Statistiques à deux variables – Ajustements linéaires

Jusqu'à présent, nous n'avons étudié que des séries statistiques à une seule variable : les tailles des individus d'une population, les cotes obtenues par des étudiants à un examen, le nombre de personnes par ménage dans une ville donnée , etc.
Nous allons maintenant étudier simultanément deux variables d'une population donnée. Le but poursuivi est de chercher une correspondance entre les deux variables : existe-t-il un lien entre la taille et la masse des personnes, entre le prix d'un article et le nombre de ventes, etc.

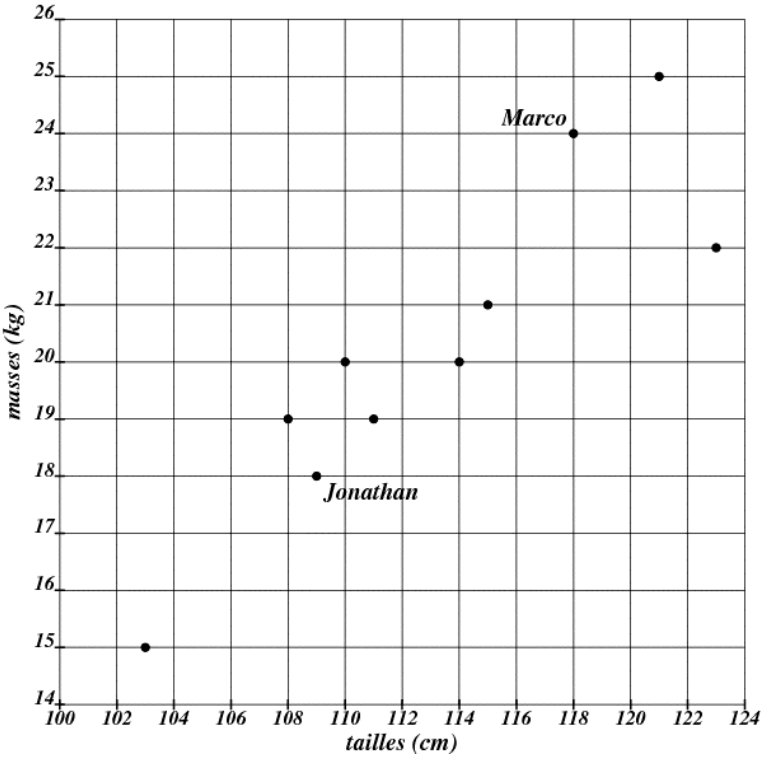
Exemple

Considérons un groupe de dix garçons de six ans. Pour chaque enfant, on a mesuré deux variables : la taille (en centimètres) et la masse (en kilos).

	Tailles	Masses
Alan	121	25
Bruno	123	22
Kevin	108	19
Marco	118	24
Bryan	111	19

	Tailles	Masses
Jonathan	109	18
Guillaume	114	20
Mourad	110	20
Giuseppe	115	21
Gauthier	103	15

Afin de mieux voir le lien entre les deux variables, reportons ces données sur un graphique avec les tailles en abscisses et les masses en ordonnées. A chaque enfant correspond ainsi un point sur le graphique. Les dix points constituent un *nuage de points expérimentaux*.



En observant le nuage, on peut supposer qu'il existe une relation linéaire entre la masse et la taille. Attention ! Cette hypothèse n'est pas toujours réaliste !

Réaliser une régression linéaire, ou un ajustement linéaire, consiste alors à trouver l'équation de la droite qui passe « le plus près possible » des points du nuage.

Première méthode : au jugé

Tracer une droite qui passe « le plus près possible » des dix points du nuage. Utiliser cette droite pour estimer, par exemple, la masse d'un enfant mesurant 105(cm).

Deuxième méthode : la droite de MAYER

C'est une des méthodes d'ajustement les plus simples. Elle consiste à :

1. découper le nuage de points en deux sous-nuages distincts et de taille identique (à une unité près) ; en pratique, on prend la première moitié des points (ceux dont les abscisses sont les plus petites) pour former le premier sous-nuage, et la deuxième moitié pour former l'autre sous-nuage ; s'il y a un nombre impair de points, on place celui du milieu dans le sous-nuage que l'on veut ;
2. calculer les points moyens respectifs de chaque sous-nuage (G_1 et G_2) ;
3. déterminer l'équation de la droite $d = G_1G_2$ appelée droite de MAYER (cette droite passe aussi par le point moyen G du nuage complet).

