Ch. V (suite) : STATISTIQUES À DEUX VARIABLES

	Contenu	Extensions / inclusions
6.7.	Diagrammes de dispersion ; droite de régression trouvée visuellement, passant par le point moyen. Données à deux variables : le concept de corrélation. Le coefficient de corrélation de Pearson : utilisation de la formule $r = \frac{s_{xy}}{s_x \cdot s_y}$ Interprétation des corrélations positives, nulles et négatives.	Dans les épreuves écrites : la valeur de s_{xy} sera donnée, si nécessaire. s_x représente l'écart-type de la variable X ; s_{xy} représente la covariance des variables X et Y. Une calculatrice à écran graphique peut être utilisée pour calculer r lorsque des données brutes sont présentées.
6.8.	La droite de régression pour y en fonction de x : utilisation de la formule $y - \overline{y} = \frac{S_{xy}}{S_x^2} \cdot (x - \overline{x})$ Utilisation de la droite de régression pour faire des prédictions.	On s'attend à une compréhension des valeurs aberrantes de la part des élèves. Les élèves doivent être conscients que la droite de régression est moins fiable lorsqu'on l'étend bien au-delà de la région occupée par les données. Une calculatrice à écran graphique peut être utilisée pour trouver l'équation de la droite de régression lorsque des données brutes sont présentées.
6.9.	Le test d'indépendance du χ^2 ; formulation de l'hypothèse nulle et de l'hypothèse alternative ; seuils de signification ; tableaux de contingence ; fréquences théoriques ; utilisation de la formule $\chi_{calc}^2 = \sum \frac{(f_o - f_e)^2}{f_e}$; degrés de liberté ; utilisation des tables pour trouver les valeurs critiques ; valeurs de p.	Inclus : tableaux de contingence h par k avec h, k ≤ 4. Dans les épreuves écrites : des questions faisant intervenir les seuils de signification usuels (1 %, 5 %, 10 %) seront posées. La calculatrice à écran graphique peut être utilisée pour trouver la valeur du χ^2 lorsque des données brutes sont présentées. Non exigé : la correction de Yates. Les valeurs de p seront utilisées dans les cas des tests unilatéraux à gauche et à droite, mais pas dans les cas des tests bilatéraux.

Formulaire :

6.7.	Coefficient de corrélation de Pearson	$r = \frac{s_{xy}}{s_x s_y}$, avec $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$, $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}}$ et s_{xy} la covariance
6.8.	Équation de la droite de régression pour y en fonction de x	$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$
6.9.	Le test statistique du χ^2	$\chi_{calc}^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, avec f_o les fréquences observées et f_e les fréquences estimées

4. Séries statistiques à deux variables

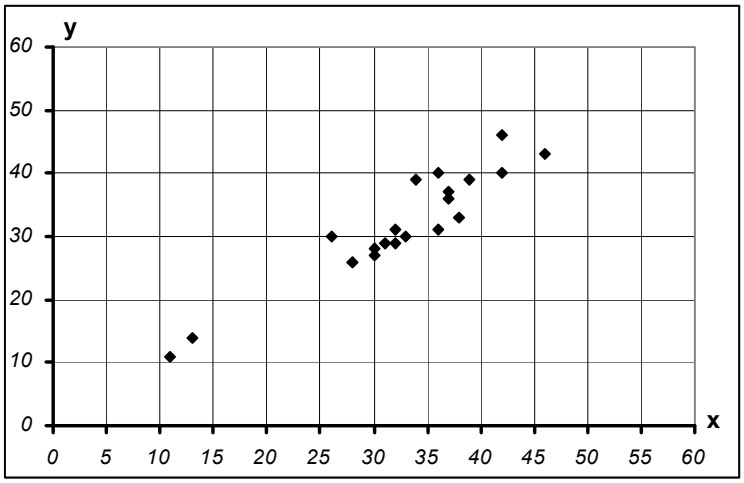
4.1. Position du problème

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.
P.ex. on peut s'intéresser simultanément

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

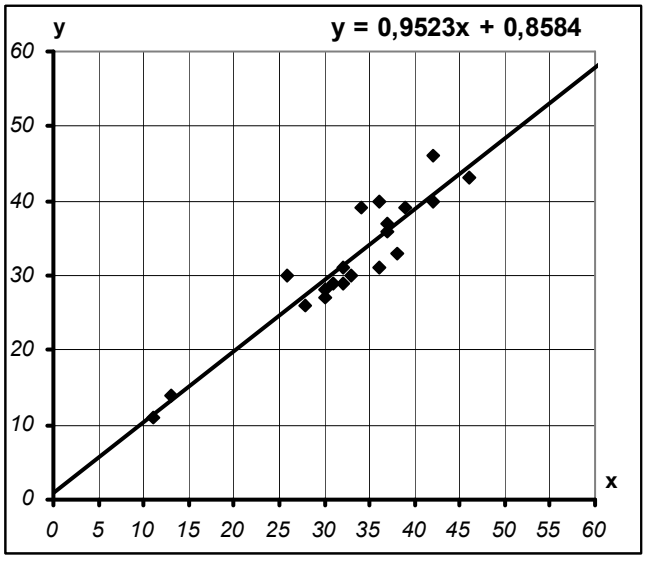
Prenons comme 1er exemple les notes obtenues en maths par les 20 élèves d'une classe de 6e au 1er trimestre, puis à la fin de l'année. Notons x_i la note du 1er trimestre obtenue par l'élève i, y_i la note à la fin de l'année obtenue par le même élève.

Élève	x	y
1	28	26
2	36	40
3	13	14
4	42	40
5	32	29
6	11	11
7	37	36
8	46	43
9	38	33
10	32	31
11	42	46
12	37	37
13	34	39
14	31	29
15	30	28
16	26	30
17	36	31
18	30	27
19	33	30
20	39	39



Dans un repère nous construisons tous les points M_i de coordonnée (x_i, y_i) .
L'ensemble de tous les points ainsi obtenus est appelé un ***nuage de points***. On parle aussi d'un ***diagramme de dispersion***. Si les points obtenus sont totalement dispersés, on dira que les variables x et y sont indépendantes.

Pour cet exemple, on a cependant l'impression que les variables X et Y sont liées dans le sens que, si X devient plus grand, il y a une tendance à ce que Y devienne également plus grand. On parle d'une ***corrélation positive*** entre les variables X et Y.



Un des objectifs de ce chapitre est de trouver une mesure pour cette corrélation, qui peut être plus ou moins forte, et qui peut également être négative. Si on regarde le nuage de points de plus près, on peut penser que les points sont pour la plupart disposés autour d'une certaine droite. Un autre objectif de ce chapitre sera de trouver la droite qui ajuste le mieux l'ensemble des données (x_i, y_i). On parle de ***l'ajustement linéaire***, la droite en question est appelée ***droite de régression linéaire.*** Cet ajustement graphique peut paraître un peu arbitraire. Il faudra préciser par le calcul une méthode qui permet de trouver la « meilleure » droite.

L'outil « trendline » du logiciel Excel fournit une droite de régression linéaire ainsi que son équation réduite de la forme $Y = mX + c$, ici : **$Y = 0{,}9523X + 0{,}8584$**.

A l'aide de cette équation, le professeur de maths peut essayer de prédire une note finale Y pour un élève à partir de la note X obtenue au 1er trimestre (peut-être l'année d'après dans une autre classe):

P.ex. pour $X = 30$: $Y = 0{,}9523 \cdot 30 + 0{,}8584 = 29{,}42$

$X = 45$: $Y = 0{,}9523 \cdot 45 + 0{,}8584 = 43{,}71$

Bien sûr l'équation de la droite de régression ne traduit qu'une tendance générale (et encore seulement pour l'échantillon considéré) ; pour un élève en particulier, le résultat en fin d'année peut ne pas du tout correspondre à la valeur attendue selon cette méthode statistique.

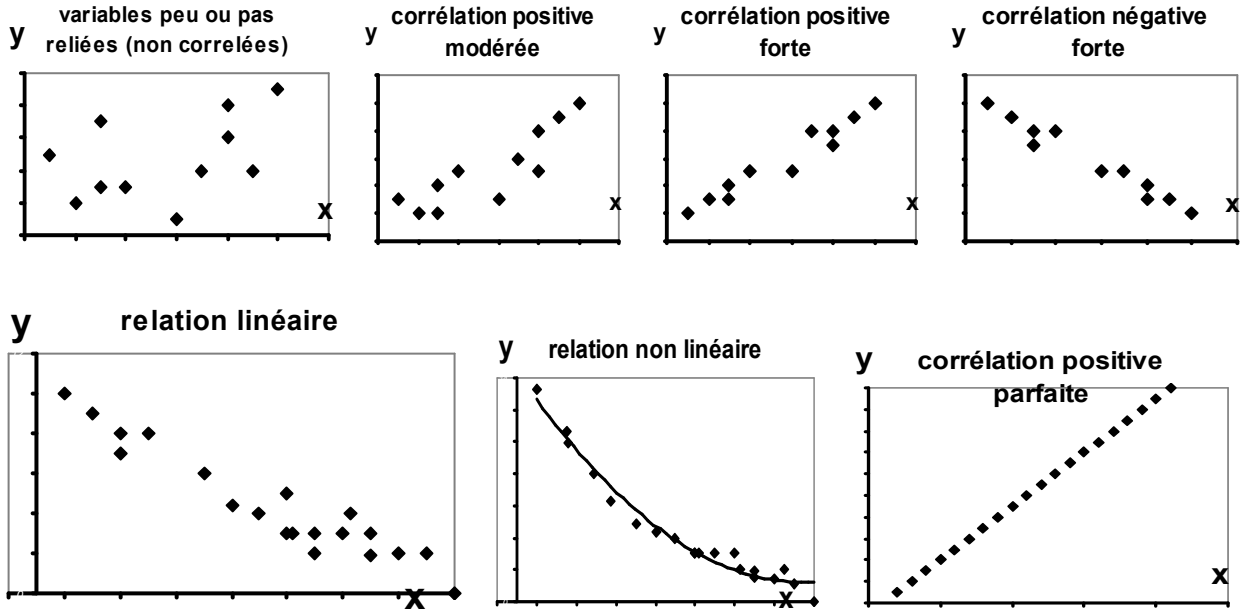
4.2. Diagrammes de dispersion . Notion de corrélation.

Sur une population donnée, on étudie deux caractères.
Pour chacun des n individus de cette population, notons x_i et y_i les valeurs prises par chacun de ces caractères, et présentons les données à l'aide de la série statistique à deux variable suivante :

Valeur x_i	x_1	x_2	…	x_n
Valeur y_i	y_1	y_2	…	y_n

Définition : Dans un repère orthogonal, l'ensemble des points M_i de coordonnées $(x_i ; y_i)$ (avec $1 \leq i \leq n$) est appelé le **nuage de points ou diagramme de dispersion** associé à cette série statistique à deux variables.

