LEÇON N^{O} 4

Statistique à une ou deux variables. Représentation et analyse des données.

Clément BOULONNE

Session 2020

Préambule

Niveau de la leçon

Collège, Seconde, Première S, Terminale STMG

Prérequis

Aucun

Références

— R. NOEL, *Statistiques descriptives*. http://amphimaths.chez-alice.fr/N1/stats_desc_poly.pdf.
— J. LEVY, *Séries statistiques*. http://jellevy.com/Classes/2nde/Statistiques/Cours/Cours_statistique.htm.
— P. BRACHET, *Statistiques : résumé de cours et méthodes*. Première S. http://www.xm1math.net/seconde/seconde_chap9_cours.pdf.
— F. GAUDON, *Statistiques à une ou deux variables, cours, terminale STG*, 15 novembre 2009. URL : http://mathsfg.net.free.fr/terminale/TSTG/TSTG2009/statistiquesdeuxvarciyrsTSTG.pdf.

Table des matières

1 Statistiques à une variable

1.1 Premières définitions et exemples

Définition 1.1 (Statistiques). La statistique étudie certaines caractéristiques : *caractères* ou *variables* d'un ensemble fini qu'on appelle *population*. Les éléments de cette population étudiée sont appelés *individus*.

Définition 1.2 (Type de variables). On peut classer en trois catégories les variables rencontrées :

Qualitative numérique et fait l'objet de calcul. Par exemple, des couleurs ou des sports favoris.

Quantitative discrète si la variable prend qu'un nombre fini de valeurs (on appelle *modalités* de telle valeur et on les notera x_i). Par exemple, le nombre de frères et sœurs (ne peut qu'être un nobre entier).

Quantitative continue si la variable prend ses valeurs dans un intervalle (*classe*). Par exemple, âge, taille et poids.

Exemple 1.3. Voici une liste de 30 notes d'un Devoir Surveillé de 2nde d'un lycée parisien :

5	10	12	13	20	14
15	8	3	4	5	1
20	14	12	3	5	19
10	4	9	10	15	12
11	12	14	20	4	0

On peut regrouper ces notes par ordre croissant et on les compte :

Note	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Effectif	1	1	0	2	3	3	0	0	1	1	3	1	4	1	3	2	0	0	0	1	3

et on peut regrouper ces notes par intervalle :

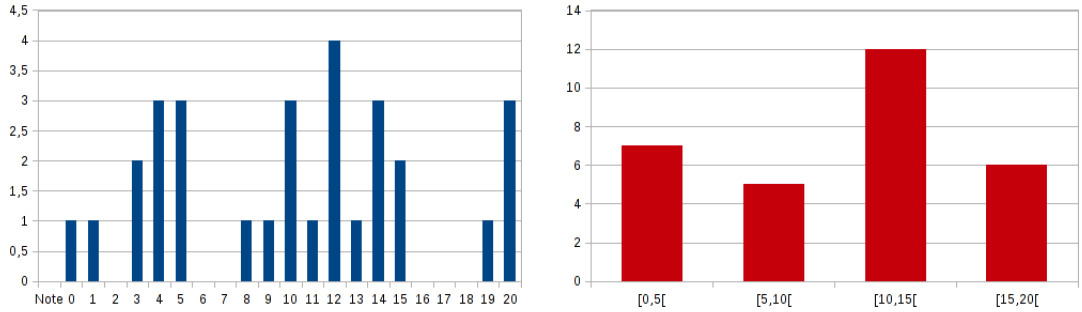
Intervalle	$[0\,,5[$	$[5\,,10[$	$[10\,,15[$	$[15\,,20[$	Total
Effectif	7	5	12	6	30

Définition 1.4 (Représentation graphique de données statistiques).
— Si le caractère est quantitatif discret, on peut utiliser le *diagramme en bâtons* pour représenter graphiquement les données statistiques. Dans un repère orthogonal, pour chaque valeur de la série statistique, on trace un trait vertical dont la hauteur est proportionnelle.
— Si le caractère est quantitatif continue, on peut utiliser le *diagramme en rectangles* pour représenter graphiquement les données statistiques. Dans un repère orthogonal, la base des

rectangles est proportionnelle à la longueur de l'intervalle et la hauteur est proportionnelle à l'effectif.

— Si le caractère est qualitatif, on utilise *les diagrammes circulaires*.

Exemple 1.5. On donne en figure 1, la représentation graphique de la série statistique des classements de notes par ordre croissant et par intervalle de 5 notes.



(a) Représentation graphique du classement des notes par ordre croissant

(b) Représentation graphique du classement par intervalles

FIGURE 1

1.2 Effectif et fréquence

Définition 1.6 (Effectif). L'*effectif* d'une classe ou d'une modalité est le nombre d'individu de cette classe ou de cette modalité. Généralement, on note n_i l'effectif de la classe numéro i (ou de la modalité x_i).

L'*effectif total* est la somme des effectifs de toutes les classes. On le note souvent N.