Appliquons cette démarche à notre exemple.

1. Premier sous-nuage : $(103,15)$, $(108,19)$, $(109,18)$, $(110,20)$ et $(111,19)$.

   Second sous-nuage : $(114,20)$, $(115,21)$, $(118,24)$, $(121,25)$ et $(123,22)$.

2. $$G_1\left(\frac{103+108+109+110+111}{5},\frac{15+19+18+20+19}{5}\right)=(108.2,18.2)$$

   $$G_1\left(\frac{114+115+118+121+123}{5},\frac{20+21+24+25+22}{5}\right)=(118.2,22.4)$$

3. - Nous avons $d \equiv y = mx + p$ avec $m = \frac{22.4-18.2}{118.2-108.2} = \frac{4.2}{10} = 0.42$;
   - exprimons que le point G_1 appartient à d : $18.2 = 0.42 \times 108.2 + p$; nous trouvons ainsi $p = 18.2 - 0.42 \times 108.2 = -27.244$;
   - la droite de MAYER a donc pour équation $\boxed{d \equiv y = 0.42x - 27.244}$ (graphique à la page suivante) ;
   - on vérifie que le point $G(113.2,20.3)$ appartient bien à cette droite.

L'équation obtenue permet de faire des *estimations*. Voici deux exemples:

- quelle serait la masse d'un enfant mesurant 105(cm) ?

  $y = 0.42 \times 105 - 27.244 = 16.856$

  On peut donc s'attendre à une masse d'environ 17(kg).

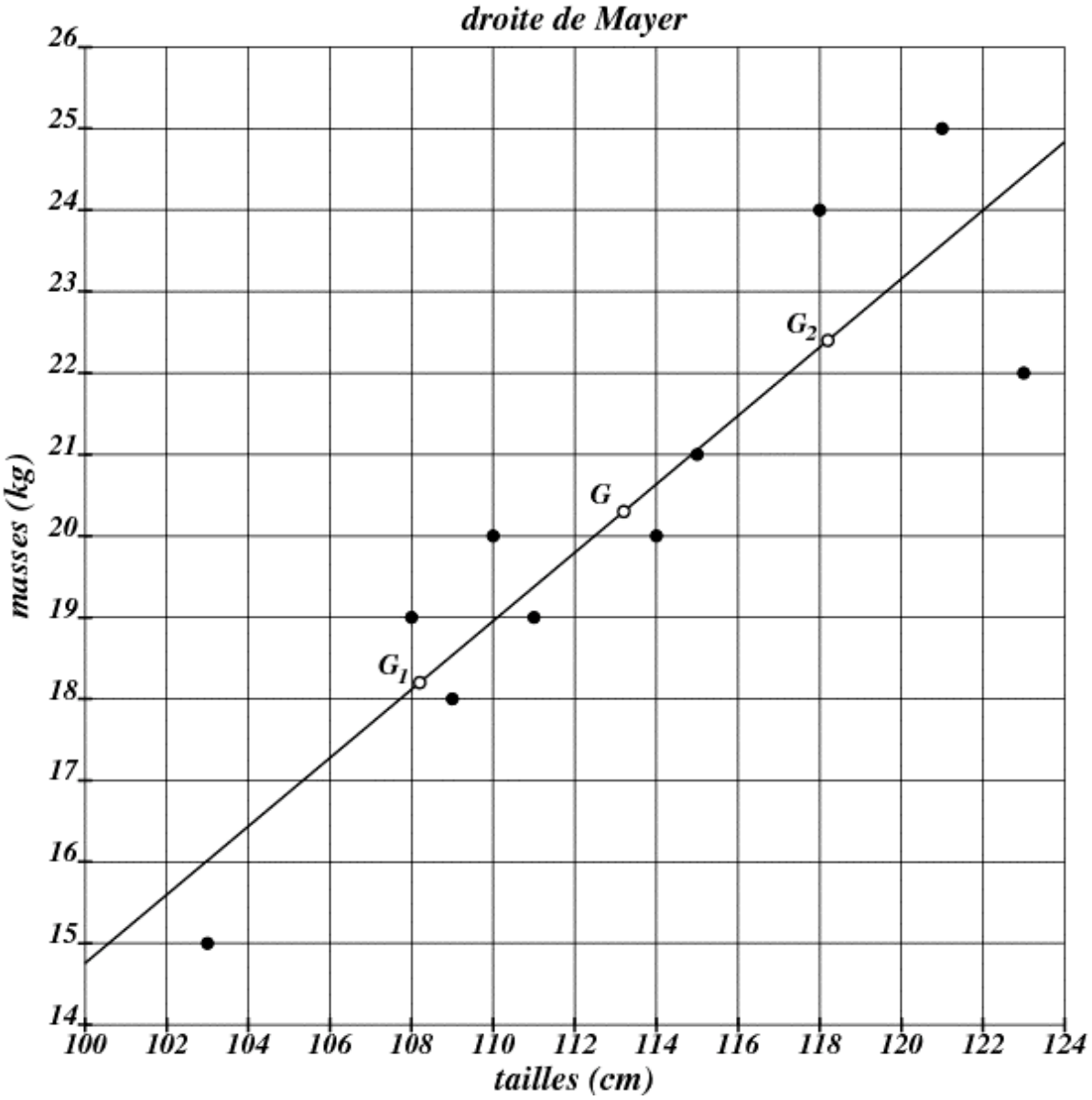
- quelle serait la taille d'un enfant dont la masse est de 23(kg) ?

