Cours de Statistiques pour Ingénieurs en Génie des Matériaux
Chapitre 4 : Séries Statistiques à Deux Variables

Introduction

Dans le domaine du génie des matériaux, il est fréquent d'étudier la relation entre deux variables. Par exemple :

— la température de traitement (°C) et la dureté du matériau (HV),
— la composition en un élément d'alliage (%) et la résistance mécanique (MPa),
— la porosité (%) et la conductivité thermique (W/mK).

L'objectif est d'analyser statistiquement ces relations à l'aide de séries statistiques à deux variables.

1 Séries statistiques à deux variables

Une **série statistique à deux variables** associe à chaque individu deux caractères X et Y, qui peuvent être :

— quantitatifs (ex : température, résistance),
— qualitatifs (ex : type de traitement),
— ou l'un quantitatif et l'autre qualitatif.

On note les observations : (x_i, y_i) pour $i = 1, \ldots, n$.

Exemple 1 : Heures d'étude et notes en mathématiques

Une entreprise de tutorat scolaire a collecté les données suivantes pour 20 élèves en difficulté en mathématiques.

Heures d'étude (x)	5	7	3	12	9	12	10	8	5	1
Note (y)	65	72	56	80	75	74	74	75	58	51
Heures d'étude (x)	7	13	10	3	7	20	18	15	8	6
Note (y)	62	83	80	61	67	93	82	85	70	65

1.1 Nuage de points

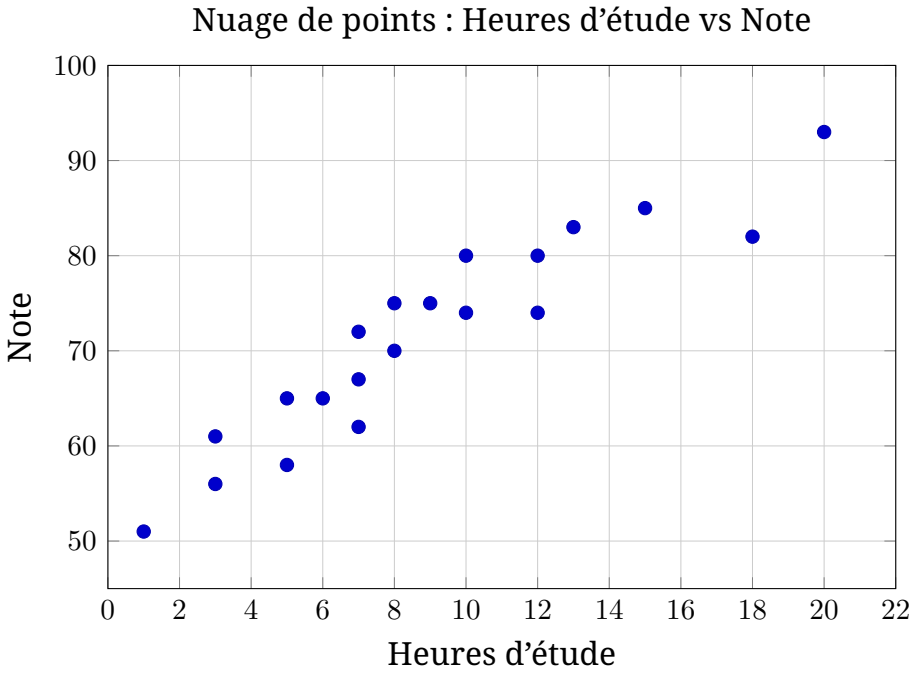


FIGURE 1 – Nuage de points représentant la relation entre le nombre d'heures d'étude et la note obtenue.

2 Covariance

La **covariance** mesure la tendance de deux variables à varier ensemble.

Formule

Soient $\bar{x}$ et $\bar{y}$ les moyennes de X et Y :

$$\bar{x} = \sum_{i=1}^{k} x_i f_{i\cdot}, \quad \bar{y} = \sum_{j=1}^{m} y_j f_{\cdot j}.$$

La covariance est définie par :

$$\mathrm{Cov}(X,Y) = \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})(y_j - \bar{y})f_{ij}.$$

Interprétation

— $\mathrm{Cov}(X,Y) > 0$: les variables ont tendance à augmenter ensemble.
— $\mathrm{Cov}(X,Y) < 0$: lorsque l'une augmente, l'autre diminue.
— $\mathrm{Cov}(X,Y) = 0$: absence de dépendance linéaire.

Exemple en génie des matériaux

Si X = température de traitement et Y = dureté du matériau :
— Covariance positive $\Rightarrow$ plus la température est élevée, plus la dureté augmente.
— Covariance négative $\Rightarrow$ sur-traitement thermique réduit la dureté.