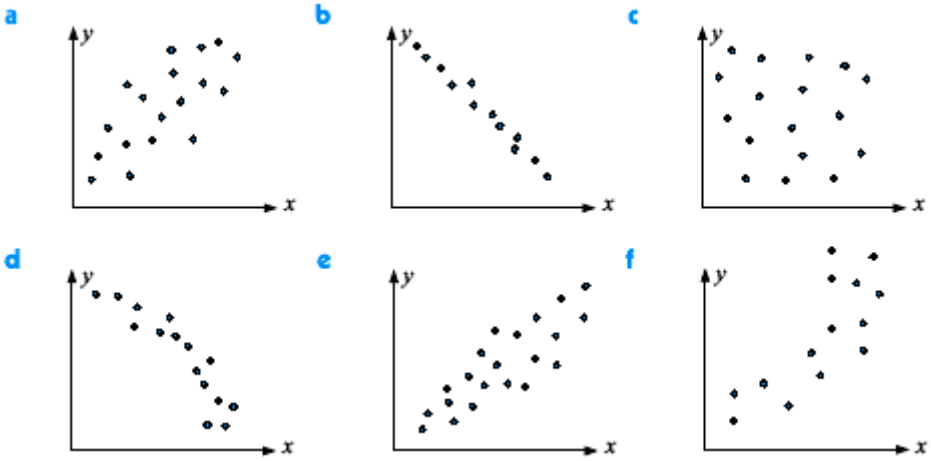
Le nuage de points peut prendre des allures différentes et traduire une relation plus ou moins importante entre les variables x et y. Il peut avoir une forme allongée, curviligne ou très dispersée.



Entre les variables x et y il existe une **relation linéaire** si les données peuvent être ajustées par une droite. Souvent, les données sont mieux ajustées par une autre courbe, par exemple une parabole ou ou courbe exponentielle. Notre étude va cependant se limiter à l'ajustement linéaire.

Exercice 768: Dites pour chacun des nuages de points

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



Exercice 769: On a regroupé les résultats pour des tests en math et en sciences pour certains étudiants :

étudiant	A	B	C	D	E	F	G	H	I	J
Math	64	67	69	70	73	74	77	82	84	85
Sciences	68	73	68	75	78	73	77	84	86	89

a) Donnez le diagramme de dispersion.
b) Décrivez la direction, la forme et le degré (l'ampleur) de la relation entre les résultats en science et en math.

Exercice 770: Le nombre des accidents de travail dans une certaine entreprise a évolué de la manière suivante entre 1994 et 2003 :

année	1994	1995	1996	1997	1998	1999	2000	2001	2002	2002
Nombre des accidents	166	131	123	162	160	130	91	82	65	53

a. Commentez le travail réalisé par la direction de l'entreprise au niveau de la sécurité.
b. Donnez le nuage de points et utilisez-le pour commenter les données.

4.3. Ajustement linéaire

Lorsque les points du nuage paraissent presque alignés, on peut chercher une relation de la forme $y = mx + c$ qui exprime de façon approchée y en fonction de x, autrement dit, une fonction affine f telle que l'égalité $y = f(x)$ s'ajuste au mieux avec les données. Graphiquement, cela signifie qu'on cherche **une droite qui passe au plus près de tous les points du nuage**.
Une telle relation permettrait notamment de faire des **prévisions**.

4.3.a. Premère approche : à main libre et avec le point moyen

Définition : Le point G de coordonnées $(\overline{x}\ ;\ \overline{y})$, avec

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \overline{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

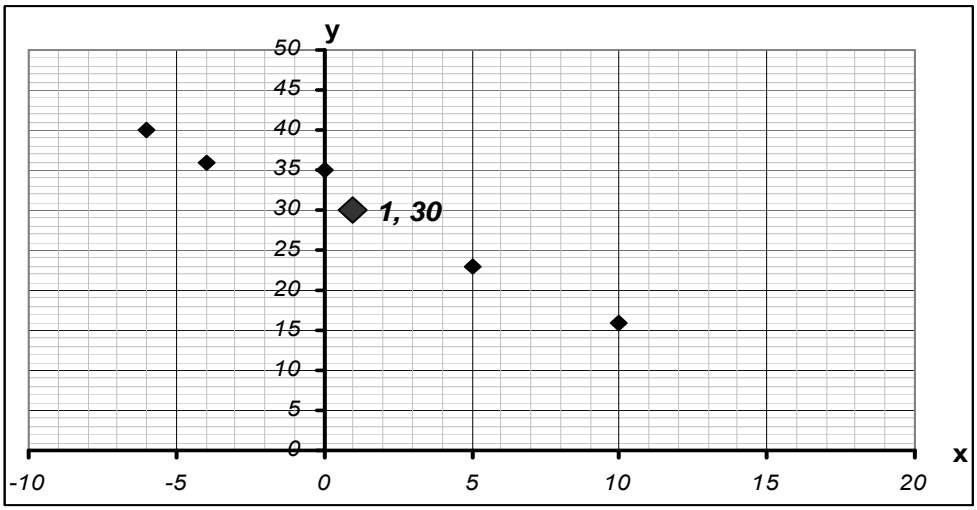
est appelé le **point moyen** du nuage de points associé à cette série statistique à deux variables.
G est donc le point qui a pour abscisse la moyenne arithmétique des abscisses, et pour ordonnée la moyenne arithmétique des ordonnées des points du nuage.

Dans une première approche, on peut essayer de tracer à main libre une droite qui ajuste bien les données ; on exige cependant que la droite doit passer par le point moyen G$(\overline{x}\ ;\ \overline{y})$, qui est en quelque sorte le « centre de gravité » du nuage de points.

Exemple :
Le tableau ci-dessous donne la consommation quotidienne Y en fuel d'une chaudière (en litres) en fonction des relevés de température extérieure X.

x (en degrés C)	-6	-4	0	5	10
y (en litres)	40	36	35	23	16

On cherche un lien (s'il existe) entre la température extérieure x et la consommation quotidienne de fuel y.



Point moyen d'un nuage : G($\overline{x}$; $\overline{y}$) avec $\overline{x} = 1/5 \cdot (x_1 + \ldots + x_5) = 1$ et $\overline{y} = 1/5 \cdot (y_1 + \ldots + y_5) = 30$

$\Rightarrow$ **G(1 ;30)**

Après avoir tracé par le point G une droite qui vous semble être la meilleure :

- Donnez une équation de cette droite
- Donnez une estimation pour la consommation de fuel par jours si la température extérieure est
  - -10° C
  - 15° C

4.3.b. La méthode de Mayer ou la méthode des moyennes discontinues

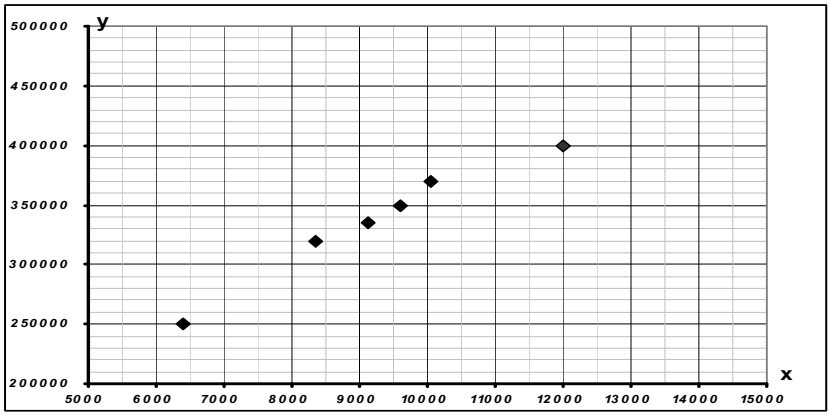
(ne figure pas au programme)

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 divise la série statistique en deux groupes de même importance. Par exemple, on met dans le premier groupe la première moitié de l'effectif (aprés avoir ordonné les valeurs de x), et dans le deuxième groupe la seconde moitié. Pour un nombre impair de données, on prend un point (du milieu) dans les deux groupes. Pour chaque groupe, on détermine le point moyen. La droite qui passe par les deux points moyens est choisie comme droite d'ajustement.

Exemple : Le tableau suivant donne le chiffre d'affaire réalisé au cours des 6 derniers mois par un site de vente en ligne en fonction du nombre de commandes reçues.

nombre de commandes x_i	6 400	8 350	9 125	9 600	10 050	12 000
chiffre d'affaire mensuel y_i (€)	250 000	320 000	335 000	350 000	370 000	400 000

Représentation du nuage de points



Calcul des coordonnées des points moyens G_1 et G_2 :

On partage le nuage de points en deux groupes de même importance suivant les valeurs croissantes de x_i, et on calcule les coordonnées des points moyens G_1 et G_2 de chaque groupe de points.

➢ Coordonnées de G_1 $(\bar{x}_1 ; \bar{y}_1)$ avec $\bar{x}_1$ = moyenne des valeurs x du premier groupe et $\bar{y}_1$ = moyenne des valeurs y du premier groupe.

$$\bar{x}_1 = \frac{6400 + 8350 + 9125}{3} \approx 7\,960 \qquad \bar{y}_1 = \frac{250000 + 320000 + 335000}{3} \approx 310\,650$$

Donc ***G_1 (7 960 ; 310 650)***

➢ Coordonnées de G_2 $(\bar{x}_2 ; \bar{y}_2)$ avec $\bar{x}_2$ = moyenne des valeurs x du deuxième groupe et $\bar{y}_2$ = moyenne des valeurs y du deuxième groupe.

$$\bar{x}_2 = \frac{9600 + 10050 + 12000}{3} \approx 10\,550 \qquad \bar{y}_2 = \frac{350000 + 370000 + 400000}{3} \approx 373\,330$$

Donc ***G_2 (10 550 ; 373 000)***

On trace la droite d'ajustement qui passe par les deux points G_1 et G_2.

Equation de la droite d'ajustement affine :

L'équation de la droite d'ajustement est de la forme $y = m\,x + c$

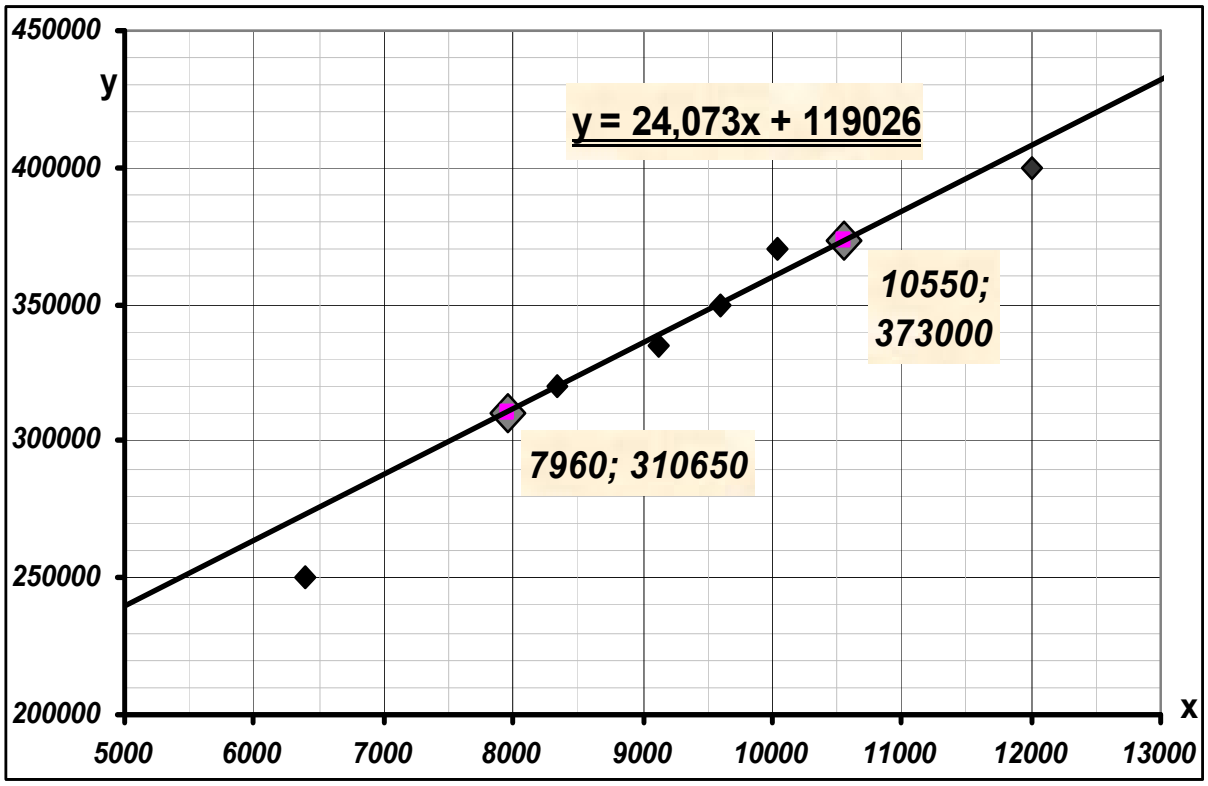
Rappel : toute droite passant par les points $A\ (x_A ; y_A)$ et $B\ (x_B ; y_B)$ a pour équation

$$y = y - y_A = m \cdot (x - x_A) \quad \text{avec } m = \frac{y_B - y_A}{x_B - x_A}$$