Exemple 1.7. Dans l'exemple précédent,

$$N = \sum_{i=1}^{5} n_i = n_1 + n_2 + n_3 + n_4 = 7 + 5 + 12 + 8 = 30.$$

Définition 1.8 (Effectif cumulé). L'*effectif cumulé* d'une modalité est la somme des effectifs des modalités qui lui sont inférieurs ou égales.

Définition 1.9 (Fréquence). La *fréquence* notée f_i de la classe i (ou de la modalité x_i) est le rapport $f_i = \frac{n_i}{N}$, la fréquence d'une classe est un nombre de l'intervalle $[0\,,1]$.

Définition 1.10. La *fréquence cumulée* d'une modalité est la somme des fréquences des modalités qui lui sont inférieures ou égales.

Exemple 1.11. Reprenons les données de l'exemple précédent. On a :

Note	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Effectif	1	1	0	2	3	3	0	0	1	1	3	1	4	1	3	2	0	0	0	1	3
Effectif cumul.	1	2	2	4	7	10	10	10	11	12	15	16	20	21	24	26	26	26	26	27	30

(par exemple, 20 personnes ont une note inférieure ou égale à 12) et

Intervalle	$[0\,,5[$	$[5\,,10[$	$[10\,,15[$	$[15\,,20[$	Total
Effectif	7	5	12	6	30
Effectif cumul.	7	12	24	30	30

(par exemple 12 personnes ont en dessous de la moyenne).

1.3 Etendue et mode d'une série statistique

Définition 1.12 (Etendue d'une série statistique). L'*étendue d'une série statistique* est la différence entre la plus petite modalité du caractère et la plus grande modalité.

Exemple 1.13. Reprenons les données de l'exemple précédent. On a :

Note	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Effectif	1	1	0	2	3	3	0	0	1	1	3	1	4	1	3	2	0	0	0	1	3

L'étendue de cette série est $20 - 0 = 20$.

Définition 1.14 (Mode d'une série statistique). Dans le cas continu, on dit qu'une classe est *modale* si elle a le plus grand effectif parmi toutes les classes.

Dans le cas discret, le mode est la valeur de plus grand effectif.

Exemple 1.15. Dans cette série statistique, on a :

Intervalle	$[0\,,5[$	$[5\,,10[$	$[10\,,15[$	$[15\,,20[$	Total
Effectif	7	5	12	6	30

La classe modale de cette série statistique est $[10\,,15[$.

1.4 Paramètre de position

1.4.1 Moyenne

Définition 1.16 (Moyenne). Dans le cas discret, on appelle *moyenne* d'une série statistique d'effectif total N, le réel

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \cdots + n_kx_k}{N}.$$

Exemple 1.17. Reprenons les données de l'exemple précédent. On a :

Note	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Effectif	1	1	0	2	3	3	0	0	1	1	3	1	4	1	3	2	0	0	0	1	3

La moyenne de la série statistique est :

$\overline{x} = \frac{1\times1+0\times2+2\times3+3\times4+3\times5+0\times6+0\times7+1\times8+1\times9+3\times10+11\times1+12\times4+13\times1+14\times3+15\times2+16\times0+17\times0+18\times0+19\times1+3\times20}{30}$

$\overline{x} = \frac{304}{30} \simeq 10,13.$

Remarque 1.18. Pour calculer la moyenne d'une série statistique continu, on prend comme valeur de caractère *le milieu de chaque classe*.

PROPRIÉTÉS 1.19.

1. Si on ajoute à toutes les valeurs d'une série statistique le même nombre b, on augmente la moyenne de cette série par b.
2. Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre a, la moyenne de cette série est aussi multipliée ou divisée par a.
3. Si une population d'effectif N est composée d'une partie d'effectif N_1 et de moyenne $\overline{x}_1$ et d'une autre partie d'effectif N_2 et de moyenne $\overline{x}_2$ alors la moyenne $\overline{x}$ de la population totale

est telle que :

$$\overline{x} = \frac{N_1\overline{x}_1 + N_2\overline{x}_2}{N}.$$

Exemple 1.20. Si, dans une classe, les 15 garçons d'une classe mesurent en moyenne 182 cm et si les 20 filles mesurent en moyenne 168 cm alors la taille moyenne d'un élève de cette classe est égale à

$$\frac{15 \times 182 + 20 \times 168}{15 + 20} = 174 \text{ cm}.$$

1.4.2 Médiane

Définition 1.21. La *médiane* est un paramètre de position qui permet de couper la population étudiée en deux groupes contenant le même nombre d'individus.

Exemple 1.22. On reprend la liste des 30 notes d'un Devoir Surveillé de 2nde d'un lycée parisien :

5	10	12	13	20	14
15	8	3	4	5	1
20	14	12	3	5	19
10	4	9	10	15	12
11	12	14	20	4	0

Pour trouver la médiane, on range les notes par ordre croissant.

