Chapitre 4

Statistiques descriptives

4.1 Statistiques à une variable

4.1.1 Vocabulaire

Effectuer une étude statistique consiste à exploiter des informations sur un ensemble, appelé **population**, constitué d'**individus**. L'étude porte sur un **caractère** (ex. âge, taille, couleur des yeux...). Le caractère peut être **qualitatif** (ex. couleur des yeux) ou **quantitatif** (ex. poids). La variable est dite **discrète** si elle ne prend que des valeurs isolées. Elle est dite **continue** si elle peut prendre toutes les valeurs d'un intervalle de $\mathbb{R}$. Le nombre d'individus d'une valeur est **l'effectif** de cette valeur. Le nombre d'individus de la population est **l'effectif total**.

4.1.2 Variables discrètes

Représentation

On représente les variables statistiques discrètres sous forme de diagramme en bâtons, histogramme ou de graphique circulaire à l'aide des fréquences de chaque valeur du caratère.

Caractéristiques

1. La moyenne pondérée

   Soient n valeurs de la variable. Si cette variable prend p valeurs disctinctes ($p \leq n$), $x_1, \cdots, x_p$, d'effectifs respectifs $n_1, \cdots, n_p$ alors la moyenne est donnée par :

   $$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i$$

   Propriété 1 *Si pour tout $i \in \mathbb{N}_p$ on opére un changement de variable affine ${x'}_i = ax_i + b$ avec ($a \in \mathbb{R}^*$ et $b \in \mathbb{R}$) alors on a $\overline{x'} = a\overline{x} + b$*

2. La variance

   La variance est un indicateur de dispersion de la variable statistique :

   $$V = \frac{1}{n}\sum_{i=1}^{p} n_i(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_i {x_i}^2 - \overline{x}^2$$

3. L'ecart-type

   L'ecart-type est égal à la racine carrée de la variance :

   $$\sigma = \sqrt{V}$$

   Propriété 2 *Si pour tout $i \in \mathbb{N}_p$ on opére un changement de variable affine ${x'}_i = ax_i + b$ avec ($a \in \mathbb{R}^*$ et $b \in \mathbb{R}$) alors on a $V_{x'} = a^2 V_x$ et $\sigma_{x'} = |a|\sigma_x$.*

4.1.3 Variables continues

Représentation

Pour leur représentation, on les regroupe en général dans des classes adjacentes d'amplitude pas forcément égale. Ceci est représenté dans le tableau ci-dessous :

Classes	$[X_0; X_1[$	$[X_1; X_2[$	$\cdots\cdots$	$[X_{p-1}; X_p]$
Centre des classes	x_1	x_2		x_p
Effectifs	n_1	n_2		n_p
Fréquences	$\frac{n_1}{n}$	$\frac{n_2}{n}$		$\frac{n_p}{n}$

4.2 Statistiques à deux variables

4.2.1 Tableau de données et nuages de points

Il existe parfois une relation entre deux caractères d'une population. On définit alors une série statistique à deux variables x et y, prenant les valeurs $x_1, \cdots, x_n$ et $y_1, \cdots, y_n$.

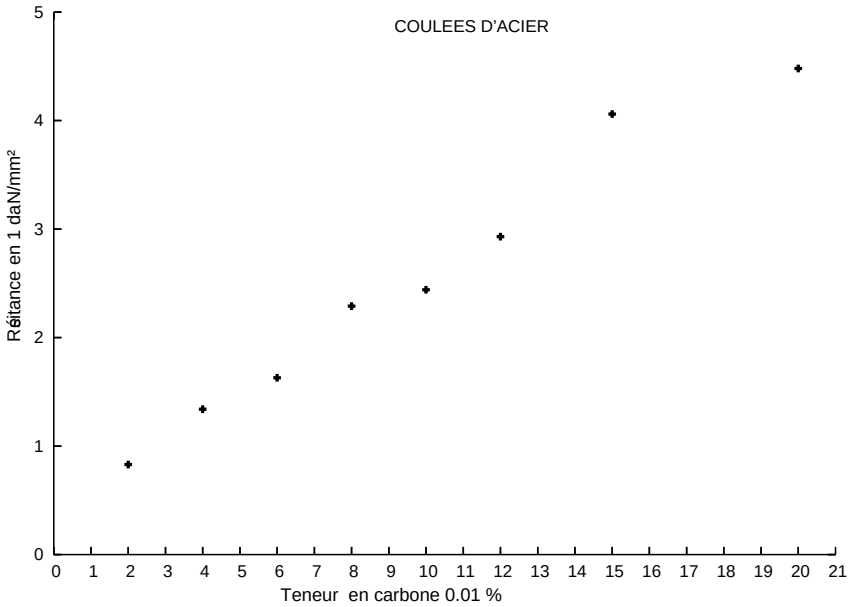
Tableau de données :

Exemple : on a mesuré, pour huit coulées d'acier, la teneur en carbone x (unité : 0,01 %) et la résistance à la traction y (unité : 1 daN/mm^2).

x	2	4	6	8	10	12	15	20
y	0,83	1,34	1,63	2,29	2,44	2,93	4,06	4,48

Nuages de points : Dans le plan muni d'un repère orthogonal $(O; \vec{i}, \vec{j})$, 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.



