Statistiques descriptives

1 Rappel sur les statistiques à une variable

1.1 Vocabulaire

On étudie une information sur un ensemble d'**individus**. Ces individus peuvent être des personnes, mais aussi des animaux, des végétaux, des pays, des pièces sortant d'une machine, etc.

L'ensemble des individus est la **population**. L'information étudiée est le **caractère** ou **variable** statistique (par ex : âge, taille, groupe sanguin, PIB, diamètre...). Le caractère peut être **qualitatif** (ex. groupe sanguin) ou **quantitatif** (ex : taille).

La variable quantitative est **discrète** si elle ne prend que des valeurs isolées. Elle est **continue** si elle peut prendre toutes les valeurs d'un intervalle de $\mathbb{R}$.

Le nombre d'individus correspondant à une valeur de la variable est **l'effectif** de cette valeur. Le nombre d'individus de la population est **l'effectif total**.

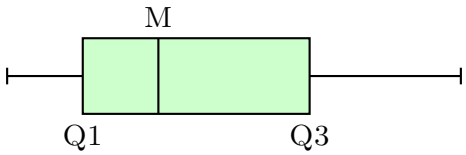
La proportion d'individus associés à une valeur, c'est à dire le quotient $\dfrac{\text{effectif de cette valeur}}{\text{effectif total}}$ est la **fréquence** de cette valeur.

1.2 Graphiques

Lorsque le caractère étudié est **quantitatif et discret**, on peut représenter la série statistique étudiée par un **diagramme en bâtons** ou en barres : la hauteur de chaque bâton est alors proportionnelle à l'effectif (ou à la fréquence) associé à chaque valeur.

Lorsque le caractère étudié est **quantitatif et continu**, et lorsque les modalités sont regroupées en classes, on peut représenter la série par un **histogramme** : l'aire de chaque rectangle est alors proportionnelle à l'effectif (ou à la fréquence) associée à chaque classe.

On représente également une série à l'aide d'un **diagramme en boîte** (ou diagramme à pattes, ou boîte à moustaches, ou whiskers plot).



Il y en a de plusieurs type, mais les bords de la boîte représentent toujours les premier et troisième quartile, et la barre à l'intérieur de la boîte représente la médiane.

Lorsque le caractère est qualitatif, il existe une grande variété de représentations graphiques (diagramme en barre, camembert, etc.)

1.3 Caractéristiques de position d'une variable quantitative

1. Considérons une variable statistique qui prend p valeurs : x_1 avec l'effectif n_1, x_2 avec l'effectif n_2, ..., x_p avec l'effectif n_p. La **moyenne pondérée** est :

$$\overline{x} = \frac{n_1 \times x_1 + n_2 \times x_2 + \cdots + n_p \times x_p}{n_1 + n_2 + \cdots + x_p} = \frac{\text{somme des } n_i \times x_i}{\text{effectif total}}$$

   Si f_1 est la fréquence associée à x_1, f_2 la fréquence associée à x_2, etc., on a aussi :

$$\overline{x} = f_1 \times x_1 + f_2 \times x_2 + \cdots + f_p \times x_p = \text{somme des } f_i \times x_i$$

C'est pratiquement la même formule, sauf qu'il est inutile de diviser par la somme des fréquences car celle-ci vaut toujours 1.

2. La **médiane** M_e d'une série statistique **ordonnée** est une valeur qui partage la population en deux groupes à peu près de même effectif.
3. Les **quartiles** Q_1, Q_2 et Q_3 d'une série sont trois valeurs de la série ordonnée qui la partagent en quatre groupes à peu près de même effectif ($\simeq 25\%$ de l'effectif total).

1.4 Caractéristiques de dispersion d'une variable quantitative

1. La **variance** d'une série est la moyenne des carrés des écarts à la moyenne :

$$Var(x) = \frac{n_1 \times (x_1 - \overline{x})^2 + n_2 \times (x_2 - \overline{x})^2 + \cdots + n_p \times (x_p - \overline{x})^2}{n_1 + n_2 + \cdots + n_p}$$

ou, en utilisant les fréquences :

$$Var(x) = f_1 \times (x_1 - \overline{x})^2 + f_2 \times (x_2 - \overline{x})^2 + \cdots + f_p \times (x_p - \overline{x})^2$$

L'écart-type est égal à la racine carrée de la variance : $\sigma(x) = \sqrt{Var(x)}$
Plus la variance d'une série est grande, plus les valeurs sont dispersées. Même chose pour l'écart-type.