0	1	3	3	4	4
4	5	5	5	8	9
10	10	10	11	12	12
12	12	13	14	14	14
15	15	19	20	20	20

Comme il y a 30 notes, la médiane correspond à la moyenne de la 15e note et de la 16e de cette liste, d'où :

0	1	3	3	4	4
4	5	5	5	8	9
10	10	**10**	**11**	12	12
12	12	13	14	14	14
15	15	19	20	20	20

$$, \quad \overline{x} = \frac{10 + 11}{2} = 10{,}5.$$

Remarque 1.23. En général, la moyenne et la médiane d'une série statistique sont deux valeurs différentes.

1.5 Paramètre de dispersion

1.5.1 Associé à la moyenne

Définition 1.24 (Variance). On appelle *variance* d'une série statistique d'effectif total N, et de moyenne $\overline{x}$, le réel :

$$V = \frac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \cdots + n_k(x_k - \overline{x})^2}{N}.$$

Définition 1.25 (Ecart-type). On appelle l'*écart-type* de la série, le réel $\sigma = \sqrt{V}$.

Exemple 1.26. Dans l'exemple des notes, on peut montrer que :

$$V = \frac{7286}{225} \simeq 32{,}115$$

et

$$\sigma = \sqrt{V} = \sqrt{32{,}115} \simeq 5{,}66.$$

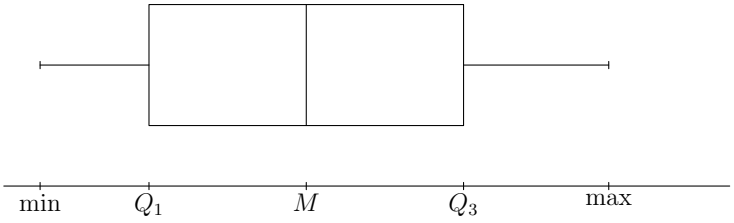
PROPRIÉTÉS 1.27.
1. Si on ajoute à toutes les valeurs d'une série statistique le même nombre b, l'écart-type reste inchangé.
2. Si les valeurs d'une série statistique sont multipliées ou divisées par un même nombre a, l'écart-type est multiplié ou divisé par $|a|$.

1.5.2 Associé à la médiane

Définition 1.28. Soit une série statistique de médiane M dont la liste des valeurs est rangée dans l'ordre croissant. En coupant la liste en deux sous-séries de même effectif,
- on appelle *premier quartile* le réel noté Q_1 égal à la médiane de la sous-série inférieure ;
- on appelle *troisième quartile* le réel noté Q_3 égal à la médiane de la sous-série supérieure.
- L'*écart-interquartile* est égal à $Q_3 - Q_1$.
- $]Q_1\,, Q_3[$ est appelé *intervalle interquartile*.

Remarque 1.29.
- 25% de la population admet une valeur du caractère entre $\min$ et Q_1,
- 25% de la population admet une valeur du caractère entre Q_1 et M,
- 25% de la population admet une valeur du caractère entre M et Q_3,
- 25% de la population admet une valeur du caractère entre Q_3 et $\max$.

Définition 1.30 (Diagramme en boites). Le *diagramme en boites* d'une série se construit de la manière suivante :



Exemple 1.31. On reprend la liste ordonnée de l'exemple précédent :

0	1	3	3	4	4
4	5	5	5	8	9
10	10	10	11	12	12
12	12	13	14	14	14
15	15	19	20	20	20

On peut immédiatement voir que $Q_1 = \frac{4+5}{2} = 4{,}5$ et $Q_3 = \frac{13+14}{2} = 13{,}5$. Donc, on a la construction du diagramme en bâtons suivant (voir la figure 2) :

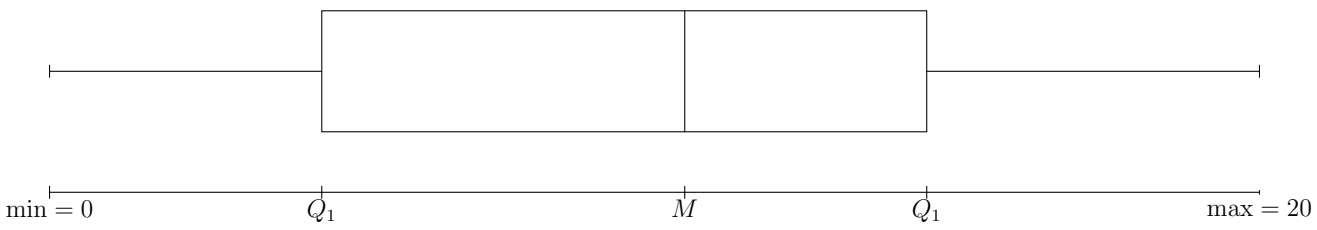


FIGURE 2 – Construction du diagramme en boîte

1.6 Interprétation

Méthode 1.32 (Résumer une série statistique). On peut résumer une série statistique, c'est-à-dire en donner une tendance globale, par

— le couple médiane-écart interquartile, qui n'est pas sensible aux valeurs extrêmes : on le privilégie donc quand on étudie une série dont les valeurs extrêmes sont moins *importantes* ou moins *significatives* que les valeurs *centrales* ;

— le couple moyenne-écart-type, qui est sensible aux valeurs extrêmes : on le privilégie donc quand on étudie une série dont les valeurs extrêmes sont aussi *importantes* ou aussi *significatives* que les autres.

