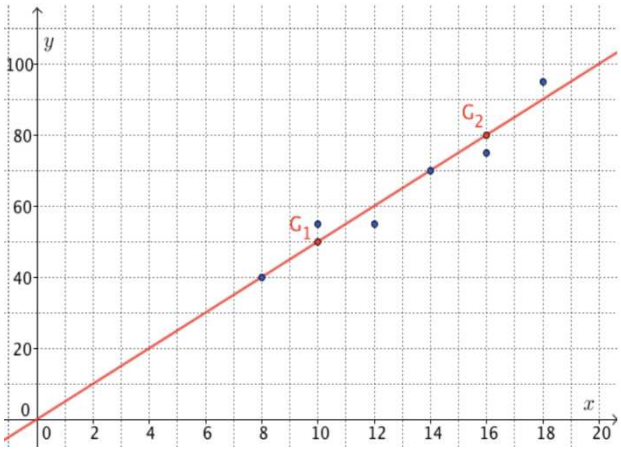
Le point moyen G_1 a pour coordonnées (10 ; 50).

$\overline{x_2} = (14 + 16 + 18) : 3 = 16$

$\overline{y_2} = (70 + 75 + 95) : 3 = 80.$

Le point moyen G_2 a pour coordonnées (16 ; 80).

b)



2) On lit graphiquement :

a) Le chiffre d'affaire à prévoir pour un budget publicitaire de 22 000 € est de 110 000 €.

b) Le budget publicitaire qu'il faudrait prévoir pour obtenir un chiffre d'affaire de 100 000 € est de 20 000€.

c) Les lectures graphiques sont réalisées ici en dehors du domaine d'étude, on parle donc d'extrapolation.

Méthode des moindres carrés

De nombreuses séries statistiques (x_i,y_i) sont reliées par des conditions du type $y=ax+b$.
En général, en raison des erreurs de mesure, les points (xi,yi) ne sont pas alignés, mais sont "presque" sur une même droite. Il faut alors choisir a et b de sorte que la droite soit la meilleure possible.

Il faut choisir une mesure de l'écart entre une droite $y=ax+b$ et le nuage de points expérimentaux $(xi,yi)1\leq i\leq n$.

On choisit en général le carré de la différence entre le point théorique et le point expérimental, c'est-à-dire $(yi-(axi+b))^2$. L'écart total est donc :

$$J(a,b)=\sum_{i=1}^{n}(y_i-ax_i-b)^2.$$

Effectuer une régression linéaire au sens des moindres carrés, c'est trouver la droite qui minimise l'écart précédent, c'est-à-dire la somme des carrés des différences

Théorème : Si la variance $Var(X)$ de la série statistique $X=(x_i)$ est non-nulle, il existe une unique droite qui minimise la quantité $J(a,b)$. Elle vérifie

$$a=\frac{Cov(X,Y)}{Var(X)} \text{ et } b=\bar{Y}-a\bar{X},$$

où $Cov(X,Y)$ désigne la covariance de X et de Y, $\bar{X}$ la moyenne de (x_i) et $\bar{Y}$ la moyenne de (y_i).

$$cov(x;y)=\frac{1}{n}\left((x_1-\bar{x})(y_1-\bar{y})+\cdots+(x_n-\bar{x})(y_n-\bar{y})\right)$$