2. L'**écart inter-quartile** est le nombre $Q_3 - Q_1$.
Plus l'écart inter-quartile d'une série est grand, plus cette série est dispersée.
3. L'**étendue** d'une variable statistique est le nombre : valeur maximum−valeur minimum.

2 Statistiques à deux variables

2.1 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 :

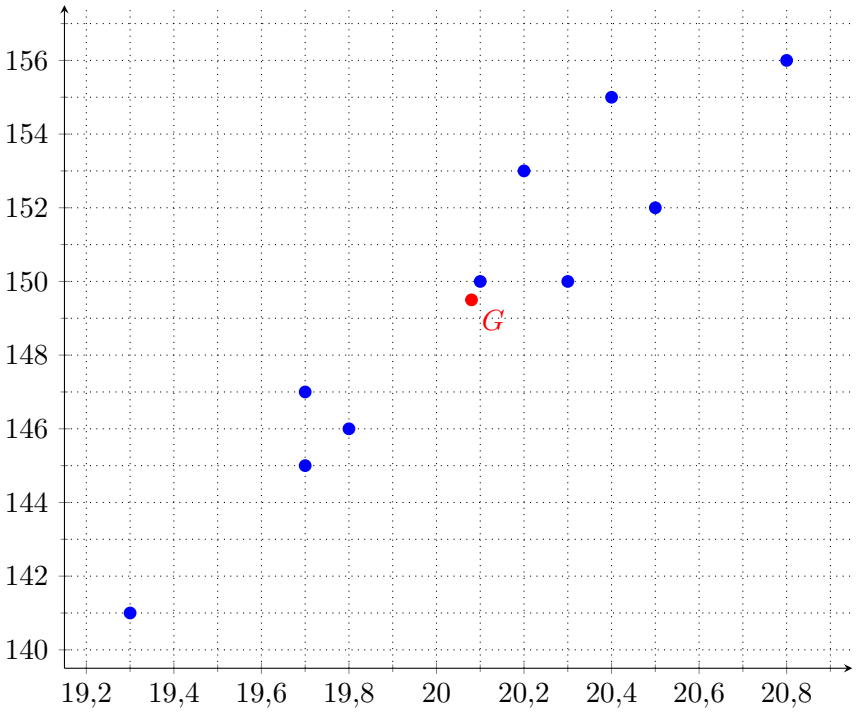
Pièce n°	1	2	3	4	5	5	7	8	9	10
Diamètre (cm)	19,3	20,5	19,8	19,7	20,2	20,4	20,3	19,7	20,8	20,1
Masse (g)	141	152	146	147	153	155	150	145	156	150

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

Calculons les coordonnées du point moyen :

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

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



3 Ajustement affine

3.1 Méthode graphique

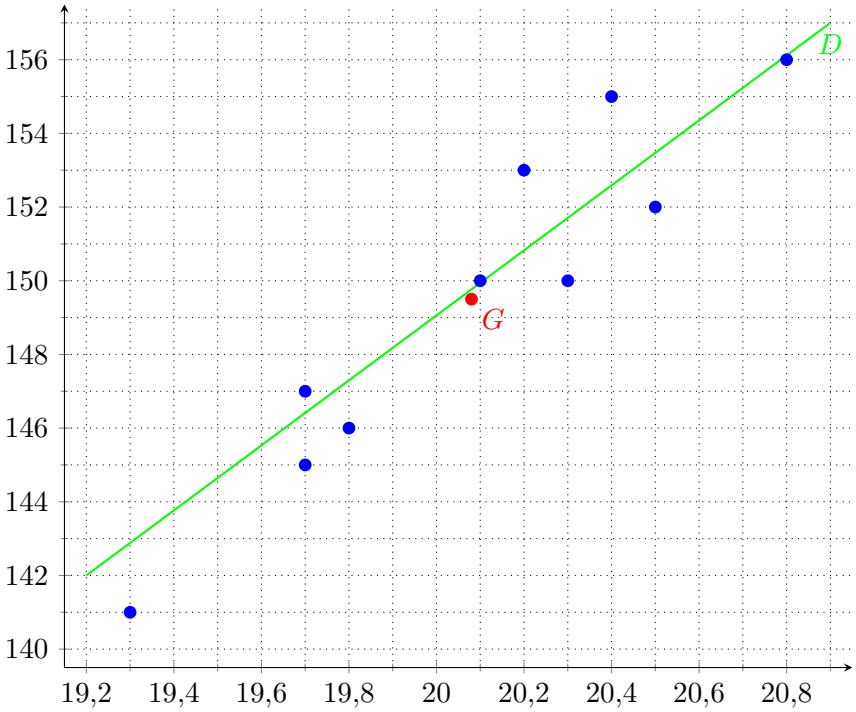
Lorsque les points M_i semblent à peu près alignés, on cherche à construire une droite qui représente au mieux la forme du nuage, c'est à dire qui passe « au plus près » de chacun des points.