Dans le cas précédent, la droite passant par les points G_1 (7 960 ; 310 650) et G_2 (10 550 ; 373 000), a pour coefficient directeur $m = \frac{373000 - 310650}{10550 - 7960} \approx 24{,}073$

La droite d'ajustement affine a donc pour équation :

$$y - 310650 = 24{,}073 \cdot (x - 7960) \Rightarrow \boxed{y = 24{,}073\,x + 119\,026}$$



V Statistiques

Exercice 771: Dans le tableau ci-dessous, on donne la taille moyenne (en cm) des nouveaux nés en fonction du nombre de l'âge gestationnel (en semaines). Données 1990

Âge gestationnel (semaines)	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
Taille (cm)	47,5	48,5	49	49,7	50	50,5	50,8	51,2	51,5	51,8	52,2	52,5	52,8	53	53,5	53,7

a. Représenter le nuage de points dans un repère orthogonal en prenant comme unités :
   - en abscisse : 1 cm pour 1 semaine (commencer la graduation à 20 semaines)
   - en ordonnée : 2 cm par unité (commencer la graduation à 45 cm)

b. On se propose de tracer la droite d'ajustement de ce nuage de points.
   - Calculer les coordonnées des points moyens G_1 et G_2
   - Tracer la droite d'ajustement passant par les points G_1 et G_2.

c. Déterminer l'équation de la droite d'ajustement.

Exercice 772: Dans le tableau ci-dessous, on donne la pluviométrie moyenne mensuelle sur le département de la Meuse au cours des 30 dernières années.

Mois	Janv	Fev	Mar	Avr	Mai	Juin	Juil	Aoû	Sept	Oct	Nov	Dec
Pluviométrie (mm)	102	82	85	69	75	82	81	68	80	97	97	124

1. Représenter le nuage de points dans un repère orthogonal en prenant comme unités :
   - en abscisse : 1 cm pour un mois (numéroter les mois de 1 à 12).
   - en ordonnée : 1 cm pour 10 mm de pluie.
2. On se propose de tracer la droite d'ajustement de ce nuage de points.
   - Calculer les coordonnées des points moyens G_1 et G_2 correspondant respectivement au premier et au second semestre.
   - Tracer la droite d'ajustement passant par les points G_1 et G_2.
3. Déterminer l'équation de la droite d'ajustement.
4. Commentez les résultats trouvés.

Exercice 773:

En prévision du lancement d'un nouveau produit, une société a effectué une enquête auprès de clients éventuels pour fixer le prix de vente de ce produit. Les résultats sont donnés dans le tableau ci-dessous :

Prix x_i de vente en euros	9	10	11	12	14	15	16	17
Nombre y_i d'acheteurs éventuels	180	160	150	130	100	90	80	70

1°) Représenter graphiquement le nuage de points $M_i(x_i ; y_i)$
unité 1cm sur l'axe des abscisses et 1cm pour 10 unités sur l'axe des ordonnées.

2°) a) Calculer les coordonnées du point moyen G_1 des quatre premiers points, puis les coordonnées du point moyen G_2 des quatre derniers points.
Placer ces points sur le graphique et tracer la droite (G_1G_2).

b) On admet que la droite (G_1G_2) est une droite d'ajustement du nuage de points.
Estimer graphiquement le prix maximum pour qu'il y ait au moins 50 acheteurs potentiels.

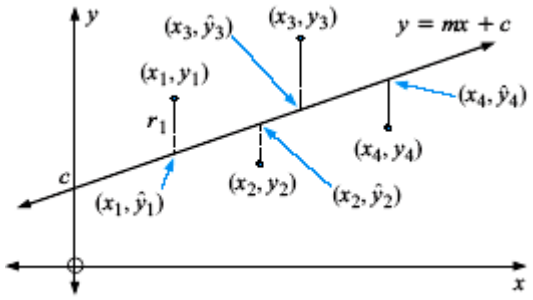
3°) a) Justifier qu'une équation de la droite (G_1G_2) est $y = -14x + 302$

b) En déduire :
- le nombre d'acheteurs que l'on peut prévoir si le prix de vente est fixé à 13 euros.
- le prix de vente pour que le nombre d'acheteurs potentiels soit supérieur ou égal à 250.

4.3.c. La méthode des moindres carrés – droite de régression

Imaginons que deux élèves aient tracé, à tâtons, des droites d'ajustement qui passent par le point moyen G. N'y en a-t-il pas une meilleure que l'autre ? Comment "mesurer" la qualité de l'ajustement ? Une méthode consiste à considérer la somme des résidus associée à une droite d'équation y = mx+c.

Pour mesurer la qualité de la droite d'ajustement d'équation y = mx + c, on considère, pour chaque valeur x_i, la différence entre la valeur observée, c'est à dire y_i, et la valeur calculée par la formule, c'est à dire $\hat{y}_i = mx_i + c$. On souhaite que la somme **des carrés** de toutes les différences : $y_i - \hat{y}_i$ appelées **erreurs**, ou **résidus**, ou **écarts verticaux**, ou **perturbations**, soit la plus petite possible.



Cette méthode, qui est la plus couramment employée, dite **méthode des moindres carrés**, consiste à choisir *m* et *c* de façon que la **somme des carrés des résidus soit la plus petite possible**.

On peut montrer qu'il existe une droite unique qui rend minimale la somme des carrés des résidus. Cette droite est appelée **droite de régression de y par rapport à x**.
Elle passe toujours par le point moyen $G(\overline{x}, \overline{y})$.

Théorème (admis):

Il existe une droite unique associée au nuage de points $(x_i ; y_i)$, avec $i = 1, 2, \ldots, n$, telle que la somme S des carrés des résidus soit minimale.

- **Cette droite, appelée droite de régression de y par rapport à x, passe par le point moyen G($\overline{x}$, $\overline{y}$) du nuage.**
- **Elle a pour équation** $y - \overline{y} = m \cdot (x - \overline{x})$ **avec** $m = \dfrac{S_{xy}}{S_x^2}$

Pour calculer la pente m de la droite :

S_x^2 **est la variance de la variable x :** $S_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$

S_{xy} **est appelée la covariance de *x* et *y* :** $S_{xy} = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y}$.

La seconde expression est plus commode pour les calculs à la main.

Exemple: Utilisez les formules pour calculer m et c pour la droite de régression passant par les points (1,3), (3,5) et (5,6). Vérifiez vos résultats
a) à l'aide de la TI 84 b) avec Excel

	x	y	xy	x^2
	1	3	3	1
	3	5	15	9
	5	6	30	25
$\sum$	9	14	48	35

$\sum x = 9,\ \sum y = 14,\ \sum xy = 48,$
$\sum x^2 = 35,\ n = 3$

$\overline{x} = \dfrac{\sum x}{n} = \dfrac{9}{3} = 3$ $\quad \overline{y} = \dfrac{\sum y}{n} = \dfrac{14}{3}$

$S_{xy} = \left(\dfrac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y} = 1/3 \bullet 48 - 3 \bullet 14/3 = 2$

$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2 = 1/3 \bullet 35 - 3^2 = 8/3$

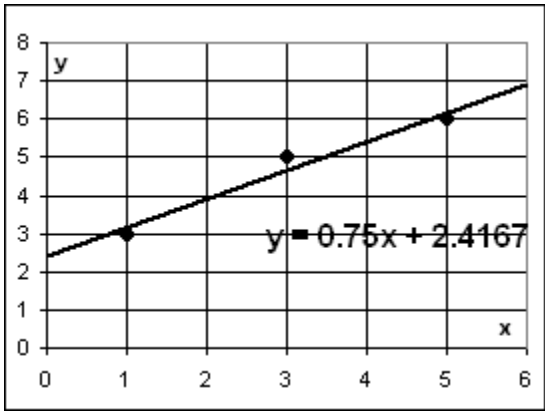
d'où : $m = \frac{S_{xy}}{S_x^2} = \frac{2}{8/3} = \frac{6}{8} = \frac{3}{4}$

et l'équation de la droite devient :

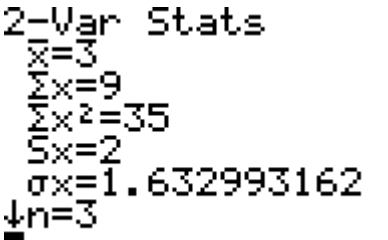
$y - \frac{14}{3} = \frac{3}{4}\cdot(x-3) \Leftrightarrow y = \frac{3}{4}x + \frac{29}{12}$

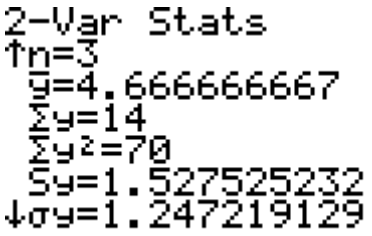
Avec Excel :

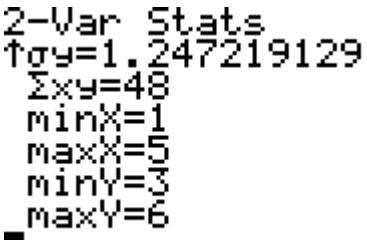
x	y
1	3
3	5
5	6

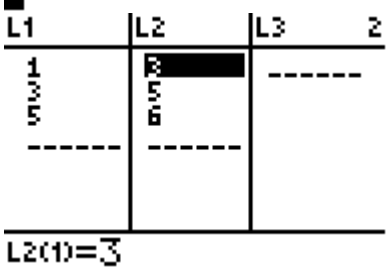


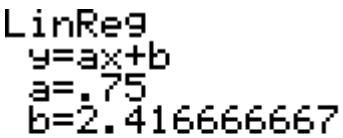
Avec la TI 84 :