  $$23 = 0.42 \cdot x - 27.244 \rightarrow x = \frac{23 + 27.244}{0.42} \approx 119.63$$

  On peut donc s'attendre à une taille d'environ 120(cm)

Insistons sur le fait que ces résultats ne sont pas à prendre au pied de la lettre. Il ne donnent que des ordres de grandeur. En particulier, il faut être prudent lors des extrapolations car même si une variation semble linéaire pour les valeurs dont on dispose, rien ne dit que la linéarité sera toujours de mise pour des valeurs plus grandes des abscisses.

Pour revenir à notre exemple, la droite obtenue semble rendre compte de façon raisonnable du lien existant entre la masse et la taille.



Troisième méthode : la méthode des moindres carrés

Cette méthode d'ajustement linéaire est celle qui est programmée dans la plupart des calculatrices scientifiques, ainsi que dans des logiciels comme GRAPHMATICA et EXCEL. Si les coordonnées des points du nuage sont (x_i, y_i) , il s'agit de déterminer une droite d'équation $y = mx + p$ telle que la somme des carrés des écarts entre les valeurs des ordonnées mesurées et les valeurs trouvées sur la droite soit minimale : $\sum (y_i - (mx_i + p))^2$ est minimale. Cette droite passe toujours par le point moyen du nuage $G(\overline{x}, \overline{y})$.

La pente de la droite est donnée par la formule suivante (admise) :

$$\boxed{m = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum\left(x_i - \overline{x}\right)^2}}$$

Une fois que l'on a trouvé la pente, il suffit d'utiliser les coordonnées de G pour trouver la valeur de p.

Un exemple permettra d'y voir plus clair. Reprenons celui des tailles et des masses des dix enfants. Nous connaissons déjà le point moyen $G(113.2,20.3)$, c'est-à-dire $\overline{x} = 113.2$ et $\overline{y} = 20.3$. Organisons le calcul de m au moyen d'un tableau.

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$	$\left(x_i - \overline{x}\right)^2$
103	15	-10.2	-5.3	54.06	104.04
108	19	-5.2	-1.3	6.76	27.04
109	18	-4.2	-2.3	9.66	17.64
110	20	-3.2	-0.3	0.96	10.24
111	19	-2.2	-1.3	2.86	4.84
114	20	0.8	-0.3	-0.24	0.64
115	21	1.8	0.7	1.26	3.24
118	24	4.8	3.7	17.76	23.04
121	25	7.8	4.7	36.66	60.84
123	22	9.8	1.7	16.66	96.04
				Σ = 146.40	Σ = 347.60

Par conséquent, $m = \frac{146,40}{347,60} \approx 0,4212$.
Calculons p via les coordonnées de G : $20,3 \approx 0,4212 \times 113,2 + p \rightarrow p \approx -27,3769$.