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 :

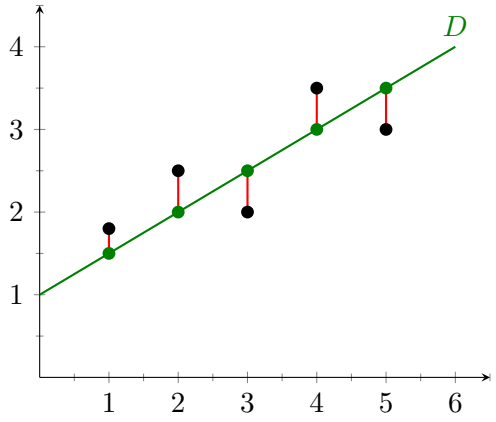
On partage le nuage de points en deux nuages d'affectif à peu près égal, de part et d'autre du point moyen. On détermine les coordonnées des points moyens G_1 et G_2 de chacun de ces nuages.

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

3.2 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 $[y_i - (ax_i + b)]^2$.

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 :

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

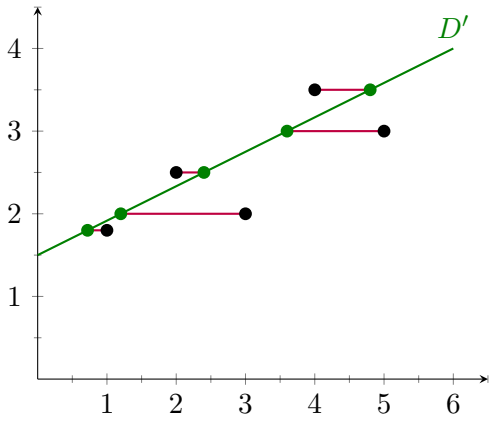
Propriété 1
Le coefficient directeur de la droite de régression est : $a = \dfrac{Cov(x\,;y)}{Var(x)}$.

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

On peut aussi choisir de minimiser les distances non pas entre points de même abscisse, mais entre point de même ordonnée. Il suffit de permuter les deux variables x et y.

Définition 5
Soit D' une droite d'équation $x = a'y + b'$. Appelons Q_i le point d'abscisse y_i situé sur la droite D' (son abscisse est donc $a'y_i + b'$).
On appelle **droite de régression de x en y** la droite D' choisie de telle sorte que la somme des $M_iQ_i^2$ soit minimale.

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



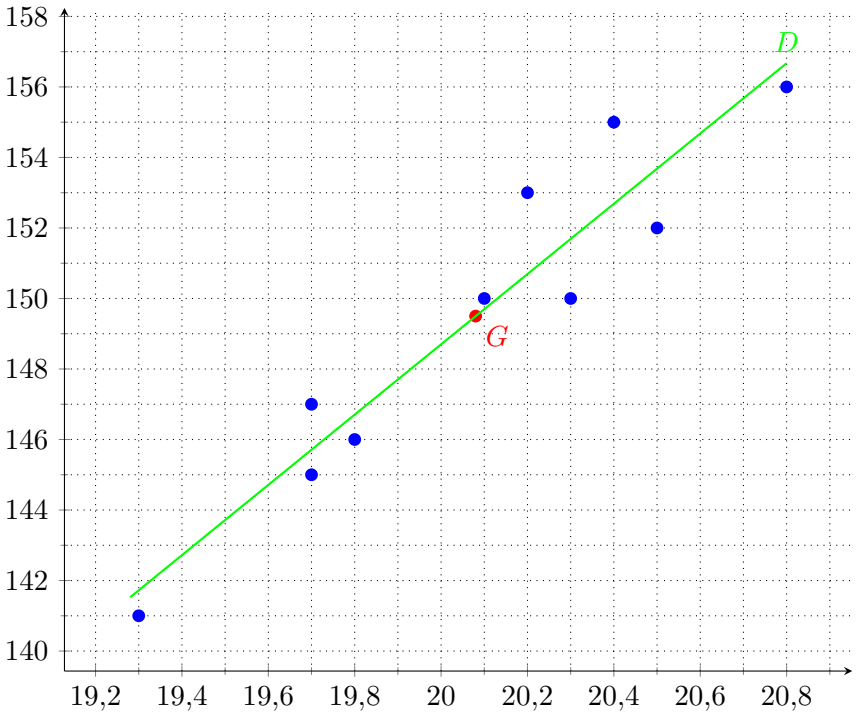
La droite de régression de x en y passe, elle aussi, toujours par le point moyen du nuage, et on trouve, sans surprise : $a' = \dfrac{Cov(x\,;y)}{Var(Y)}$.