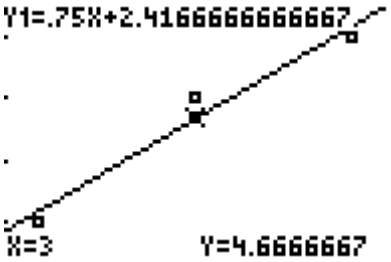




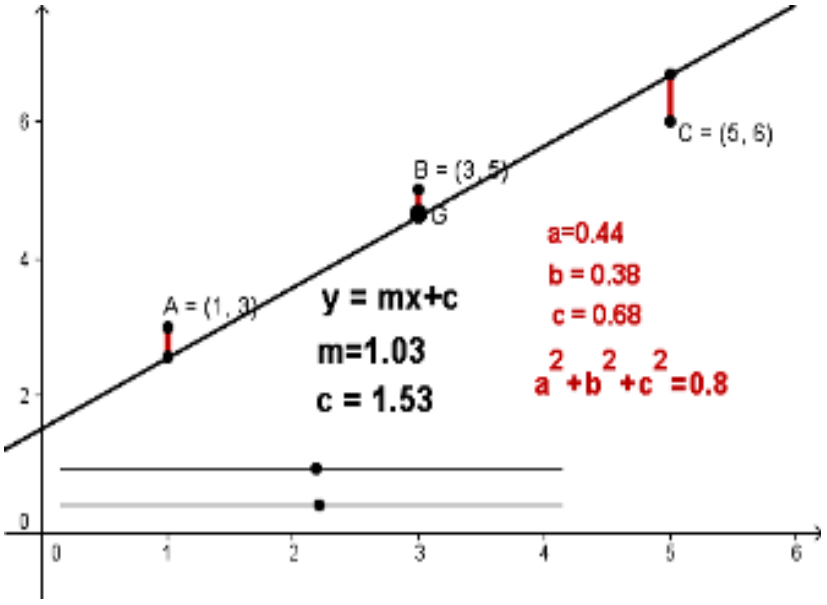




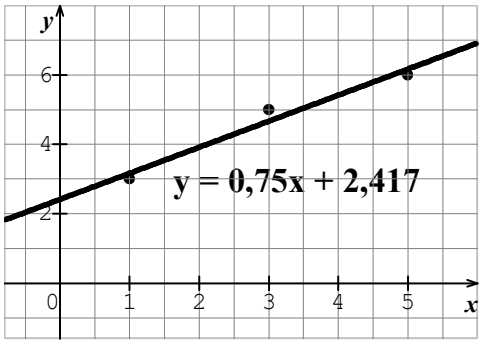




Avec Geogebra



Avec Sinequanon :



Exercice 774: Le tableau suivant donne l'évolution du prix d'un paquet de café en francs au 31 décembre de l'année 1900 + x.

Rang x_i de l'année	70	80	88	94	96	98	99	100
Prix yi en francs	3	5,5	10	15,5	19,3	19,4	20	21

a. Représenter le nuage de points associé à cette série statistique $(x_i; y_i)$.
b. Déterminer par la méthode des moindres carrés une équation de la droite d'ajustement affine de y en x.
c. Tracer cette droite sur le graphique.
d. En supposant que ce modèle mathématique reste valable jusqu'à l'an 2002, donner une estimation du prix, en euros, arrondi au centime, d'un paquet de café au 31/12/2002. On rappelle qu'un euro vaut 6,55957 francs.

Exercice 775: Le tableau suivant donne les pressions sanguines (pression systolique) mesurées auprès de huit femmes et leurs âges respectifs :

Âge (x)	60	42	68	72	42	36	55	49
Pression sanguine (y)	155	140	152	160	125	118	155	145

a. Représentez les données par un nuage de points
b. Déterminez la droite de régression de y par rapport à x et ajoutez-là sur le diagramme.
c. Utilisez la droite d'ajustement pour déterminer la pression sanguine pour une femme
   i. âgée de 45 ans ii. âgée de 85 ans.
d. Quelle est la différence dans la façon d'utiliser la droite d'ajustement en i) et ii) ?

Exercice 776: La direction commerciale d'une entreprise industrielle a augmenté régulièrement ses dépenses publicitaires pendant plusieurs années et voudrait y comparer la progression de son chiffre d'affaires. Elle dispose pour cela des données suivantes :

Année	Dépenses publicitaires x (en francs)	Chiffres d'affaires y (en milliers de francs)
1960	73200	35261
1961	74700	35771
1962	76200	36791
1963	77700	37301
1964	79200	37556
1965	80700	38066
1966	82200	38831

a. Tracez le nuage de points correspondant. En déduire la possibilité d'un ajustement linéaire.
b. Cherchez la droite de régression de y par rapport à x.
c. Quel devrait être, d'après l'ajustement trouvé, le montant des dépenses publicitaires pour atteindre un chiffre d'affaires de 45000 francs ?

Exercice 777: **Alignement de points et lien de causalité**

On considère le tableau suivant :

t_i : Année	1995	1996	1997	1998	1999	2000	2001
x_i : Nombre d'inscrits dans un club de belote	48	53	57	62	68	73	77
y_i : Nombre de hamburgers vendus dans un restaurant de Moscou	7000	7450	8000	8500	9050	9550	10000

1. Placer dans un repère le nuage de points $(t_i ; x_i)$ et constater que sa forme allongée justifie un ajustement affine.
2. Placer dans un autre repère le nuage de points $(t_i ; y_i)$ et constater que sa forme allongée justifie un ajustement affine.
3. **a)** placer dans un troisième repère le nuage de points $(x_i ; y_i)$ et constater que sa forme allongée justifie un ajustement affine.
b) Vérifier que la droite de régression de y en x admet comme équation $y = 103x + 2057$ et tracer cette droite. Commentez le résultat trouvé.

V Statistiques

Exercice 778: ***Comparaison de deux ajustements affines : droite de Mayer et droite de régression***

Le tableau suivant donne le PNB (en euros, par habitants) ainsi que le nombre d'hôpitaux (pour 1 million d'habitants) dans quelques pays européens.

Pays	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
[illegible]	5123	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3250	[illegible]	[illegible]

1. Représenter le nuage de points associé à la série statistique (*X*, *Y*).

Unités graphiques : • En abscisses : 1 cm pour 1000 euros. • En ordonnées : 1cm pour 200 hôpitaux.
On prendra pour origine le point (5000 ; 600).

2. Déterminer les coordonnées du point moyen G de ce nuage de points. Placer G sur le graphique.
3. Un premier ajustement affine : la droite de Mayer

Dans cette question, on considère deux sous-nuages : celui constitué des points correspondants aux pays *P*1, *P*2, *P*3 et *P*4 et celui constitué des points correspondants aux pays *P*5, *P*6, *P*7 et *P*8.

   a. Calculer les coordonnées des points moyens *G*1 et *G*2 des deux sous-nuages. Placer les points *G*1 et *G*2 sur le graphique.
   b. Démontrer qu'une équation de la droite (*G*1*G*2) sous la forme $y = mx + p$ est :
   $y = 0{,}15x - 199$ (On détaillera les calculs). (On arrondira *m* à 10^{-2} près et *p* à l'unité près)

La droite (*G*1*G*2) s'appelle la "droite de Mayer". Représenter cette droite sur le graphique.

4. Un deuxième ajustement affine : la droite de régression
   a. Déterminer une équation de la droite de régression de *y* en *x* par la méthode des moindres carrés. On notera *D* cette droite. Représenter *D* sur la graphique.
   b. Laquelle des deux droites (G_1G_2) et *D* réalise-t-elle le meilleur ajustement affine?
5. Estimations. À l'aide de l'équation de la droite (*D*) (ou à défaut celle de (G_1G_2)), et en détaillant les calculs, répondre aux deux questions suivantes :
   a. Un pays a un PNB de 23400 € par habitant. Quelle estimation peut-on faire du nombre d'hôpitaux (par million d'habitants) dans ce pays ? (On arrondira à l'unité près)
   b. Un pays a 3500 hôpitaux par million d'habitants. À combien peut-on estimer son PNB (en €, par habitants) ? (On arrondira à l'euro près)

Exercice 779: Cherchez la droite de régression de y par rapport à x sachant que

a $\overline{x} = 6.12,\ \overline{y} = 5.94,\ s_{xy} = -4.28,\ s_x = 2.32$

b $\overline{x} = 21.6,\ \overline{y} = 45.9,\ s_{xy} = 12.28,\ s_x = 8.77$

Exercice 780: On donne :

$$n = 6,\ \sum x = 61,\ \sum y = 89,\ \sum xy = 1108,\ \sum(x-\overline{x})^2 = 138 \text{ and } \sum(y-\overline{y})^2 = 284$$

a. Trouvez : i. La moyenne de X ii. La moyenne de Y
b. Trouvez : i. L'écart-type de X ii. L'écart-type de Y.
c. Trouvez la covariance de X et Y.
d. Trouvez la droite de régression de y par rapport à X.

Exercice 781:

Le prix de vente des terrains à bâtir dans la même commune rurale est donné par le tableau suivant :

Année	1980	1985	1987	1990	1995	1997	2000
Rang de l'année x_i	0	5	7	10	15	17	20
Prix du m² en francs y_i	58,8	60,9	62,1	67,5	71,7	73	73,8

1. Quelle est, en pourcentage, l'augmentation du prix du m² entre 1980 et 2000 ?
2. Représentez le nuage de points $M_i(x_i\,;\,y_i)$ dans un repère orthogonal où 5 cm représentent 10 ans en abscisse, 5 cm représentent 10 francs en ordonnées.
3. Déterminez le point moyen G du nuage et placez-le sur le graphique.
4. On considère que la position des points sur le graphique justifie un ajustement affine par la méthode des moindres carrés. Ecrire une équation de la droite d'ajustement affine de *y* en *x*, notée (D) [les coefficients sont arrondis à 0,01]. Tracer (D).
5. Estimer à 1 millier de francs près le prix d'un terrain de 1500m² en 2003.

4.4. Corrélation

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 correlées à des degrés très différents, et la droite trouvée ne s'ajuste pas nécessairement bien aux données. Un nombre qui décrit la validité de la droite d'ajustement et qui mesure le degré de dépendance linéaire entre les variables x et y est le coefficient de corrélation (de Pearson).

Définition :

Le **coefficient de corrélation linéaire de Pearson r** est le nombre $\mathbf{r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}}$ où

* σ_{xy} désigne la covariance de X et Y : $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

* σ_x est l'écart type de X : $\sigma_x = \sqrt{V(x)}$ avec $V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

* σ_y est l'écart type de Y : $\sigma_y = \sqrt{V(y)}$ avec $V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$

Autre calcul : $\mathbf{r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$

Remarque : La notation BI est plutôt Sxy, Sx, Sy.