Dans les deux cas, on utilise un indicateur de position (la médiane ou la moyenne) et un indicateur de dispersion (l'écart interquartile ou l'écart-type.

Remarque 1.33. La moyenne $\overline{x}$ et l'écart-type σ s'expriment dans la même unité que les valeurs de la série.

Cela a un sens de parler des intervalles $[\overline{x}-\sigma\,,\overline{x}+\sigma]$, $[\overline{x}-2\sigma\,,\overline{x}+2\sigma]$..., qui sont souvent utilisés en statistiques.

2 Statistiques à deux variables

2.1 Vocabulaire

Définition 2.1.

— Soient x et y deux caractères quantitatifs d'une même population. À chaque individu de la population, on associe un couple $(x_i;y_i)$ où x_i et y_i pour $1\leq i\leq n$ avec n entier naturel sont les valeurs prises respectivement par x et y. L'ensemble de ces couples constitue une *série statistique à deux variables* x et y.

— Dans un repère $(O,\vec{\imath},\vec{\jmath})$, l'ensemble des points M_i de coordonnées $(x_i;y_i)$ est appelé *nuage de points* associé à la série statistique.

— Soit une série statistique à deux variables x et y de moyennes $\overline{x}$ et $\overline{y}$. Le point G de coordonnées $(\overline{x};\overline{y})$ avec :

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n}\quad\text{et}\quad\overline{y}=\frac{y_1+y_2+\cdots+y_n}{n}$$

est appelé le *point moyen* de nuage de points associé à la série statistique.

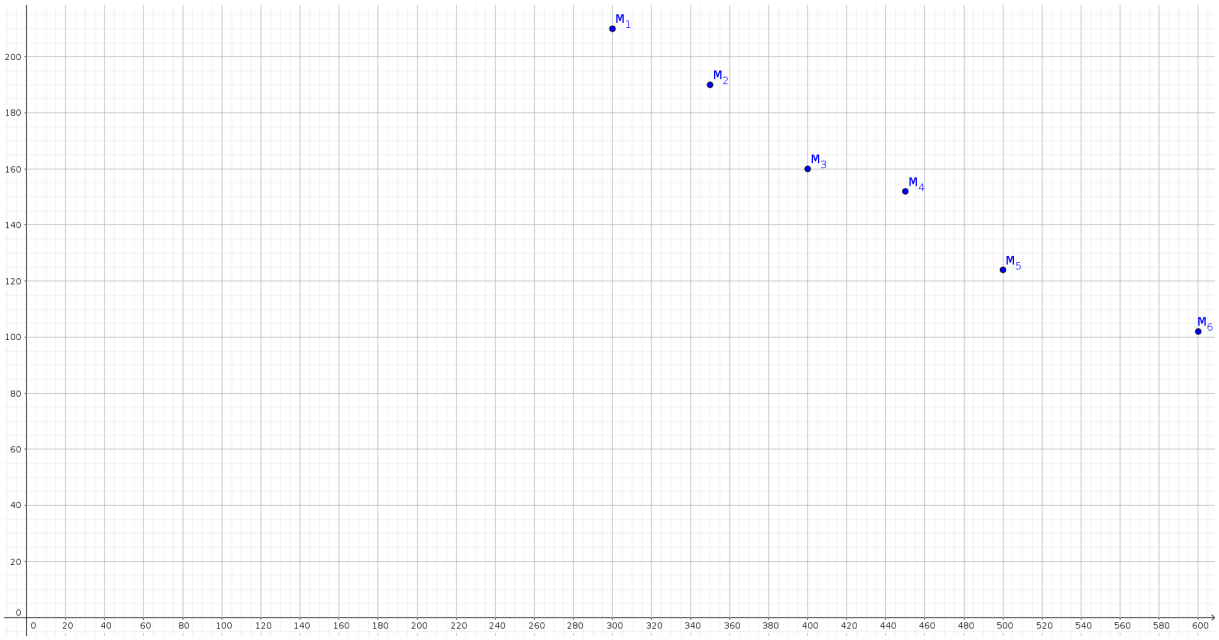
Exemple 2.2. Un magasin réalise une étude sur l'influence du prix de vente sur le nombre de machines à laver vendues au cours d'une année. Le tableau suivant donne les résultats de cette étude :

Prix x_i en euros	300	350	400	450	500	600
Nombre de machines vendues	210	190	160	152	124	102

Le nuage de points associé à cette série est constitu des points M_i pour i allant de 1 à 6 dont les coordonnées sont $(300;210)$, $(350;190)$, …, $(600;102)$.