$$var(x)=\frac{1}{n}((x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2)$$

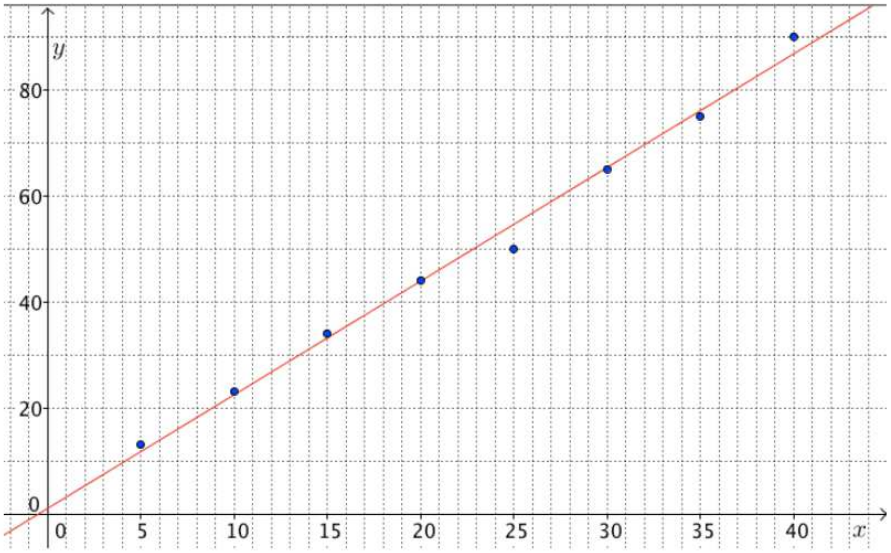
Application 8

On considère la série statistique à deux variables données dans le tableau suivant :

x_i	5	10	15	20	25	30	35	40
y_i	13	23	34	44	50	65	75	90

1) Dans un repère, représenter le nuage de points $(x_i\ ;\ y_i)$.
2) a) Déterminer une équation de la droite d'ajustement par la méthode des moindres carrés.
b) Vérifier à l'aide de la calculatrice.
b) Représenter la droite d'ajustement de y en x.
3) Estimer graphiquement la valeur de x pour $y = 70$. Retrouver ce résultat par calcul. S'agit-il d'une interpolation ou d'une extrapolation ?

Solution



2) a) On commence par calculer, les moyennes $\bar{x}$ et $\bar{y}$:

$$\bar{x} = \frac{5 + 10 + \cdots + 40}{8} = 22{,}5$$

$$\bar{y} = \frac{13 + 23 + \cdots + 90}{8} = 49{,}25$$

Par la méthode des moindres carrés, la droite d'ajustement de y en x a pour équation $y = ax + b$ avec :

$$a = \frac{cov(x;y)}{var(x)}$$

$$= \frac{\frac{1}{8}\left((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_8 - \bar{x})(y_8 - \bar{y})\right)}{\frac{1}{8}\left((x_1 - \bar{x})^2 + \cdots + (x_8 - \bar{x})^2\right)}$$

$$= \frac{(5 - 22{,}5)(13 - 49{,}25) + (10 - 22{,}5)(23 - 49{,}25) + \cdots + (40 - 22{,}5)(90 - 49{,}25)}{(5 - 22{,}5)^2 + (10 - 22{,}5)^2 + \cdots + (40 - 22{,}5)^2}$$

$$\approx 2.138$$

Et $b = \bar{y} - a\bar{x} \approx 49{,}25 - 2{,}138 \times 22{,}5 = 1{,}145$

Une équation de la droite d'ajustement est : $y = 2{,}138x + 1{,}145$
Pour le tracé, on considère l'équation : $y = 2{,}1x + 1{,}1$

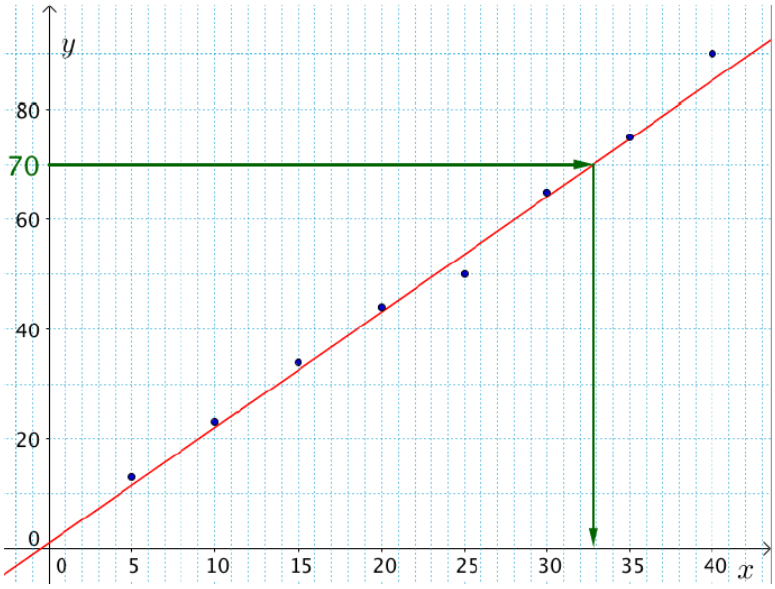
b) Avec TI :
- Appuyer sur « **STAT** » puis « **Edite** » et saisir les valeurs de x_i dans L1 et les valeurs de y_i dans L2.
- Appuyer à nouveau sur « **STAT** » puis « **CALC** » et « **RegLin(ax+b)** ».
- Saisir **L1,L2**