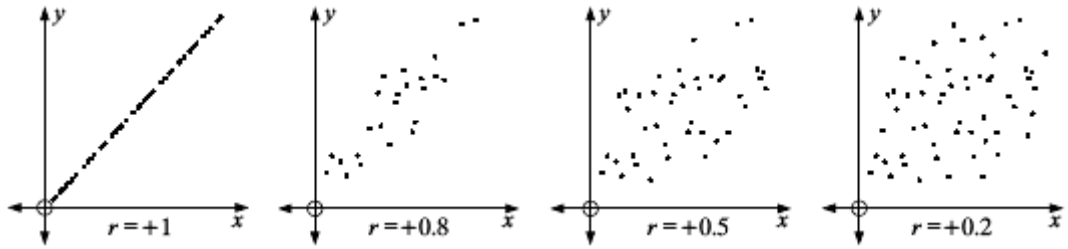
Propriétés du coefficient de corrélation linéaire :

- **$-1 \le r \le 1$ r est toujours compris entre -1 et +1.**
- Si r > 0 : entre x et y il y a une ***corrélation positive*** (dépendance linéaire positive).
- Si r < 0 : entre x et y il y a une ***corrélation négative***.
- Si r = 0 ou voisin de 0 : il n' y a pas de dépendance linéaire entre x et y.
- Si r = 1 ou r = -1 : les points du nuage sont rigoureusement alignés ; la dépendance linéaire est parfaite.
- Si r $\ge$ 0,95 alors la corrélation linéaire entre X et Y est forte. Dans ce cas un ajustement affine est justifié (Les points du nuage sont dans une situation proche de l'alignement).

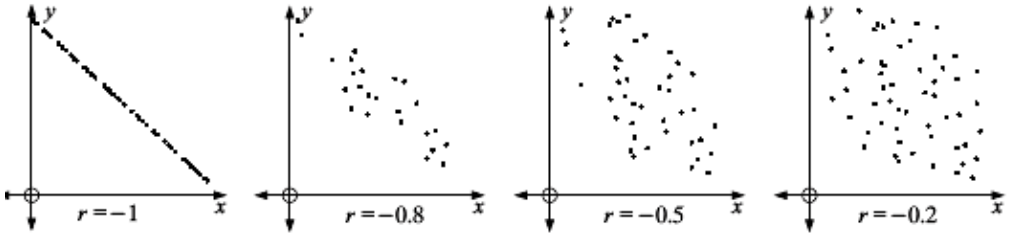
(Remarque : Ce seuil varie suivant les auteurs.)

Le plus souvent, on calcule d'abord le coefficient r pour voir si le degré de dépendance linéaire est suffisamment élevé pour justifier la recherche de la droite de régression.

Corrélation positive



Corrélation négative



V Statistiques

Exemple de calcul (à la main)

	Taille (en cm) (x_i)	Pointure du soulier (en cm) (y_i)	x_i^2	y_i^2	$x_i y_i$
Etudiant 1					
Etudiant 2					
Etudiant 3					
Etudiant 4					
Etudiant 5					
Etudiant 6					
Etudiant 7					
Etudiant 8					
Etudiant 9					
Etudiant 10					
Σ					
$\frac{1}{n}\Sigma$					

Var x = S_x^2 = **S_x =**

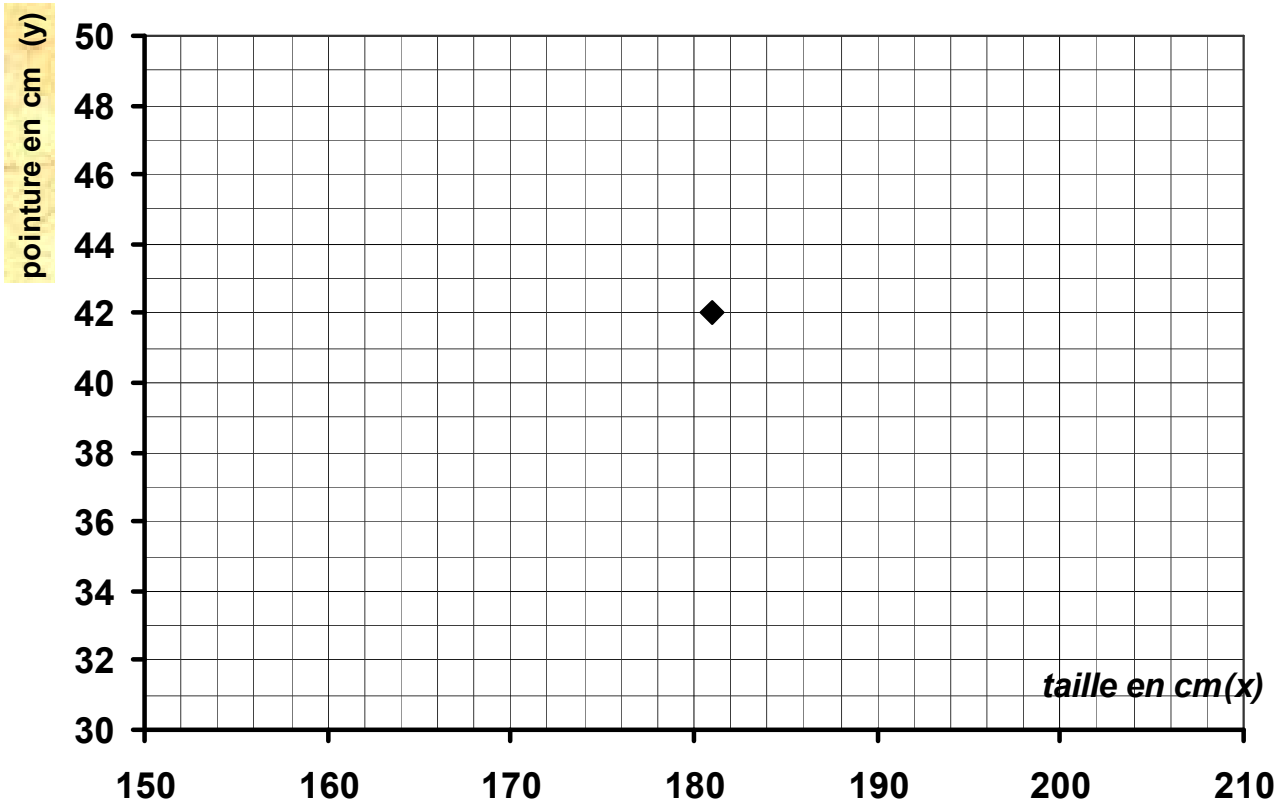
Var y = S_y^2 = **S_y =**

Cov (x,y) = S_{xy} = **r =**

Droite de régression de y sur x : **pente = a =**

Point moyen : G( )

Equation :



V Statistiques

SÉRIES STATISTIQUES À 2 VARIABLES : RÉSUMÉ

Définition : deux caractères X et Y dont on dispose, souvent généralement présentés sous forme de tableau :

X	x_1	x_2	x_3	...	x_n
Y	y_1	y_2	y_3	...	y_n

Ce que l'on peut vous demander de faire (ou de calculer) :

1) Nuage de points : dans un repère, c'est l'ensemble des points $M_i(x_i ; y_i)$ pour $1 \leq i \leq n$.

2) Point moyen d'un nuage : c'est le point $G(\bar{x} ; \bar{y})$ où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

3) Calcul du coefficient de corrélation linéaire : $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ où :

• σ_{xy} désigne la covariance de X et Y : $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

• σ_x est l'écart type de X : $\sigma_x = \sqrt{V(x)}$ avec $V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

• σ_y est l'écart type de Y : $\sigma_y = \sqrt{V(y)}$ avec $V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$

Si $|r|$ est proche de 1 alors la corrélation linéaire entre X et Y est forte. Un ajustement affine est alors justifié.

(les points du nuage sont dans une direction privilégiée (presque alignés)).

Si $r^2 = 1$ (c'est-à-dire $r = 1$ ou $r = -1$) alors les points du nuage sont alignés.

4) Déterminer un ajustement affine (uniquement dans le cas où $|r| \geq$ [illegible]) :

• Droite de régression de y en x par la méthode des moindres carrés :

c'est la droite d'équation $y = ax + b$ avec : $a = \frac{\sigma_{xy}}{V(x)}$ et $b = \bar{y} - a\bar{x}$

• Droite de régression de x en y (peu utilisée) par la méthode des moindres carrés :

c'est la droite d'équation $x = a'y + b'$ avec : $a' = \frac{\sigma_{xy}}{V(y)}$ et $b' = \bar{x} - a'\bar{y}$

Les droites de régression passent toujours par le point moyen du nuage.

Intersection entre les deux : $G(\bar{x} ; \bar{y})$

5) Déterminer un autre type d'ajustement (quadratique, exponentiel ou autre ...)

Dans ce cas, l'énoncé vous guide à l'aide d'un changement de variable ($z_i = \sqrt{y_i}$ ou $z_i = \ln y_i$ ou autre ...)

afin de se ramener à un ajustement affine entre X et Z.

6) Calculer la somme des résidus quadratiques (méthode des moindres carrés) :

Soit D une droite d'ajustement d'équation $y = ax + b$. La somme des résidus quadratiques est :

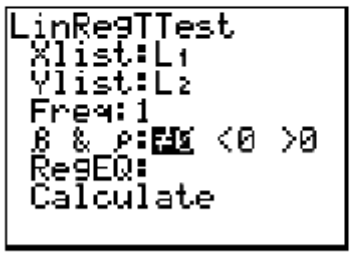
$S = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2$ (somme des carrés des "écarts verticaux" entre les points du nuage et la droite D)

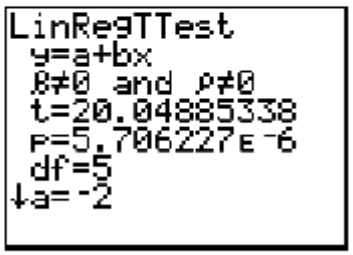
Dans le cas de la droite de régression de y en x : $S = nV(y)(1 - r^2)$.

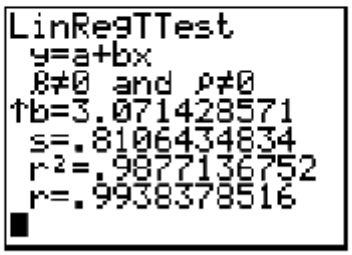
Usage de la calculatrice

Droite de régression

- Saisissez les données dans « L1 » et « L2 » (ou dans n'importe quelle autre liste).
- Utilisez « STAT – TESTS – E: LinRegTTest – ENTER » puis saisissez les noms des listes utilisées. Déplacez le curseur vers le bas jusqu'à « Calculate » puis appuyez sur la touche « ENTER ».
- La droite de régression est obtenue ainsi que le coefficient de corrélation, *r*.

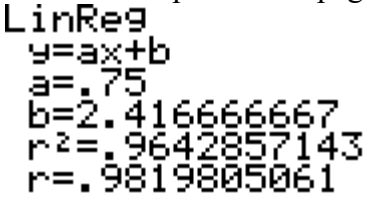






On peut également utiliser « STAT – CALC LinReg(ax + b) » . Néanmoins, avant de l'utiliser, « DiagnosticOn » doit être réglé (cette option se trouve dans le menu « CATALOG »).

Pour l'exemple de la page 38 :



Si les données sont fournies, les élèves devraient être capables d'écrire le coefficient de corrélation r et l'équation de la droite de régression directement à partir de la calculatrice. Néanmoins, ils devraient également être familiers avec les formules et savoir comment les utiliser.