Exemple 2 : Tableau avec fréquences

On a relevé, sur un échantillon de pièces métalliques, le temps de traitement thermique X (en heures) et la dureté Y (en HV) obtenue. Certaines combinaisons (x_i, y_i) apparaissent plusieurs fois. On résume les données dans le tableau suivant :

Temps de traitement X (h)	1	2	3	4	5
Dureté Y (HV)	180	210	235	250	270
Effectif n_i	4	7	10	6	3

Interprétation : La combinaison $(X = 1, Y = 180)$ a été observée 4 fois, $(X = 2, Y = 210)$ a été observée 7 fois, etc. L'effectif total est $N = 4 + 7 + 10 + 6 + 3 = 30$.

1. Calcul des moyennes

$$\bar{x} = \frac{\sum n_i x_i}{N} = \frac{4 \cdot 1 + 7 \cdot 2 + 10 \cdot 3 + 6 \cdot 4 + 3 \cdot 5}{30} = \frac{4 + 14 + 30 + 24 + 15}{30} = \frac{87}{30} = 2{,}9$$

$$\bar{y} = \frac{\sum n_i y_i}{N} = \frac{4 \cdot 180 + 7 \cdot 210 + 10 \cdot 235 + 6 \cdot 250 + 3 \cdot 270}{30}$$
$$= \frac{720 + 1470 + 2350 + 1500 + 810}{30} = \frac{6850}{30} \approx 228{,}333$$

2. Calcul de la covariance

$$\mathrm{Cov}(X, Y) = \frac{1}{N} \sum n_i (x_i - \bar{x})(y_i - \bar{y})$$

On calcule d'abord les écarts :

x_i	1	2	3	4	5
$x_i - \bar{x}$	$-1{,}9$	$-0{,}9$	$0{,}1$	$1{,}1$	$2{,}1$
$y_i - \bar{y}$	$-48{,}333$	$-18{,}333$	$6{,}667$	$21{,}667$	$41{,}667$

Puis :

$$\begin{aligned}\sum n_i(x_i-\bar{x})(y_i-\bar{y}) &= 4(-1{,}9)(-48{,}333)+7(-0{,}9)(-18{,}333)+10(0{,}1)(6{,}667)\\ &\quad +6(1{,}1)(21{,}667)+3(2{,}1)(41{,}667)\\ &= 4\cdot 91{,}833+7\cdot 16{,}500+10\cdot 0{,}667+6\cdot 23{,}834+3\cdot 87{,}500\\ &= 367{,}332+115{,}500+6{,}667+143{,}004+262{,}500\\ &= 895{,}003\end{aligned}$$

$$\mathrm{Cov}(X,Y)=\frac{895{,}003}{30}\approx 29833{,}4$$

3 Tableau d'effectifs à double entrée

Lorsqu'on étudie deux variables quantitatives discrètes X et Y, on peut organiser les données sous la forme d'un **tableau à double entrée** (aussi appelé tableau de contingence).

- Les **lignes** correspondent aux différentes valeurs de X - Les **colonnes** correspondent aux différentes valeurs de Y - Chaque case contient l'**effectif** n_{ij} (nombre d'individus ayant simultanément x_i et y_j)

On ajoute généralement : - une ligne des effectifs marginaux de Y - une colonne des effectifs marginaux de X - l'effectif total N

3.1 Exemple concret

On a mesuré, sur un échantillon de 50 pièces, la durée de traitement thermique X (en heures) et la dureté obtenue Y (en HV). Les résultats sont regroupés dans le tableau suivant :

	Dureté Y (HV)				
Temps X (h)	200	220	240	260	**Total**
2	4	3	1	0	8
4	2	6	5	1	14
6	0	3	8	4	15
8	0	1	4	8	13
Total	6	13	18	13	**50**

3.2 Calculs préliminaires

3.2.1 1. Effectifs marginaux et moyennes

Les effectifs marginaux de X sont déjà donnés dans la dernière colonne. Ceux de Y sont dans la dernière ligne.

$$\bar{x}=\frac{\sum n_{i\cdot}x_i}{N}=\frac{8\cdot 2+14\cdot 4+15\cdot 6+13\cdot 8}{50}=\frac{16+56+90+104}{50}=\frac{266}{50}=5{,}32$$

$$\bar{y} = \frac{\sum n_{\cdot j} y_j}{N} = \frac{6 \cdot 200 + 13 \cdot 220 + 18 \cdot 240 + 13 \cdot 260}{50}$$
$$= \frac{1200 + 2860 + 4320 + 3380}{50} = \frac{11760}{50} = 235{,}2$$