Le point moyen associé à ce nuage de points est le point G de coordonnées $(\overline{x};\overline{y})$ données par :

$$\overline{x}=\frac{300+350+\cdots+600}{6}=\frac{2600}{6}\approx 433,3\quad\text{et}\quad\overline{y}=\frac{210+190+\cdots+102}{6}=\frac{938}{6}\approx 156,3.$$



2.2 Ajustement d'un nuage de points

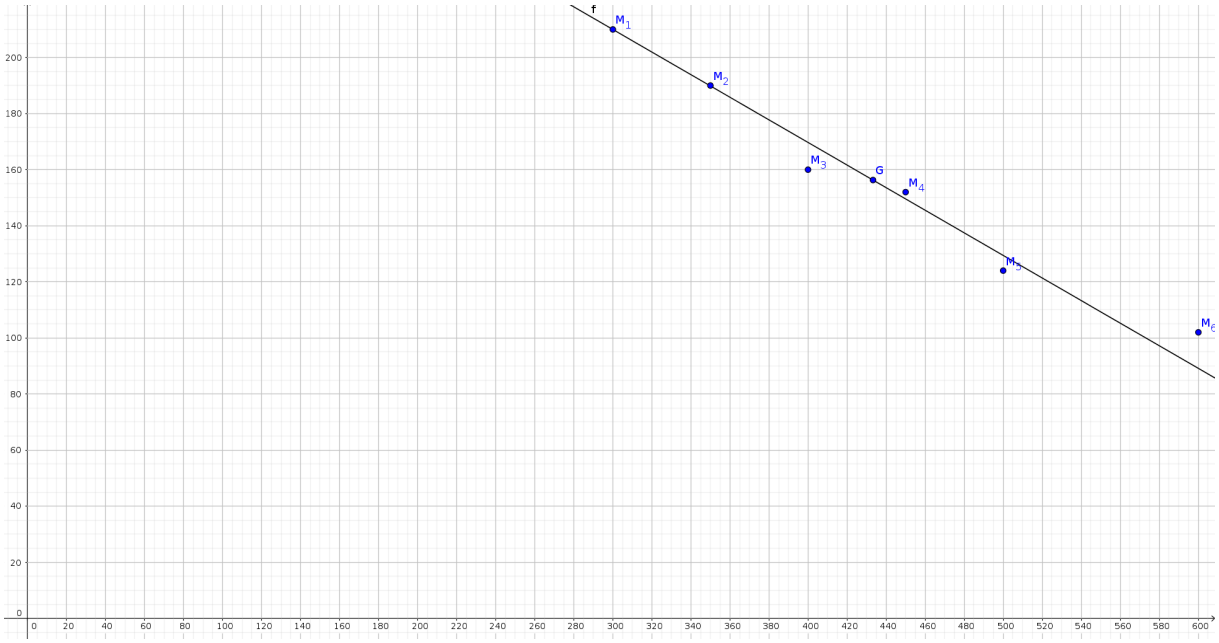
Définition 2.3. Toute droite passant par le point moyen du nuage et « résumant approximativement » le nuage est appelée *droite d'ajustement affine* du nuage de points.

Remarque 2.4. Il existe d'autres types d'ajustement : dans certains cas, on peut observer que visiblement une droite ne convient pas mais que le nuage de points semble être approché par un autre type de courbe, parabole par exemple. En outre, certains nuages peuvent ne pas sembler être approchables par une quelconque courbe auquel cas les deux variables ne sont pas reliées entre elles.

2.3 Détermination d'une équation de droite d'ajustement affine

2.3.1 Méthode graphique au jugé

PROPRIÉTÉ 2.5. On trace « au jugé » une droite passant par le point moyen du nuage qui « semble résumer » le nuage de points. C'est une méthode simple mais qui dépend de la droite tracée.



2.3.2 Méthode de Mayer

PROPRIÉTÉ 2.6. On sépare le nuage en deux sous nuages et on calcule les coordonnées des points moyens des deux sous nuages. La droite de Mayer est al droite passant par ces deux points. On peut montrer qu'elle passe aussi par le point moyen du nuage.

Exemple 2.7. Dans l'exemple précédent, on définit les deux sous nuages constitués des points M_1, M_2 et M_3 pour le premier et des points M_4, M_5 et M_6 pour le second nuage.