La droite d'ajustement a donc pour équation :

$$\boxed{d \equiv y = 0,4212 \cdot x - 27,3769}$$

Pour notre exemple, la droite obtenue par la méthode des moindres carrés a une équation quasi identique à celle de la droite de MAYER. Ce n'est pas toujours le cas !

Coefficient de corrélation

Il s'agit d'un réel r donné par la formule : $\boxed{r = \frac{\frac{1}{n} \cdot \sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sigma_x \cdot \sigma_y}}$.

Dans cette formule, $\overline{x}$ et $\overline{y}$ sont les moyennes respectives des valeurs de x et de celles de y, tandis que σ_x et σ_y sont les écarts types respectifs des valeurs de x et de celles de y.
On peut montrer que le coefficient de corrélation est un réel compris entre -1 et 1.
Plus il est proche de 1 en valeur absolue, plus la corrélation entre y et x est forte, c'est-à-dire plus la variable y dépend linéairement de la variable x.
La relation fonctionnelle est parfaite entre y et x si $r = 1$ ou $r = -1$.
Plus le coefficient est proche de 0, plus la corrélation est faible, voire inexistante (si $r = 0$, les valeurs de y ne dépendent pas de celles de x).

Si l'on dispose déjà de la pente m de la droite de régression, on peut utiliser la formule suivante pour calculer le coefficient de corrélation :

$$\boxed{r = m \cdot \frac{\sigma_x}{\sigma_y}}$$

Revenons à notre exemple.

Nous savons déjà que $m \approx 0{,}4212$. Les écarts types de x et de y sont les suivants (pour les calculer, revoir le cours de statistique descriptive à une variable) :

$$\sigma_x \approx 5{,}8958 \text{ et } \sigma_y \approx 2{,}7586 .$$

Utilisons la seconde formule pour calculer le coefficient : $r \approx 0{,}4212 \times \frac{5{,}8958}{2{,}7586} \approx 0{,}9001$.

La corrélation est forte entre la masse et la taille des dix garçons de l'échantillon. L'aspect du nuage de points le laissait présager !

Exercices

1. Le tableau ci-dessous donne l'évolution du nombre de déchetteries (centres de collecte traitements des déchets) en Bretagne entre 1991 et 1996.

Années x_i	1991	1992	1993	1994	1995	1996
Nombre de déchetteries y_i	18	33	53	69	83	95

   a) Représenter ces données dans un repère orthogonal (2(cm) pour une année en abscisse et 1(cm) pour 5 déchetteries en ordonnée).

   b) Déterminer l'équation de la droite de Mayer associée à ce nuage de points.

   c) Vérifier que le point moyen du nuage appartient bien à cette droite.

   d) Utiliser la droite de Mayer pour prévoir le nombre de déchetteries en 2005 .

   e) Selon la droite de Mayer, en quelle année la Bretagne comptait-elle 200 déchetteries ?

---

2. Un vendeur d'appareils électroménagers propose à ses clients huit modèles de lave-linge. Le tableau ci-dessous donne, pour chacun des huit modèles, le prix de vente à l'unité en euros et le nombre d'appareils vendus au cours du mois précédent.

Prix x_i	250	300	350	400	500	550	600	700
Nombre d'appareils vendus y_i	105	95	80	76	62	56	49	29

   a) Représenter ces données dans un repère orthogonal (2(cm) pour 100 euros en abscisse et 1(cm) pour 10 appareils vendus en ordonnée).

   b) Déterminer l'équation de la droite de Mayer associée à ce nuage de points.

   c) Utiliser la droite de Mayer pour prévoir le nombre d'appareils vendus si leur prix est de 650 euros.

3. Créé en 1971, le PAC (Port Autonome de Charleroi) est constitué de 24 zones portuaires implantées le long de la Sambre et du canal Charleroi-Bruxelles.
Voici le tableau donnant, année après année, le tonnage total transporté via le PAC.

