Statistiques descriptives

1 Statistiques à deux variables

1.1 Définition - Nuage de points

Le principe est le suivant : on étudie à la fois deux caractères sur une même population. L'objectif est le plus souvent de déterminer s'il y a un lien statistique entre les deux, ce qu'on appelle une **corrélation**.

On définit alors une série statistique à deux variables x et y, prenant les valeurs x_1, ..., x_p et y_1, ..., y_q. À chaque couple de valeurs $(x_i\,;y_j)$, on peut faire correspondre son effectif n_{ij}.

Définition 1
Dans le plan muni d'un repère, on peut associer à chaque couple $(x_i;y_i)$ de la série statistique, le point M_i de coordonnées $(x_i;y_i)$. L'ensemble des points M_i s'appelle le **nuage de points** représentant la série statistique.

Définition 2
On appelle point moyen d'un nuage le point G de coordonnées $(\overline{x}\,;\overline{y})$.

Exemple : on mesure le diamètre et la masse d'une série de pièces produites par une machine ; on obtient le tableau suivant :

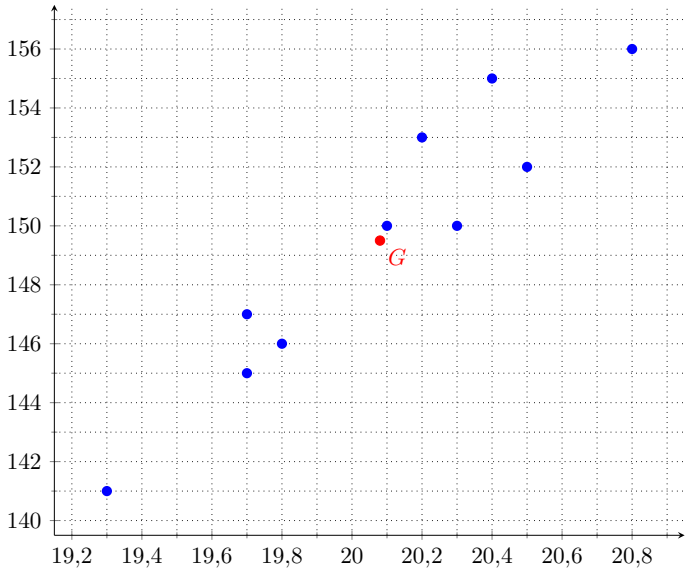
Pièce n°	1	2	3	4	5	5	7	8	9	10
Diamètre (cm)	19,3	20,5	19,8	19,7	20,2	20,4	20,3	19,7	20,8	20,1
Masse (g)	141	152	146	147	153	155	150	145	156	150

On va donc placer les points de coordonnées $(19,3\,;141)$, $(20,5\,;152)$, etc.

Calculons les coordonnées du point moyen :

$$\overline{x} = \frac{19,3+20,5+19,8+19,7+20,2+20,4+20,3+19,7+20,8+20,1}{10} = 20,08$$

$$\overline{y} = \frac{141+152+146+147+153+155+150+145+156+150}{10} = 149,5$$



2 Ajustement affine

2.1 Méthode graphique

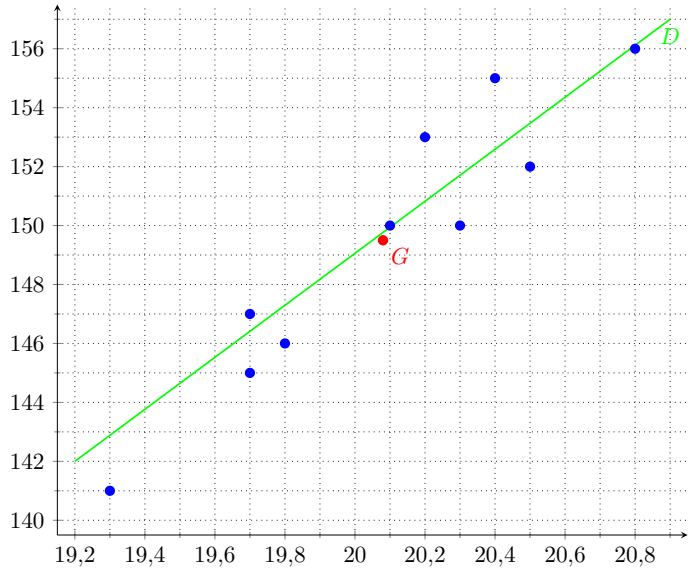
Lorsque les points M_i semblent à peu près alignés, on cherche à construire une droite qui représente au mieux la forme du nuage, c'est à dire qui passe « au plus près » de chacun des points.

L'équation de cette droite constitue une formule qui relie, de façon approchée, les deux variables x et y.

a. Ajustement à la règle :

On trace au jugé une droite D en s'efforçant d'équilibrer le nombre de points situés de part et d'autre. Ensuite on détermine son équation par lecture graphique.

Illustration, à partir de l'exemple précédent :



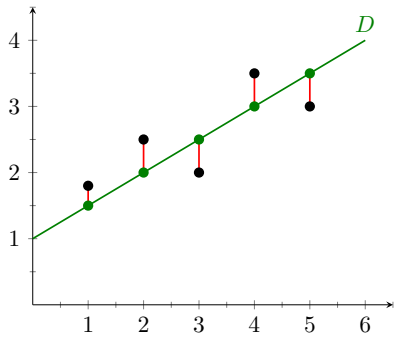
b. Ajustement par la méthode de Mayer :

On partage le nuage de points en deux nuages d'affectif à peu près égal, de part et d'autre du point moyen. On détermine les coordonnées des points moyens G_1 et G_2 de chacun de ces nuages.