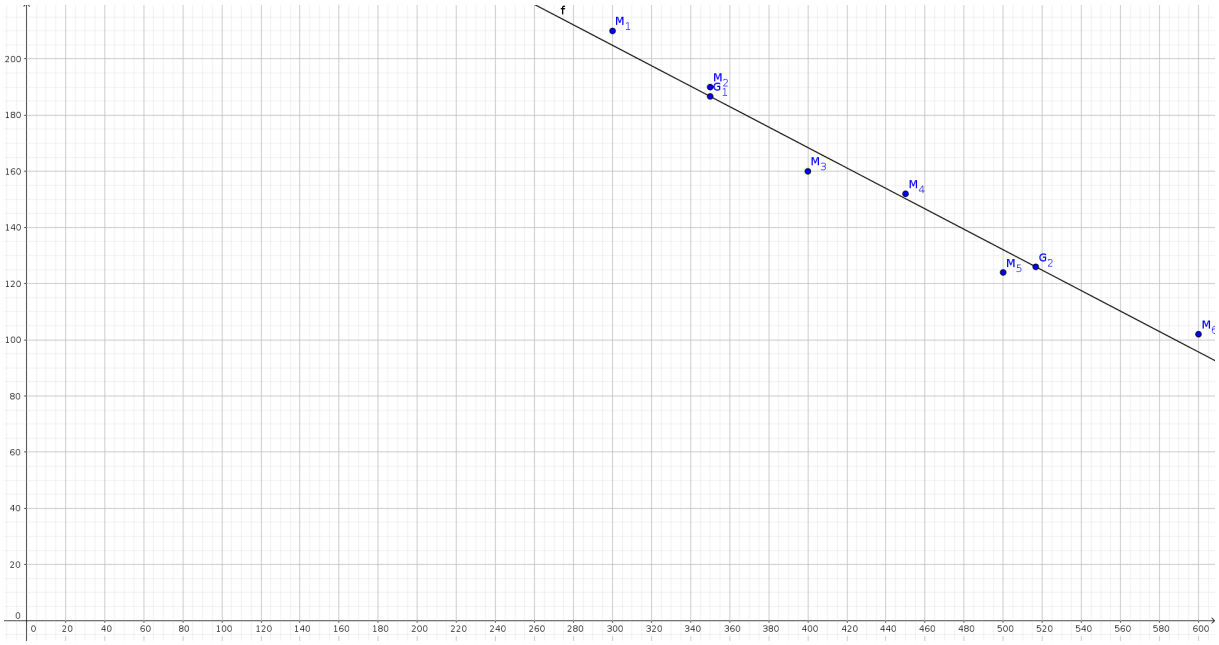
Le point moyen G_1 du premier nuage a donc pour coordonnées :

$$\overline{x}_1 = \frac{300+350+400}{3} = \frac{1050}{3} = 350 \quad \text{et} \quad \overline{y}_1 = \frac{210+190+160}{3} = \frac{560}{3} \approx 186,67.$$

Le point moyen G_2 du deuxième nuage a pour coordonnées :

$$\overline{x}_2 = \frac{450+500+600}{3} = \frac{1550}{3} \approx 516,67 \quad \text{et} \quad \overline{y}_2 = \frac{152+124+102}{3} = \frac{378}{3} = 126.$$

La droite de Mayer est alors la droite (G_1G_2).



2.3.3 Méthode des moindres carrés

PROPRIÉTÉ 2.8. Avec les notations précédentes, étant donné un nuage de n points M_n, il existe une droite passant par le point moyen G et telle que la somme des carrés des écarts (ou *résidus*) $P_1M_1^2 + P_2M_2^2 + \cdots + P_nM_n^2$ soit minimale. Cette droite est appelée *droite de régression de y en x*. On peut montrer que son équation réduite est $y = mx + p$ avec :

$$m = \frac{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})}{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_p - \overline{x})^2}$$

et

$$p = \overline{y} - m\overline{x}.$$

Exemple 2.9. On considère la série statistique à deux variables suivante :

x_i	5	10	15	20	25
y_1	13	23	34	44	50

On calcule $\overline{x}$ et $\overline{y}$:

$$\overline{x} = \frac{5 + 10 + 15 + 20 + 25}{5} = 15 \quad \text{et} \quad \overline{y} = \frac{13 + 23 + 34 + 44 + 50}{5} = 32,8.$$

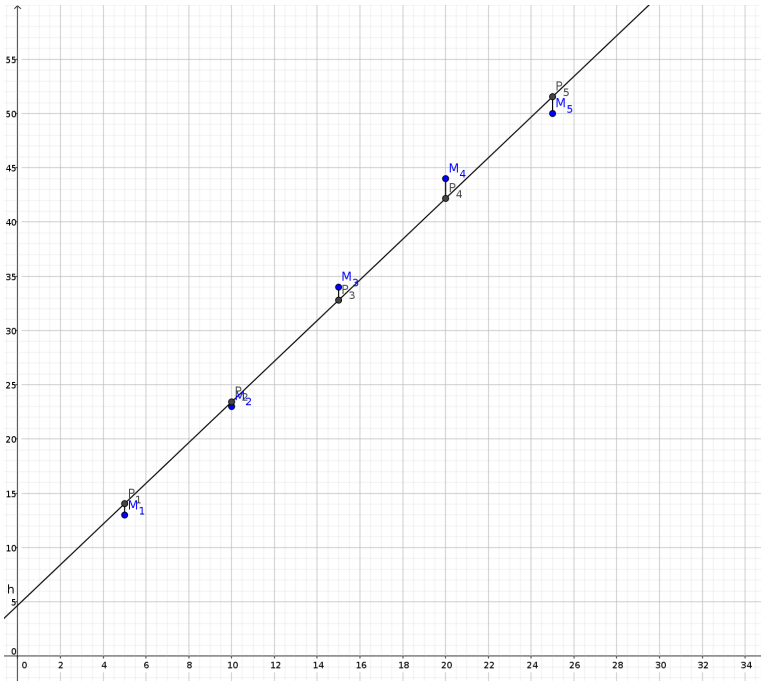
On peut dresser le tableau suivant :

i	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
1	5	13	-10	$-19,8$	198	100
2	10	23	-5	$-9,8$	49	25
3	15	34	0	$1,2$	0	0
4	20	44	5	$11,2$	$57,5$	25
5	25	50	10	$17,2$	172	100

On a ainsi :

$$m = \frac{198 + 49 + 0 + 57,5 + 172}{100 + 25 + 0 + 25 + 100} = \frac{468,5}{250} = 1,875$$

et $p = 32,8 - 15 \times 1,875 = 4,675$.