3.3 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.



3.4 coefficient de corrélation linéaire

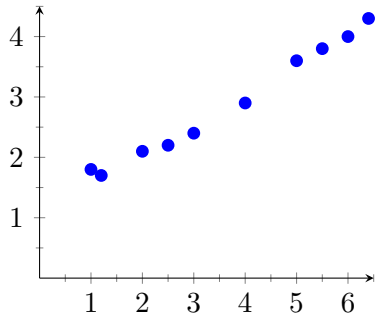
Définition 6
Le coefficient de corrélation linéaire d'une série statistique double de variables x et y est le nombre :

$$r = \frac{Cov(x\,;y)}{\sqrt{Var(x) \times Var(Y)}} = \frac{Cov(x\,;y)}{\sigma(x)\sigma(y)}$$

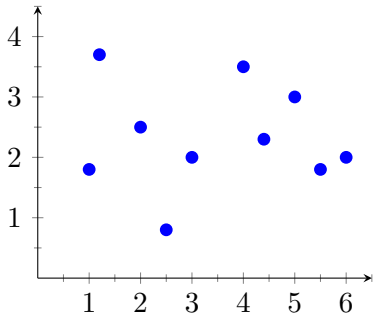
Propriété 2
Le coefficient de corrélation linéaire est un nombre toujours compris entre -1 et 1 ; il sert à mesurer la pertinence d'un ajustement affine :

- Plus r est proche de 1 ou -1, plus l'ajustement affine est justifié : les points du nuage sont presque alignés.
- Plus r est proche de 0, moins l'ajustement linéaire est justifié : les points ne sont pas alignés du tout, ou alors la droite de régression est quasiment parallèle à l'un des axes de coordonnées.

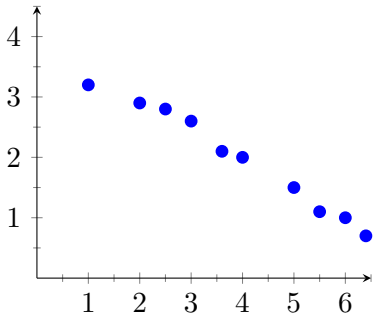
r proche de 1



r proche de 0



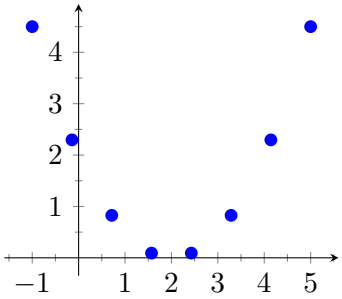
r proche de -1



La calculatrice donne en général le coefficient de corrélation avec les coefficients a et b de la droite d'ajustement (sauf sur les TI82.)

Remarques :

- Deux variables x et y peuvent être corrélées sans que cette corrélation soit linéaire : les points sont disposés à peu près suivant une courbe, mais pas suivant une droite.



- Une corrélation n'indique pas forcément une causalité. Les points peuvent se trouver alignés, alors que les deux variables étudiées n'ont aucun rapport.
  Consulter à ce sujet le site : http://www.tylervigen.com/spurious-correlations
  Examinant les données de la période 2000-2009, l'auteur trouve par exemple une corrélation quasi parfaite entre le taux de divorce dans le Maine et la consommation moyenne de margarine par habitant.