Année	Tonnage
1971	529082
1972	554970
1973	901865
1974	1305058
1975	989560
1976	1264517
1977	1627602
1978	2006768
1979	2192472
1980	2025042

Année	Tonnage
1981	1988203
1982	1754933
1983	2041570
1984	2535821
1985	2683679
1986	2533561
1987	2712692
1988	3202213
1989	3362878
1990	3073204

Année	Tonnage
1991	3208894
1992	3230451
1993	2737043
1994	3365424
1995	3699490
1996	3795893
1997	4503952
1998	4306313
1999	4746972
2000	5401444

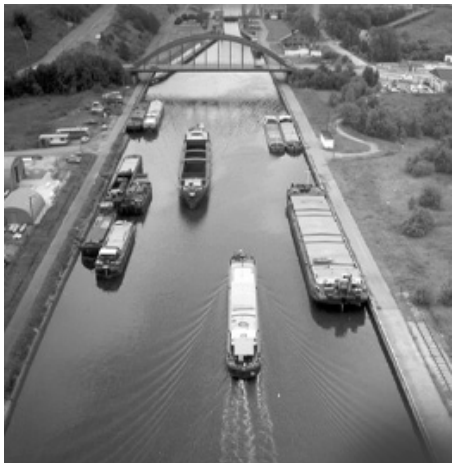
Année	Tonnage
2001	5571727
2002	5597610
2003	6045867
2004	5977990
2005	5949245
2006	7075905
2007	6626363

a) Encoder ces données dans un tableur afin de réaliser un graphique avec les années en abscisse et les tonnages en ordonnée.

b) Découper le nuage en deux sous-nuages : de 1971 à 1989 pour le premier, et de 1990 à 2007 pour le second.
Calculer le point moyen de chaque sous- nuage.

c) Déterminer l'équation de la droite de Mayer associée à ce nuage de points.

d) D'après la droite de Mayer, donner une estimation du tonnage total des marchandises transportées par le PAC en 2010.



e) En quelle année le PAC devrait-il assurer le transport de 10 millions de tonnes de marchandises ?

f) Utiliser le tableur afin de réaliser un ajustement linéaire par la méthode des moindres carrés. Comparer avec la droite de Mayer.

g) Calculer le coefficient de corrélation entre le tonnage y_i et l'année x_i .

L'entreprise sidérurgique Carinox (ArcelorMittal) bénéficie de la proximité de la zone portuaire de La Praye à Châtelet.

4. Sur une même verticale, la pression atmosphérique diminue lorsque l'altitude augmente, conformément au tableau suivant (x : altitude en (km) ; y : pression en (cm) de mercure).

x_i	0	1	2	4	6	10
y_i	76	67	59	46	35	20

a) Représenter le nuage de points sur un graphique.

b) Déterminer l'équation de la droite de Mayer associée à ce nuage de points.

c) A l'aide de la droite de Mayer, déterminer à quelle altitude la pression n'est plus que de 40(cm) de mercure.

d) Réaliser un ajusteement linéaire par la méthode des moindres carrés. Comparer avec la droite de Mayer.

e) Calculer le coefficient de corrélation entre la pression y_i et l'altitude x_i .

---

5. Le tableau suivant donne l'évolution du pourcentage de moins de vingt ans dans la population française entre 1970 et 1995 .

Année x_i	1970	1980	1990	1995
% de moins de 20 ans y_i	33,2	30,6	27,8	26,1

a) Représenter le nuage de points sur un graphique.

b) Déterminer l'équation de la droite de Mayer associée à ce nuage de points.

c) Si l'évolution enregistrée depuis 1970 se poursuit conformément à la droite de Mayer, quel devrait être le pourcentage de moins de vingt ans dans la population française de 2009 ?

---

6. On possède des specimens fossiles d'un animal disparu et ces specimens sont de tailles différentes. On estime que si ces animaux appartiennent à la même espèce, il doit exister une relation linéaire entre la longueur de deux de leurs os : le fémur et l'humérus.
Voici les données de ces longueurs en centimètres.

Fémur	38	56	59	64	74
humérus	41	63	70	72	84

a) Représenter le nuage de points sur un graphique. Pensez-vous que les 5 specimens peuvent appartenir à la même espèce et ne différer en taille que parce que certains sont plus jeunes que d'autres ?

b) Déterminer l'équation de la droite des moindres carrés associée à ces données.

c) Calculer le coefficient de corrélation. Commenter.