On peut obtenir la droite de regression linéaire avec la TI-82 en allant dans le menu `Stats > Edit`. On entre les valeurs x_i dans la colonne `L1` et les valeurs y_i dans la colonne `L2`. Puis dans le mode prinicipal, on va dans `Stat > Calc`, on choisit `LinReg(ax+b)` et on tape à l'écran les deux listes `L1` et `L2`.

Compléments :

Définition 2.10 (Covariance). On appelle *covariance* du couple (X, Y), le réel :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Définition 2.11 (Coefficient de corrélation linéaire). On appelle *coefficient de corrélation linéaire*, le réel :

$$r = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

PROPRIÉTÉ 2.12.

1. $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ d'après la formule de Koenig.
2. La covariance est une forme bilinéaire symétrique positive.
3. $|\mathrm{Cov}(X,Y)| \leq \sigma(X)\sigma(Y)$ et donc $|\rho(X,Y)| \leq 1$.
4. $|r| = |\rho(X,Y)| = 1$ si et seulement si les points du nuages sont alignés.

THÉORÈME 2.13 (MÉTHODE DES MOINDRES CARRÉS). La droite d'équation

$$y - \overline{Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)^2}(x - \overline{X})$$

passe par le point moyen et est la droite d'équation réduite de la forme $y = ax + b$ qui minimise la somme :

$$\sum_{i=1}^{n} f_i(ax_i + b - y_i)^2$$

pour $(a, b) \in \mathbb{R}^2$. Autrement dit :

$$a = \frac{\mathrm{Cov}(X, Y)}{\sigma(X)^2} \quad \text{et} \quad b = \overline{Y} - \overline{X}\frac{\mathrm{Cov}(X, Y)}{\sigma(X)^2}$$

réalisent ce minimum sur $\mathbb{R}^2$.

Démonstration du théorème 2.13, première méthode. $\Diamond$ On pose

$$S(a, b) = \sum_{i=1}^{n} [y_i - ax_i - b]^2$$

et on introduit $z = y - ax - b$, on peut alors réécrire $S(a, b)$ comme

$$S(a, b) = \sum_{i=1}^{n} z_i^2.$$

Or, on sait que

$$\mathrm{Var}(z) = \frac{1}{n}\sum_{i=1}^{n}(z_i - z)^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2 - \overline{z}^2$$

et, par linéarité de la moyenne $\overline{z} = \overline{y} - a\overline{x} - b$. Donc, minimiser $S(a, b)$ revient à minimiser $\sum z_i^2 = n(\mathrm{Var}(z) + \overline{z}^2)$.

On va donc minimiser $n\,\mathrm{Var}(z)$. On a :

$$z_i - \overline{z} = y_i - ax_i - b - (\overline{y} - a\overline{x} - b) = (y_i - \overline{y}) - a(x_i - \overline{x}).$$

D'où :

$$\begin{aligned} n\,\mathrm{Var}(z) &= \sum_{i=1}^{n}(z_i - \overline{z}) = \sum_{i=1}^{n}[(y_i - \overline{y}) - a(x_i - \overline{x})]^2 \\ &= \sum_{i=1}^{n}[(y_i - \overline{y})^2 - 2a(x_i - \overline{x})(y_i - \overline{y}) + a^2(x_i - \overline{x})^2] \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2a\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) + a^2\sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned}$$

Or

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

On a finalement :

$$\mathrm{Var}(z) = \mathrm{Var}(x)a^2 - 2\,\mathrm{Cov}(x, y) + \mathrm{Var}(y).$$

On reconnaît un trinôme du second degré. On va l'écrire sous forme canonique :

$$\begin{aligned} \mathrm{Var}(z) &= \left(\sigma(x)a - \frac{\mathrm{Cov}(x, y)}{\sigma(x)}\right)^2 + \mathrm{Var}(y) - \left(\frac{\mathrm{Cov}(x, y)}{\sigma(x)}\right)^2 \\ &= \left(\sigma(x)a - \frac{\mathrm{Cov}(x, y)}{\sigma(x)}\right)^2 + \frac{\mathrm{Var}(x)\,\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2}{\mathrm{Var}(x)}. \end{aligned}$$

Ainsi, $\mathrm{Var}(z)$ est minimal lorsque $\left(\sigma(x)a - \frac{\mathrm{Cov}(x,y)}{\sigma(x)}\right)^2 = 0$, c'est-à-dire $a = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$ et le minimum de $\mathrm{Var}(z)$ est

$$\frac{\mathrm{Var}(x)\,\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2}{\mathrm{Var}(x)}.$$

On va maintenant minimiser $\overline{z}^2$. On a : $\overline{z} = \overline{y} - a\overline{x}$. Donc $\overline{z}$ est minimal si $b = \overline{y} - a\overline{x}$ et le minimum de $\overline{z}$ est 0.