3.2.2 2. Covariance

$$\mathrm{Cov}(X, Y) = \frac{1}{N} \sum_i \sum_j n_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

Après développement du calcul (en utilisant les écarts à la moyenne), on obtient :

$$\mathrm{Cov}(X, Y) \approx 48{,}74$$

3. Variance de X

$$V(X) = \frac{1}{N} \sum n_{i\cdot} (x_i - \bar{x})^2 \approx 4{,}6176$$

4 Fréquences marginales et fréquences conditionnelles

À partir d'un tableau à double entrée, on peut calculer deux types de fréquences très utiles pour l'analyse : les **fréquences marginales** et les **fréquences conditionnelles**.

4.1 1. Fréquences marginales

Les **fréquences marginales** décrivent la distribution de chaque variable *prise séparément*, en ignorant l'autre variable.

- Fréquence marginale de $X = x_i$:

$$f_{i\cdot} = \frac{n_{i\cdot}}{N}$$

- Fréquence marginale de $Y = y_j$:

$$f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

Application à l'exemple

Temps X (h)	**Dureté Y (HV)** 200	220	240	260	**Total**
2	0,08	0,06	0,02	0	0,16
4	0,04	0,12	0,10	0,02	0,28
6	0	0,06	0,16	0,08	0,30
8	0	0,02	0,08	0,16	0,26
Total	0,12	0,26	0,36	0,26	**1,00**

4.2 2. Fréquences conditionnelles

Les **fréquences conditionnelles** décrivent la distribution d'une variable *sachant* la valeur de l'autre variable. Elles permettent d'étudier la liaison entre X et Y.

On distingue deux types :

- Fréquence conditionnelle de Y sachant $X = x_i$:

$$f_{j|i} = \frac{n_{ij}}{n_{i.}}$$

- Fréquence conditionnelle de X sachant $Y = y_j$:

$$f_{i|j} = \frac{n_{ij}}{n_{.j}}$$

Application à l'exemple

Distribution de la dureté Y pour les pièces ayant subi un traitement de 6 heures ($X = 6$) :

$$\begin{aligned} f(Y = 200 \,|\, X = 6) &= \frac{0}{15} = 0 \\ f(Y = 220 \,|\, X = 6) &= \frac{3}{15} = 0{,}20 \quad (20\,\%) \\ f(Y = 240 \,|\, X = 6) &= \frac{8}{15} \approx 0{,}533 \quad (53{,}3\,\%) \\ f(Y = 260 \,|\, X = 6) &= \frac{4}{15} \approx 0{,}267 \quad (26{,}7\,\%) \end{aligned}$$

Interprétation : Parmi les pièces traitées pendant 6 heures, plus de la moitié (53,3 %) ont une dureté de 240 HV.

Distribution du temps de traitement X pour les pièces ayant une dureté de 260 HV ($Y = 260$) :

$$\begin{aligned} f(X = 2 \,|\, Y = 260) &= \frac{0}{13} = 0 \\ f(X = 4 \,|\, Y = 260) &= \frac{1}{13} \approx 0{,}077 \quad (7{,}7\,\%) \\ f(X = 6 \,|\, Y = 260) &= \frac{4}{13} \approx 0{,}308 \quad (30{,}8\,\%) \\ f(X = 8 \,|\, Y = 260) &= \frac{8}{13} \approx 0{,}615 \quad (61{,}5\,\%) \end{aligned}$$

Interprétation : Parmi les pièces les plus dures (260 HV), la grande majorité (61,5 %) ont subi un traitement de 8 heures.

4.3 Intérêt des fréquences conditionnelles

Les fréquences conditionnelles permettent de répondre à des questions du type :

— « Quelle est la répartition de la dureté *parmi* les pièces traitées pendant 6 heures ? »

— « Quel temps de traitement est le plus fréquent *parmi* les pièces de dureté 260 HV ? »

Elles sont essentielles pour étudier la **dépendance** entre les deux variables. Si les fréquences conditionnelles varient fortement selon la valeur de l'autre variable, cela indique une liaison entre X et Y.

5 Coefficient de corrélation

La covariance dépend de l'unité des variables. Pour la normaliser :

$$r = \frac{\mathrm{Cov}(X, Y)}{s_X s_Y},$$

où s_X et s_Y sont les écarts-types.

6 Conclusion

Ce chapitre propose les outils fondamentaux pour analyser les relations entre deux variables en ingénierie des matériaux. Les notions de distributions marginales, conditionnelles et de covariance permettent de comprendre la corrélation entre paramètres de procédé et propriétés du matériau.