Exercice 782:

Le tableau suivant recense, par clinique, le nombre de postes de personnel non médical en fonction du nombre de lits de la clinique :

1. Construire le nuage de points $M_i(x_i ; y_i)$ correspondant à cette série statistique.
Unités graphiques : en abscisse : 1 cm pour 10 lits en ordonnée : 1 cm pour 20 postes.
2. Calculer les coordonnées du point moyen G du nuage et le placer sur le graphique.
3. Calculer le coefficient de corrélation linéaire r. Un ajustement affine est-il justifié ?
4. Déterminer une équation de la droite de régression D de y en x par la méthode des moindres carrés. Tracer la droite D sur le graphique. (Marquer les points utilisés pour tracer D)
5. Une clinique possède 25 lits. En utilisant les résultats de la question 4, à combien peut-on estimer, par calcul, le nombre de postes de personnel non médical ? Illustrer sur le graphique.

Exercice 783:

Un hypermarché dispose de 20 caisses. Le tableau ci-dessous donne le temps moyen d'attente à une caisse en fonction du nombre de caisses ouvertes :

Nombre de caisses ouvertes X	3	4	5	6	8	10	12
Temps moyen d'attente (en minutes) Y	16	12	9,6	7,9	6	4,7	4

1. Construire le nuage de points $M_i(x_i ; y_i)$ correspondant à cette série statistique.Unités graphiques :
en abscisse : 1 cm pour une caisse ouverte en ordonnée : 1 cm pour une minute d'attente.

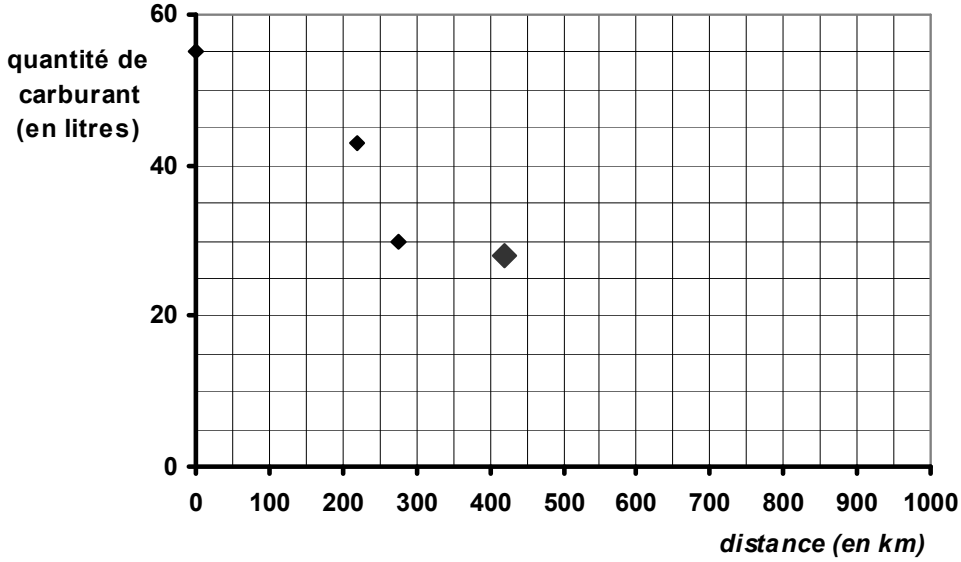
2. Calculer les coordonnées du point moyen G du nuage et le placer sur le graphique.
3. Un ajustement affine.
   a) Calculer le coefficient de corrélation linéaire r.
   b) Déterminer, l'équation de la droite de régression D de y en x par la méthode des moindres carrés. Tracer la droite D sur le graphique. (Marquer les points utilisés pour tracer D)
   c) Estimer à l'aide d'un calcul utilisant l'équation de la droite D :
      i) Le nombre de caisses à ouvrir pour que le temps moyen d'attente à une caisse soit de 5 minutes.
      ii) Le temps moyen d'attente à la caisse lorsque 15 caisses sont ouvertes.
      iii) Pensez-vous que, dans le cas de la question ii), l'ajustement affine soit fiable ?

Exercice 784: ***Exercice d'examen (Epreuve 1)***

Le tableau suivant donne la quantité de carburant restant dans le réservoir d'une voiture et le nombre de kilomètres parcourus après avoir rempli le réservoir.

Distance parcourue (en km)	0	220	276	500	680	850
Quantité de carburant restant dans le réservoir (litres)	55	43	30	24	10	6

a. Sur le diagramme de dispersion ci-dessous, placez les autres points.



b. La distance moyenne parcourue est de 421 km ($\bar{x}$), et la quantité moyenne de carburant restant dans le réservoir est de 28 litres ($\bar{y}$). Ce point est placé sur le diagramme de dispersion.
c. Esquissez la droite de régression.

Une voiture a parcouru 350 km.

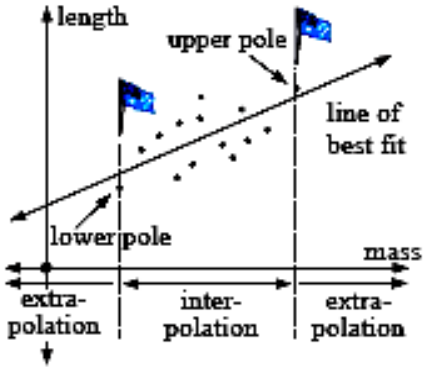
d. Utilisez votre droite de régression pour estimer la quantité de carburant qui reste dans le réservoir de cette voiture.

Deux remarques importantes :

➢ Une forte corrélation entre deux grandeurs x et y ne signifie pas nécessairement qu'il y a un **lien de causalité** entre ces grandeurs. Par exemple, il est possible que les deux grandeurs soient des effets d'une même cause. Par exemple, on peut constater une forte corrélation entre les notes en latin et en mathématiques dans un groupe d'étudiants, ce qui ne veut pas dire pour autant que la bonne note obtenue en latin favorise une bonne note en mathématiques ou vice-versa.
(voir aussi l'exercice 68)

➢ **Interpolation et extrapolation**

Lorsqu'on utilise une droite de régression pour faire des prévisions, il faut distinguer entre **l'interpolation et l'extrapolation**. On fait une interpolation, si on fait des prévisions à l'intérieur du champ des données utilisées pour trouver la droite de régression. On fait une extrapolation, si on va au-delà du champ des données. Il est clair que les prévisions obtenues par extrapolation sont en général beaucoup moins fiables, voire deviennent parfois absurdes.



Exercice 785: (Exercice d'examen) On décide de sélectionner un échantillon aléatoire de 10 élèves afin de voir s'il existe une relation linéaire entre la taille d'un individu et la pointure de son soulier. Les résultats sont présentés dans le tableau ci-dessous.

Taille (cm) (x)	Pointure (y)
175	8
160	9
180	8
155	7
178	10
159	8
166	9
185	11
189	10
173	9

a. Donnez l'équation de la droite de régression de la pointure (y) sur la taille (x), en donnant votre réponse sous la forme y = mx + c.
b. Utilisez votre réponse de la partie (a) pour prédire la pointure d'un élève qui mesure 162 cm.
c. Donnez la valeur du coefficient de corrélation.
d. Décrivez la corrélation entre la taille et la pointure du soulier.

Exercice 786:

Les tomates d'une plantation sont traitées avec un engrais chimique pour favoriser la croissance et la résistance aux maladies. Une étude sur l'efficacité de cet engrais a donné les résultats suivants :

Concentration (en mL/L)	3	5	6	8	9	11
Nombre de tomates par plante	67	90	103	120	124	150

a. Représentez les données par un nuage de points.
b. Utilisez la méthode des moindres carrés pour déterminer une équation de la droite de régression.
c. Interprétez la pente et l'ordonnée à l'origine de cette droite.
d. Utilisez l'équation trouvée pour prédire le rendement de tomates pour une concentration de 7 mL/L. La prévision est-elle raisonnable ?
e. Si on utilisait une concentration de 50 mL/L, est-ce qu'on aurait assuré une bonne récolte de tomates ? Expliquez.

5. Test d'indépendance du KHI-CARRE

5.1. Position du problème

Supposons par exemple que nous voulions savoir s'il existe un lien entre les intentions de vote des électeurs d'un certain pays et leur âge. Pour cela nous avons réalisé un sondage parmi 500 électeurs, et nous avons résumé les résultats dans le tableau suivant appelé

tableau de contingence ou table des fréquences (effectifs) croisées:

Âge des électeurs → *Parti ↓*	**18 à 35**	**36 à 50**	**50 +**	**Total**
A	**23**	**32**	**45**	100
B	**45**	**93**	**78**	216
C	**92**	**58**	**34**	184
Total	160	183	157	**500**

Le problème à résoudre est celui de la liaison entre les deux variables de classification : l'âge des électeurs et le parti pour lequel ils ont voté. Plus précisément : L'intention de vote dépend-elle de l'âge d'un électeur ?
Il est difficile d'évaluer la nature et le degré d'un lien statistique entre deux caractères. Nous nous limitons à la question la plus simple : Est-ce qu'il existe un lien statistique quelconque entre l'âge des électeurs et leurs intentions de vote ? Pour montrer qu'il existe un lien, nous partons de l'hypothèse contraire, qui est
l'hypothèse nulle H_o : il n'y a aucun lien entre l'âge des électeurs et leurs intentions de vote.
L'idée est de montrer que, dans cette hypothèse, il est très improbable de trouver les résultats ci-dessus. Si on peut montrer p.ex. que cette probabilité est inférieure à 5 %, alors on peut conclure qu'il existe un lien statistique avec une probabilité supérieure à 95 %.

Le tableau de contingence ci-dessus donne les ***fréquences observées***. On commence alors par calculer les ***fréquences théoriques ou attendues***, c'est-à-dire les fréquences qu'on aurait dû observer s'il n'y avait aucun lien entre les deux caractères.

Si l'hypothèse nulle est vraie, les cellules de la table des fréquences croisées vont adopter un schéma typique qu'on appellera « comportement théorique ». Quel est-il ? Revenons pour cela à la table ci-dessus : si effectivement, il n'y a pas de raison de penser que l'âge a une influence sur le comportement des électeurs, alors on doit trouver les mêmes pourcentages d'électeurs des différents partis dans les différentes catégories d'âge.

Par exemple 100 électeurs sur un total de 500 ont voté pour le parti A : $\frac{100}{500} = 0{,}2 = 20\ \%$.

Théoriquement, on devrait donc trouver dans chaque catégorie d'âge 20 % des personnes qui votent pour A. Cela fait :

Entre 18 et 35 ans : 20 % de 160 = 0,2 · 160 = 32 électeurs

Entre 36 et 50 ans : 20 % de 183 = 0,2 · 183 = 37 électeurs (valeur arrondie)

Au-dessus de 50 ans : 20 % de 157 = 0,2 · 157 = 31 électeurs (valeur arrondie) etc.

On peut aussi partir de la répartition des électeurs dans les catégories d'âge. Par exemple, 160 électeurs sur 500 sont âgées entre 18 et 35 ans : $\frac{160}{500} = 0{,}32 = 32\ \%$.

Théoriquement, on devrait alors avoir entre 18 et 35 ans 0,32 · 100 = 32 électeurs du parti A.