D'où la droite de régression de y en x a pour équation $y = ax + b$ où

$$a = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \quad \text{et} \quad b = \overline{y} - a\overline{x}.$$

□

Démonstration du théorème 2.13, seconde méthode. ◊ Soit f la fonction définie sur $\mathbb{R}^2$ par :

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

C'est une fonction polynôme de degré 2 que l'on peut écrire sous la forme :

$$\begin{aligned} f(a,b) = \left(\sum_{i=1}^{n} f_i x_i^2\right) a^2 + b^2 + 2\left(\sum_{i=1}^{n} f_i x_i\right) ab \\ - 2\left(\sum_{i=1}^{n} f_i x_i y_i\right) a - 2\left(\sum_{i=1}^{n} f_i y_i\right) b + \sum_{i=1}^{n} f_i y_i^2 \\ f(a,b) = \overline{X^2}a^2 + b^2 + 2\overline{X}ab - 2\overline{XY}a - 2\overline{Y}b + \overline{Y^2}. \end{aligned}$$

Les dérivées partielles sont données par :

$$\frac{\partial f}{\partial a}(a,b) = 2\overline{X^2}a + 2\overline{X}b - 2\overline{XY} \quad \text{et} \quad \frac{\partial f}{\partial b}(a,b) = 2\overline{X}a + 2b - 2\overline{Y}$$

Elles s'annulent simultanément en l'unique point critique défini par :

$$\begin{aligned} a_0 &= \frac{-\overline{X}\cdot\overline{Y} + \overline{XY}}{\overline{X^2} - \overline{X}^2} = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)^2} \\ b_0 &= \frac{\overline{XY}\cdot\overline{X} + \overline{Y}\cdot\overline{X^2}}{\overline{X^2} - \overline{X}^2} = \overline{Y} - \overline{X}\frac{\mathrm{Cov}(X,Y)}{\sigma(X)^2}. \end{aligned}$$

Les dérivées partielles secondes sont données par :

$$\frac{\partial^2 f}{\partial a^2}(a,b) = 2\overline{X^2} \quad \text{et} \quad \frac{\partial^2 f}{\partial a \partial b}(a,b) = 2\overline{X}$$

$$\frac{\partial^2 f}{\partial b^2}(a,b) = 2.$$

Avec les notations de Monge, au point (a_0, b_0), on a :

$$rt - s^2 = 4\overline{X^2} - 4\overline{X}^2 = 4\sigma(X)^2 > 0$$

ce qui assure qu'on a bien un minimum local en (a_0, b_0). De plus, un développement limité à l'ordre 2 au voisinage de (a_0, b_0) donne :

$$\begin{aligned} f(a,b) = f(a_0,b_0) + \frac{1}{2}\frac{\partial^2 f}{\partial a^2}(a_0,b_0)(a-a_0)^2 \\ + \frac{1}{2}\frac{\partial^2 f}{\partial a \partial b}(a_0,b_0)(a-a_0)(b-b_0) + \frac{1}{2}\frac{\partial^2 f}{\partial b^2}(a_0,b_0)(b-b_0)^2 \geq f(a_0,b_0) \end{aligned}$$

puisque les termes d'ordre supérieur sont nuls (fonction polynôme de degré 2) et la forme quadratique est strictement positive ($rt - s^2 > 0$) et ainsi on a bien un maximum global sur $\mathbb{R}^2$.

La droite d'équation réduite $y = a_0 x + b_0$ est la droite proposée dans l'énoncé et passe clairement par le point moyen de la série statistique. □

Définition 2.14 (Droite d'ajustement). — La droite définie ci-dessus est appelée *droite d'ajustement* (ou *droite de régression* de Y en X.

— La somme

$$\sum_{i=1}^{n} f_i(ax_i + b - y_i)^2$$

est appelée *résidu quadratique*.

Remarques 2.15. 1. La droite d'équation

$$x - \overline{X} = \frac{\mathrm{Cov}(X,Y)}{\sigma(Y)^2}(y - \overline{Y})$$

minimise la somme

$$\sum_{i=1}^{n} f_i(ay_i + b - x_i)^2$$

et s'appelle droite d'ajustement de X en Y.

2. Notons $Z = (1, \ldots, 1)$ le caractère constant égal à 1 sur la population commune à X et Y. Ajuster Y en X revient à considérer le projeté orthogonal de Y sur le sous-espace (X, Z) de l'espace euclidien $\mathbb{R}^n$ pour le produit scalaire canonique.
3. Lorsque $|r| = |\rho(X,Y)| > 0{,}9$ (valeur dépendant des auteurs et des besoins), on considère que l'ajustement affine de Y en X est satisfaisant (sinon, il faut déterminer un autre type d'ajustement).