La droite (G_1G_2), appelée droite de **Mayer**, constitue en général une bonne droite d'ajustement.

2.2 Méthode des moindres carrés - Droite de régression

Définition 3
Soit D une droite d'équation $y = ax + b$. Appelons P_i le point d'abscisse x_i situé sur la droite D (son ordonnée est donc $ax_i + b$).
On appelle **droite de régression de y en x** la droite D choisie de telle sorte que la somme des $M_iP_i^2$ soit minimale.



Sur le dessin, c'est la droite qui minimise la somme des carrés des distances en rouge.

Les points M_i ont pour coordonnées $(x_i\,;y_i)$; les points P_i ont pour coordonnées $(x_i\,;ax_i+b)$.

La quantité minimisée est donc la somme des $[y_i-(ax_i+b)]^2$.

Propriété : On démontre que la droite de régression de y en x passe toujours par le point moyen du nuage.

Expression des coefficients :

Définition 4
On appelle **covariance** de la série statistique à deux variables x et y le nombre :

$$Cov(x,y)=\sigma_{xy}=\frac{\text{Somme des } (x_i-\bar{x})(y_i-\bar{y})}{\text{effectif total}}$$

Propriété 1
Le coefficient directeur de la droite de régression est : $a=\dfrac{Cov(x\,;y)}{Var(x)}$.

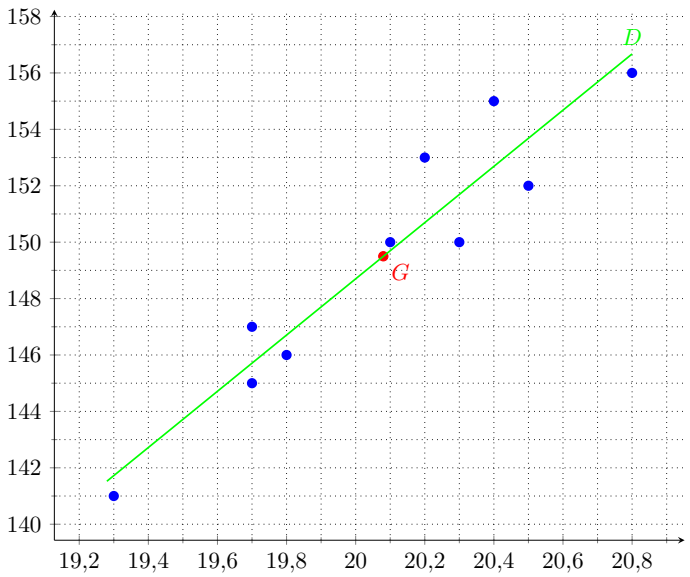
Cette formule n'est pas à savoir, car en pratique, la calculatrice ou l'ordinateur donnent directement l'équation de la droite de régression.

2.3 Utilisation de la calculatrice

On veut obtenir les coefficients a et b de la droite de régression après avoir entré les valeurs.

T.I.	Casio
Touche STAT	Menu STAT
Menu EDIT	Entrer les valeurs x_i dans $List1$
Entrer les valeurs x_i dans L_1	Entrer les valeurs y_i dans $List2$
Entrer les valeurs y_i dans L_2	Menu STAT
Touche STAT	Item CALC
Menu CALC	Régler les paramètres avec set
Item LinReg	Item REG
LinReg L_1, L_2	Choisir X

Si on reprend notre exemple, on trouve : $a\simeq 9,967$ et $b\simeq -50,64$. On peut donc tracer D, la droite de régression de y en x, d'équation $y=9,967x-50,64$. On vérifie qu'elle passe par le point moyen.



3 Changement de variable et ajustement affine

Un ajustement affine n'est pas toujours très approprié lorsque les points ne semblent pas du tout alignés. Dans ce cas, on essaie souvent de trouver un ajustement affine entre les valeurs x_i et les valeurs $f(y_i)$ pour une fonction f bien choisie. Dans les cas habituels, on prend $f(y) = \ln(y)$, $f(y) = \mathrm{e}^y$, $f(y) = y^n$ (pour un certain entier n) ou $f(y) = \sqrt{y}$.

On peut aussi appliquer une fonction aux valeurs x_i, ou aux deux séries, x_i et y_i.

Exemple :

x_i	1	3	4	6	9	11	12
y_i	5	12	15	25	68	115	160
$z_i = \ln(y_i)$	1,61	2,48	2,71	3,22	4,22	4,74	5,08

Le nuage des points $(x_i\ ; y_i)$ n'a pas du tout une forme de droite. On applique aux y_i la fonction ln. Les points $x_i\ ; z_i$ obtenus semblent à peu près alignés. On peut donc chercher un ajustement affine.

La droite de régression a pour équation : $z = 0,304x + 1,434$, soit $\ln(y) = 0,304x + 1,434$.

En appliquant la fonction exponentielle, on obtient : $y = \mathrm{e}^{0,304x+1,434}$, soit (environ) : $y = 4,2\mathrm{e}^{0,304x}$.