En procédant de cette sorte, on obtient la table de contingence des fréquences « théoriques » ou « calculées » suivante (arrondies à l'unité) :

Âge des électeurs → *Parti ↓*	**18 à 35**	**36 à 50**	**50 +**	**Total**
A	**32**	**37**	**31**	100
B	**69**	**79**	**68**	216
C	**59**	**67**	**58**	184
Total	160	183	157	**500**

Nous constatons que les totaux de ligne et de colonne n'ont pas changé. Ces totaux sont appelés ***totaux marginaux*** (ils se situent en marge du tableau). D'autre part, le tableau renferme des proportions égales à plusieurs points de vue :

$$\frac{32}{160}=\frac{37}{183}=\frac{31}{157}=\frac{100}{500}=\text{pourcentage réalisé par A}$$

$$\frac{69}{160}=\frac{79}{183}=\frac{68}{157}=\frac{216}{500}=\text{pourcentage réalisé par B}$$

$$\frac{59}{160}=\frac{67}{183}=\frac{58}{157}=\frac{184}{500}=\text{pourcentage réalisé par B}$$

et aussi

$$\frac{32}{100}=\frac{69}{216}=\frac{59}{184}=\frac{160}{500}=\text{pourcentage des électeurs entre 18 et 35 ans}$$

$$\frac{37}{100}=\frac{79}{216}=\frac{67}{184}=\frac{183}{500}=\text{pourcentage des électeurs entre 36 et 50 ans}$$

$$\frac{31}{100}=\frac{68}{216}=\frac{58}{184}=\frac{157}{500}=\text{pourcentage des électeurs au – dessus de 50 ans}$$

On comprend aisément que, plus la table des fréquences observées est proche de la table des fréquences calculées (représentation idéale de l'hypothèse d'indépendance), plus on acceptera l'hypothèse que les variables de classement n'ont pas de lien ; à l'inverse, plus les fréquences observées s'écartent des fréquences calculées plus on doutera de l'hypothèse nulle et plus on acceptera la dépendance.
Il faut donc trouver une mesure de proximité entre les cellules des deux tables.

L'expression $\chi^2 = \sum_{i=1}^{k} \frac{(f_{oi}-f_{ci})^2}{f_{ci}}$ où f_{oi} représente la fréquence observée de la cellule i et f_{ci} la fréquence calculée de la cellule i, (k étant le nombre de cellules de la table) mesure la distance entre les deux tableaux pour toutes les cellules.

Cette expression est d'autant plus petite que l'hypothèse nulle H_0 est vraie ; si les deux tables sont identiques, elle vaut 0. Plus cette valeur est élevée, plus on sera amené à douter de H_0. Dans l'exemple ci-dessus :

$$\chi^2=\frac{(23-32)^2}{32}+\frac{(32-37)^2}{37}+\frac{(45-31)^2}{31}+\frac{(45-69)^2}{69}+\frac{(93-79)^2}{79}$$
$$+\frac{(78-68)^2}{68}+\frac{(92-59)^2}{59}+\frac{(58-67)^2}{67}+\frac{(34-58)^2}{58}$$

$\chi^2 = \underline{\mathbf{51,43}}$ La question est maintenant de connaître la valeur de χ^2 à partir de laquelle on laisse tomber l'hypothèse nulle pour conclure qu'il y a une dépendance entre les deux facteurs.

Calcul en Excel

Fréquences observées

Âge des électeurs → *Parti ↓*	**18 à 35**	**36 à 50**	**50 +**	**Total**
A	**23**	**32**	**45**	100
B	**45**	**93**	**78**	216
C	**92**	**58**	**34**	184
Total	160	183	157	**500**

Fréquences calculées

Âge des électeurs → *Parti ↓*	**18 à 35**	**36 à 50**	**50 +**	**Total**
A	**32**	**37**	**31**	100
B	**69**	**79**	**68**	216
C	**59**	**67**	**58**	184
Total	160	183	157	**500**

Calcul du Khi-carré:

2,531	0,676	6,323	9,530
8,348	2,481	1,471	12,299
18,458	1,209	9,931	29,598
29,337	4,366	17,724	**51,427**

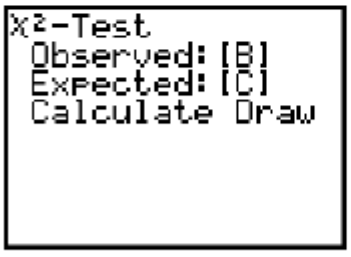
Avec la TI 84

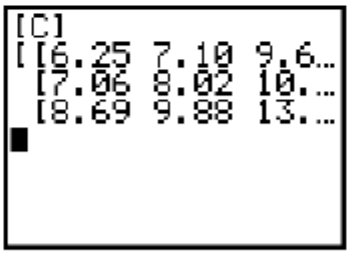
Test du chi-carré

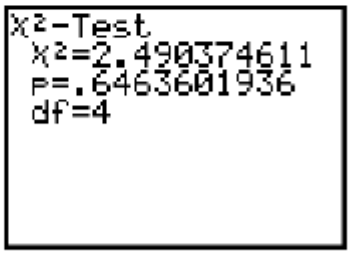
Saisissez les fréquences observées dans une matrice. Dans l'exemple, les fréquences observées suivantes sont saisies dans la matrice B :

5 6 12
8 10 8
9 9 14

- Utilisez « STAT – TESTS – C: X^2 – Test – ENTER ». Saisissez [B] après « Observed » et [C] après « Expected » (allez à « MATRIX » afin de saisir ces lettres). Placez le curseur sur « Calculate » et appuyez sur la touche « ENTER » pour la valeur du chi-carré et le nombre de degrés de liberté. La calculatrice trouve automatiquement les fréquences théoriques.







À moins qu'un autre type de question soit posé, la valeur du chi-carré, le nombre de degrés de liberté et les fréquences théoriques peuvent être obtenus à l'aide de la calculatrice. Néanmoins, les élèves devraient connaître la formule et savoir comment l'utiliser.

5.2. Test d'indépendance

La distribution du Chi-carré ou χ^2 est utilisée dans un grand nombre de tests statistiques et particulièrement dans les analyses des tables de fréquences croisées appelées aussi tables de contingence. Ces tables de type bivarié (on « croise » deux variables), extrêmement répandues dans les analyses de données dans le domaine des sciences sociales, permettent de vérifier s'il existe des relations de dépendance entre les lignes et le colonnes d'un tableau de fréquences. Elles sont applicables à tous les types de variables : nominales, ordinales et de ratio (catégorisées). Elles s'intéressent en fait à tester l'indépendance de deux variables (dans une enquête, elles permettent de vérifier que les réponses fournies à une question peuvent être liées aux réponses données à une autre).

5.2.a. Autre exemple

Dans l'enquête Moniteur de Sécurité menée à Liège en 1998, on désire analyser les réponses concernant la victimisation et le sexe ; on a ainsi construit la table suivante :

Victimes selon le Sexe	Hommes	Femmes	Totaux colonnes
Non-Victimes	**70**	**85**	155
Victimes	**79**	**82**	161
Totaux Lignes	149	167	316

Table des fréquences observées

Le problème qu'il faut résoudre est celui de la liaison entre les variables de classification : dans l'exemple ci-dessus, il faut se demander s'il y a une relation entre la victimisation et le sexe au niveau de la population d'où a été extrait l'échantillon.

5.2.b. Le test du χ^2

Pour vérifier qu'il y a une relation entre les variables utilisées dans la construction de la table des fréquences, on va s'intéresser au comportement des cellules et tenter de découvrir un « profil ». Par exemple, si on suppose qu'il existe une relation parfaite entre la victimisation et le sexe, du type « seules les femmes sont victimes », on n'a aucune peine à imaginer la structure de la table : on doit observer 179 Hommes Non-victimes et 167 Femmes victimes, ce qui revient à dire que toutes les fréquences qui ne se trouvent pas sur la diagonale principale de la table (la diagonale principale est constituée des cellules sur la direction « nord-ouest » - « sud-est ») sont nulles.
En réalité, le test basé sur la logique du χ^2 part de l'hypothèse inverse : on part de l'hypothèse qu'il n'existe pas de relation entre les variables de classement et donc,

***l'hypothèse nulle H_0 : Il y a* INDEPENDANCE.**

On a formulé cette hypothèse dans le but de la rejeter. Par exemple, pour décider qu'une pièce est truquée on formule l'hypothèse qu'elle est honnête. Ou pour décider qu'un procédé de fabrication est meilleur qu'un autre, nous formulons l'hypothèse qu'il n'y a aucune différence entre les procédés.

Toute hypothèse qui diffère de Ho est appelée ***hypothèse alternative*** : **H_1 : il y a dépendance**.

Si l'hypothèse nulle est vraie, les cellules de la table des fréquences croisées vont adopter un schéma typique qu'on appellera ***« comportement théorique »***. Il est calculé comme dans l'exemple précédent.

Victimes selon le Sexe	Hommes	Femmes	Totaux colonnes
Non-Victimes	**73**	**82**	155
Victimes	**76**	**85**	161
Totaux Lignes	149	167	316

Table des fréquences calculées (ou théoriques)

La comparaison des deux tables révèle de nouveau que les totaux marginaux n'ont pas changé (l'application de l'hypothèse d'indépendance ne modifie pas l'échantillon) ; seule la distribution des cellules s'est modifiée. La table des fréquences calculées illustre la parfaite indépendance des variables de classement : la connaissance de l'appartenance d'une observation à la catégorie « Hommes » ou « Femmes » ne permet pas de la classer dans la catégorie des « Victimes » ou des « Non-victimes ». Dans une table de fréquences croisées, l'indépendance des critères de classification entraîne que la fréquence d'une cellule est égale au produit des totaux marginaux correspondant à celle-ci divisé par la taille de l'échantillon.

On obtient : $\chi^2 = \frac{(70-73)^2}{73} + \frac{(85-82)^2}{82} + \frac{(79-76)^2}{76} + \frac{(82-85)^2}{85} = 0.457$

Exercice 787: Trouvez χ^2_{calc} pour les tables de contingence suivantes :

a Factor M

	M_1	M_2	
N_1	31	22	53
N_2	20	27	47
	51	49	100

b Factor S

	S_1	S_2	
R_1	28	17	
R_2	52	41	

c Factor A

	A_1	A_2	
B_1	24	11	
B_2	16	18	
B_3	25	12	

d Factor T

	T_1	T_2	T_3	T_4	
D_1	31	22	21	16	
D_2	23	19	22	13	

Exercice 788: Vérifiez vos réponses à l'aide d'une calculatrice.

5.2.c. Seuil de signification

Au cours d'un test d'hypothèse, il faut fixer le ***niveau ou seuil de signification*** de ce test, c.à.d. la probabilité maximale que nous acceptons pour faire une erreur.

En pratique (et au BI), on utilise souvent les seuils 1 %, 5 % et 10 %.

5.2.d. Degrés de liberté

La décision d'accepter ou de rejeter l'hypothèse dépend aussi du ***nombre de degrés de liberté (df) :*** Les degés de liberté sont donnés par la formule :

$$\boxed{Df = (L-1)\cdot(C-1)}$$

où L est le nombre de lignes de la table de contingence, et C le nombre de colonnes.