On cherche alors à trouver une fonction f telle que la courbe d'équation $y = f(x)$ "passe le plus près possible" des points du nuage.

Points moyens

On appelle point moyen d'un nuage de n points $M_i(x_i; y_i)$, le point G défini par :

$$\begin{cases} x_G = \overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \\ y_G = \overline{y} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i \end{cases}$$

4.3 Ajustement affine

4.3.1 Méthode graphique

a. Ajustement à la règle :

Lorsque les points M_i semblent alignés, on recherche une équation du type $y = ax + b$, pour cela on trace au jugé une droite D en s'efforcant d'équilibrer le nombre de points situés de part et d'autre. Ensuite on détermine par lecture graphique les réels a et b.

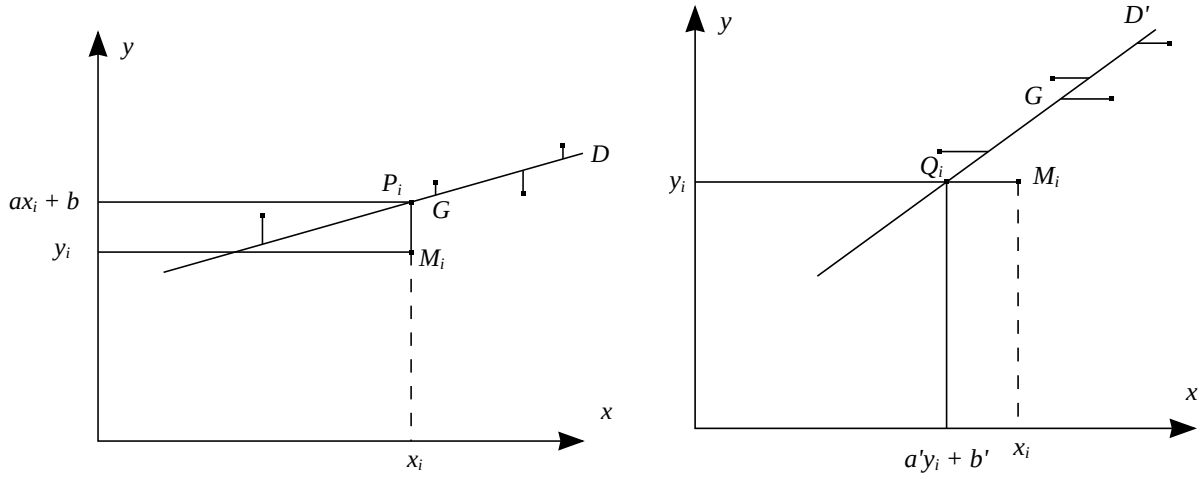
b. Ajustement par la méthode de Mayer :

On partage le nuage de points en deux nuages, on détermine les coordonnées G_1 et G_2 des points moyens respectifs du premier nuage et du deuxième nuage.

La droite (G_1G_2), droite de **Mayer**, constitue une "bonne" droite d'ajustement si le nuage est allongé.

4.3.2 Ajustement affine : Méthode des moindres carrés

a. Droites de régression :



Soient D une droite d'ajustement d'équation $y = ax + b$, et $M_i(x_i; y_i)$ un nuage de points. On pose P_i le point d'abscisse x_i situé sur la droite D.

On appelle **droite de régression de** y **en** x la droite D telle que

la somme $\sum_{i=1}^{n} {M_iP_i}^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$ soit minimale.

Soient D' une droite d'ajustement d'équation $x = a'y + b'$, et $M_i(x_i; y_i)$ un nuage de points. On pose Q_i le point d'ordonnée y_i situé sur la droite D'.

On appelle **droite de régression de** x **en** y la droite D' telle que

la somme $\sum_{i=1}^{n} {M_iQ_i}^2 = \sum_{i=1}^{n} [x_i - (a'y_i + b')]^2$ soit minimale. **b. Covariance :**

La covariance de la série statistique double de caractères x et y est le réel :

$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)(y_i - y) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y}.$

c. Equations des droites de regression :

On montre que :

- La droite de régression D de y en x a pour équation $y = ax + b$ avec : $a = \frac{\sigma_{xy}}{[\sigma_x]^2}$ et $b = \bar{y} - a\bar{x}$

  La droite D passe par le point moyen $G(\bar{x}; \bar{y})$ du nuage.

- La droite de régression D' de x en y a pour équation $x = a'y + b'$ avec : $a' = \frac{\sigma_{xy}}{[\sigma_y]^2}$ et $b' = \bar{x} - a'\bar{y}$

  La droite D' passe par le point moyen $G(\bar{x}; \bar{y})$ du nuage.

4.4 Coefficient de corrélation linéaire

Pour apprécier la qualité d'un ajustement affine, on introduit un nouveau paramètre.

Définition :

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

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$$