Avec CASIO :
- Aller dans le menu « **STAT** ».
- Saisir les valeurs de x_i dans *List1* et les valeurs de y_i dans *List2*.

- Sélectionner « **CALC** » puis « **SET** ».
- Choisir *List1* pour *2Var XList* et *List2* pour *2Var YList* puis « **EXE** ».
- Sélectionner « **REG** » puis « **X** » et « **aX+b** ».

La calculatrice nous renvoie : $a = 2.138095238$ et $b = 1.142857143$
Une équation de la droite d'ajustement est : $y = 2{,}1x + 1{,}1$

Pour tracer la droite, il suffit de calculer les coordonnées de deux points de la droite d'ajustement :
- Si $x = 0$ alors $y = 2{,}1 \times 0 + 1{,}1 = 1{,}1$ donc le point de coordonnées $(0\,;1{,}1)$ appartient à la droite d'ajustement.
- Si $x = 10$ alors $y = 2{,}1 \times 10 + 1{,}1 = 22{,}1$ donc le point de coordonnées $(10\,;22{,}1)$ appartient à la droite d'ajustement.



3) - Pour $y = 70$, on lit graphiquement $x \approx 33$.

- Par calcul, si $y = 70$, alors $70 = 2{,}1x + 1{,}1$

Soit $2{,}1x = 70 - 2{,}1$

$2{,}1x = 68{,}9$

$x = \dfrac{68{,}9}{2{,}1} \approx 32{,}8$

- Les calculs sont réalisés dans domaine d'étude, on parle donc d'interpolation.

5- Coefficient de corrélation

Définition : Le **coefficient de corrélation** de x et y est donné par :

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

Interprétation :

Le coefficient de corrélation ρ_{xy} est un nombre compris entre -1 et 1 qui mesure la relation entre les deux variables x et y. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte.

- Si $\rho_{xy} > 0$, les valeurs prises par y ont tendance à croître quand les valeurs de x augmentent.
- Si $\rho_{xy} < 0$, les valeurs prises par y ont tendance à décroître quand les valeurs de x augmentent.
- Si $\rho_{xy} = 0$, les variations des variables x et y sont indépendantes.

Application 9

Calculer le coefficient de corrélation de l'exemple précédent et interpréter le résultat

Solution

$$cov(x,y) = \frac{1}{8}\big((5-22{,}5)(13-49{,}25) + \cdots + (40-22{,}5)(90-49{,}25)\big) \approx 280{,}625$$

$$var(x) = \frac{1}{8}((5-22{,}5)^2 + \cdots + (40-22{,}5)^2) \approx 131{,}25$$

$$var(y) = \frac{1}{8}((13-49{,}25)^2 + \cdots + (90-49{,}25)^2) \approx 604{,}4375$$

Soit :

$$\rho_{xy} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} \approx \frac{280{,}625}{\sqrt{131{,}25 \times 604{,}4375}} \approx 0{,}996$$

Le coefficient de corrélation est proche de 1 donc la corrélation entre les deux variables est forte. Les points du nuage sont proches de la droite d'ajustement.