- Pour l'exemple 5.2.a., on a : $Df = (2-1)\cdot(2-1) = 1$

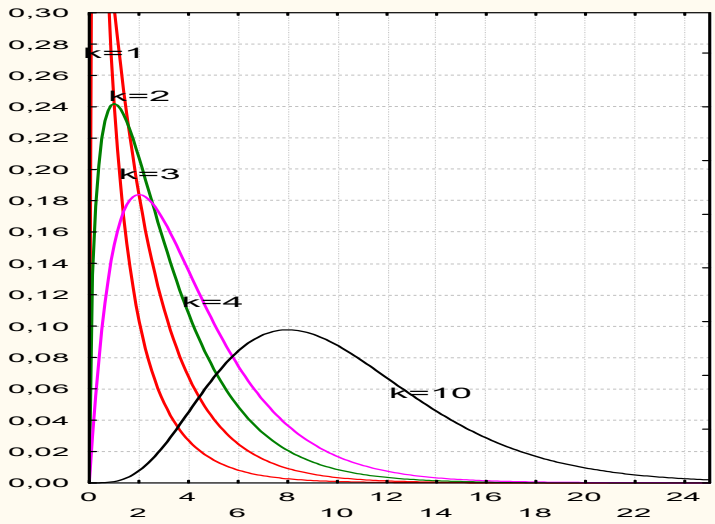
Il peut paraître étrange de n'avoir qu'un degré de liberté alors qu'on a quatre cellules. Cependant, si nous connaissons les totaux des lignes et des colonnes, la donnée d'une fréquence détermine toutes les autres fréquences.

- Pour l'exemple 5.1 : $Df = (3-1)\cdot(3-1) = 4$.

5.2.e. Test d'indépendance

Pour arriver finalement à une décision concernant l'hypothèse H_o, il faut regarder la distribution de CHI-CARRE dont les valeurs dépendent du nombre de degrés de liberté.

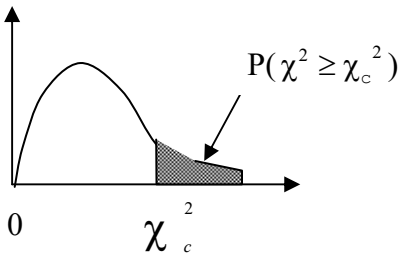
La forme de la fonction de densité de la loi pour les degrés de liberté variant de 1 à 4 puis 10 est représentée ci-dessous



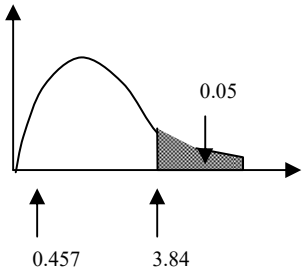
Chaque probabilité correspond à une surface sous la courbe. Par exemple, la probabilité pour que la valeur du χ^2 calculée dépasse une certaine valeur critique correspond à l'aire sous la courbe à droite de cette valeur critique. La table ci-dessous présente quelques-unes de ces probabilités :

degrés de liberté	Probabilités							
	0.9	**0.5**	**0.3**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**
1	0.016	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.211	1.386	2.408	3.219	4.605	5.991	7.824	9.21
3	0.584	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	1.064	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	1.61	4.351	6.064	7.289	9.236	11.07	13.388	15.086
6	2.204	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	2.833	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	3.49	7.344	9.524	11.03	13.362	15.507	18.168	20.09
9	4.168	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	4.865	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	5.578	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	6.304	11.34	14.011	15.812	18.549	21.026	24.054	26.217
13	7.042	12.34	15.119	16.985	19.812	22.362	25.472	27.688
14	7.79	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	8.547	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	9.312	15.338	18.418	20.465	23.542	26.296	29.633	32
17	10.085	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	10.865	17.338	20.601	22.76	25.989	28.869	32.346	34.805
19	11.651	18.338	21.689	23.9	27.204	30.144	33.687	36.191
20	12.443	19.337	22.775	25.038	28.412	31.41	35.02	37.566
21	13.24	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	14.041	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	14.848	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	15.659	23.337	27.096	29.553	33.196	36.415	40.27	42.98
25	16.473	24.337	28.172	30.675	34.382	37.652	41.556	44.314
26	17.292	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	18.114	26.336	30.319	32.912	36.741	40.113	44.14	46.963
28	18.939	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	19.768	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	20.599	29.336	33.53	36.25	40.256	43.773	47.962	50.892

Ex. : Pour 5 degrés de liberté, la valeur 11.07 a 5% de chances d'être dépassée.

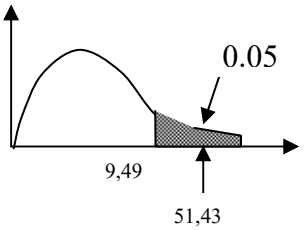


Exemple 5.2.a :
H_0 : Il y a indépendance
H_1 : Les variables sont dépendantes
a = 0.05
La valeur 0.457 provient d'une distribution d'échantillonnage χ^2 à (2-1)(2-1) = 1 degré de liberté :



La valeur d'échantillon trouvée est en-dessous de la valeur-critique, elle se trouve donc dans la région d'acceptation de H_0 : les variables victimisation et sexe ne sont pas liées.

Exemple 5.1 :
H_0 : Il y a indépendance
H_1 : Les variables sont dépendantes
a = 0.05
La valeur 51,43 provient d'une distribution d'échantillonnage χ^2 à 4 degrés de liberté .



La valeur d'échantillon trouvée est au-dessus de la valeur-critique, elle se trouve donc dans la région du rejet de H_0 respectivement d'acceptation de H_1 : Les variables âge et comportement de vote sont liés au seuil de signification 5 %.

Test d'indépendance : Marche à suivre

- On définit l'hypothèse nulle H_o : les variables sont indépendantes et l'hypothèse alternative H_1 : les variables sont dépendantes.
- On calcule le nombre des degrés de liberté : df = (L-1)(C-1)
- On fixe le seuil de signification : 10 %, 5 % ou 1 %.
- On formule l'inégalité de rejet de Ho : $\chi^2_{calc} > k$, où k est obtenue à partir du tableau des valeurs critiques.
- A partir du tableau de contingence des fréquences observées, on construit le tableau correspondant des fréquences calculées.
- On calcule $\chi^2_{calc} = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ci})^2}{f_{ci}}$.
- Nous acceptons ou rejetons H_o.
- Au niveau de signification 5 %, nous pouvons aussi utiliser les valeurs de p pour prendre notre décision : Si p < 0,05 : Ho acceptée
  Si p > 0,05 : Ho rejeté

Exercice 789: Trouvez les nombres des degrés de liberté pour les tables de contingence de l'exercice 13.

Exercice 790: On veut savoir si la couleur des cheveux dépend de la couleur des yeux. Pour un échantillon de 200 personnes choisies au hasard, on a établi le tableau de contingence suivant :

	Cheveux foncés	Cheveux clairs	
Yeux bruns	**54**	**32**	86
Yeux bleus	**42**	**72**	114
	96	104	200

Testez l'indépendance des deux variables au niveau de signification 0,05.

V Statistiques

Exercice 791: Un médicament est testé dans un groupe de 100 adultes qui sont tous infectés par un certain virus. 76 adultes sont guéris après deux semaines. Dans un autre groupe de 100 adultes qui sont également infectés, le médicament n'est pas donné. Parmi eux, 64 sont guéris après deux semaines. Peut-on dire que le médicament guérit cette maladie virale ?

Exercice 792: On a réalisé un sondage sur la peine capitale :

		Groupes d'âge			
		18-32	**33-45**	**46-60**	
Pour ou contre	**Pour**	22	25	16	63
	contre	98	115	84	297
		120	140	100	360

Peut-on affirmer, au seuil de signification de 5 %, que la position par rapport à la peine capitale est indépendante de l'âge ?

Exercice 793: On a réalisé une enquête sur le lien qui existe entre le niveau des études et de degré de satisfaction avec le travail :

	Êtes-vous satisfait avec votre travail ?		
	oui	**non**	
Avez-vous terminé des études universitaires **oui**	272	618	890
non	238	292	530
	510	910	1420

Peut-on affirmer, au seuil de signification de 5 %, que les réponses données aux deux questions sont indépendantes ?

Exercice 794: Les sociologues pensent pour la plupart que la consommation d'alcool chez les jeunes dépend de la classe sociale. Les résultats établis sur un échantillon de 230 jeunes sont donnés dans le tableau suivant :

		Consommation d'alcool		
		souvent	**parfois**	**aucune**
Parents	**Lower class**	10	21	9
	Middle class	24	90	24
	Upper class	13	19	7

Utilisez un seuil de signification de 5 % pour tester cette hypothèse.

Exercice 795: Le tableau suivant est le résultat d'une étude sur le lien éventuel entre le quotient d'intelligence (IQ) et la consommation de cigarettes.

	Quotient d'intelligence			
	bas	**moyen**	**élevé**	**Très élevé**
Consommation de cigarettes **Non-fumeur**	283	486	226	38
Moyen fumeur	123	201	58	18
Grand fumeur	100	147	64	8

Testez au seuil de signification de 1 % s'il y a un lien entre le Q_i et la consommation de cigarettes.

Exercices d'examen

Exercice 796: Un chercheur a consulté 500 hommes et femmes afin de savoir si la couleur de leur voiture était indépendante du sexe. Les couleurs étaient rouge, vert, bleu, noir et argent. Un test d'indépendance du χ^2 a été mené au seuil de signification de 5 % et la valeur expérimentale trouvée a été de 8,73.

a. Ecrivez l'hypothèse nulle.
b. Trouvez le nombre de degrés de liberté pour ce test.
c. Donnez la valeur critique pour ce test.
d. La couleur de la voiture est-elle indépendante du sexe ? Donnez une raison claire pour justifier votre réponse.

Exercice 797: Tom effectue un test d'indépendance du khi-carré afin de savoir s'il existe un lien entre le temps de préparation avant un tir de pénalité (court, moyen ou long) et le résultat du tir (compte un but ou ne compte pas un but). Tom fait le test au seuil de signification de 10 %.

a. Ecrivez l'hypothèse nulle.
b. Trouvez le nombre de degrés de liberté pour ce test.
c. La valeur de p pour ce test est de 0,073. Quelle conclusion Tom peut-il tirer ? Justifiez votre réponse.

Exercice 798: La couleur des yeux et le sexe de 500 élèves ont été notés , et les résultats sont indiqués dans le tableau ci-dessous.

	bleu	brun	vert
Masculin	18	152	50
Féminin	40	180	60

On croit que la couleur des yeux est liée au sexe dans une école de Banff. On décide de tester cette hypothèse à l'aide d'un test du χ^2, au seuil de signification de 5 %.

a. Ecrivez l'hypothèse nulle pour cette expérience.
b. Montrez que le nombre de degrés de liberté est 2.
c. Donnez une valeur critique du $\chi 2$ correspondant au nombre de degrés de liberté.
d. Calculez la valeur expérimentale du $\chi 2$ pour ces données.
e. Est-il possible de conclure que la couleur des yeux est liée au sexe dans cette école ? Donnez une raison claire pour justifier votre réponse.

Exercice 799: On décide de sélectionner un échantillon aléatoire de 10 élèves afin de voir s'il existe une relation linéaire entre la taille d'un individu et la pointure de son soulier. Les résultats sont présentés dans le tableau ci-dessous.

Taille (cm) (x)	**Pointure (y)**
175	8
160	9
180	8
155	7
178	10
159	8
166	9
185	11
189	10
173	9

a. Donnez l'équation de la droite de régression de la pointure (y) sur la taille (x), en donnant votre réponse sous la forme $y = mx + c$.
b. Utilisez votre réponse de la partie (a) pour prédire la pointure d'un élève qui mesure 162 cm.
c. Donnez la valeur du coefficient de corrélation.
d. Décrivez la corrélation entre la taille et la pointure du soulier.