Cours de Statistiques pour Ingénieurs en Génie des Matériaux
Chapitre 1 : Introduction aux Notions de Base

Objectifs de la séance

— Comprendre les concepts de population, échantillon, variables et modalités.

— Identifier et distinguer les différents types de variables statistiques : qualitatives, quantitatives, discrètes et continues.

— Appliquer ces notions à des exemples concrets en génie des matériaux.

Introduction

La statistique descriptive a pour but :

— de dégager les propriétés essentielles que l'on peut déduire d'une accumulation de données ;

— de donner une image concise et simplifiée de la réalité.

Le résultat d'une observation, d'une mesure, n'est pas égale à la valeur théorique calculée ou espérée par l'ingénieur ; la répétition d'une même mesure, réalisée dans des conditions qui semblent identiques, ne conduit pas toujours aux mêmes résultats. Ces fluctuations, dues à des causes nombreuses, connues ou inconnues, contrôlées ou non, créent des difficultés aux ingénieurs et aux scientifiques. Quel résultat doivent-ils prendre ? Quel degré de confiance peuvent-ils accorder à la décision prise ? Les réponses à une enquête varient d'un individu à un autre ; quelles conclusions valables peut-on tirer d'un sondage ? Les méthodes de la statistique descriptive apportent des réponses à ces problèmes.

Pour être soumis à un traitement statistique, un tableau de données doit comporter au moins une variable de nature aléatoire. Une définition simple du caractère aléatoire d'une variable est qu'elle peut prendre au hasard des valeurs différentes.En génie des matériaux, les statistiques sont utilisées pour étudier les propriétés des matériaux (ex. : résistance, dureté), optimiser les procédés de fabrication, ou évaluer la fiabilité des produits.

Ce premier chapitre donne les définitions et les propriétés des principales notions utiles pour comprendre et traiter un problème de statistique.

1 Notions de population et d'échantillon

1.1 Population et Individu

Population : Ensemble complet des individus, objets ou événements étudiés.

On peut aussi dire *ensemble statistique*. Cette terminologie vient de ce que les premiers statisticiens étaient intéressés par les recenssements civiles et militaires.

Unité statistique : chaque individu.

Une population doit être correctement définie afin que l'appartenance d'un individu à cette population soit reconnue sans ambiguïté.

Exemple :
Une usine fabrique des tiges métalliques utilisées dans l'assemblage de certaines structures. Pour étudier la résistance à la traction de ces tiges, on mesure cette résistance pour un lot de 100 tiges.
Population statistique : l'ensemble des 100 tiges ou des 100 mesures.
Unité statistique : chacune des tiges ou chacune des 100 mesures.
Propriété étudiée : la résistance à la traction de tiges métalliques.

1.2 Échantillon

Échantillon : groupe restreint, ou sous-ensemble, issu de la population.

Pour définir un tel échantillon, une méthode consiste à prélever, au hasard, un sous-ensemble d'individus, en utilisant, par exemple, des tables de nombres au hasard

— Exemple : 50 échantillons d'aluminium prélevés pour tester leur résistance à la traction.

— Importance : Étudier un échantillon est souvent plus pratique et économique qu'étudier toute la population.

1.3 Caractères et Variables Aléatoires

Caractères

On s'intéresse à certaines particularités ou caractères des individus d'une population statistique :

— un seul caractère étudié, série numérique à une dimension,

— deux caractères étudiés, série numérique à deux dimensions,

— plus de deux caractères, on doit utiliser les techniques de l'analyse multidimensionnelle (qui dépasse le cadre de ce cours.

Les caractères étudiés peuvent être :

— le poids, la taille, le niveau d'études, la catégorie socioprofessionnelle, le lieu d'habitation..., dans le secteur des sciences humaines,

— le poids, la masse, la composition..., dans le secteur des sciences techniques.

Modalités

Un caractère peut prendre différentes **modalités**. Ces modalités doivent être incompatibles et exhaustives afin que l'appartenance ou la non-appartenance d'un individu à une modalité soit définie sans ambiguïté.

Un caractère peut être :

— quantitatif, les modalités sont mesurables ou repérables,

— qualitatif, les modalités ne sont pas mesurables.

Exemple : Pour la variable "type de revêtement", les modalités peuvent être : peinture, galvanisation, anodisation.

Variables statistiques

Variable statistique : un caractère faisant l'objet d'une étude statistique.

Elle peut donc être qualitative ou quantitative.

Exemple : Dureté d'un alliage, température de fusion, type de revêtement.

Une variable quantitative est appelée :

— discrète si elle prend un nombre fini de valeurs souvent entières,

— continue si elle prend toutes les valeurs d'un intervalle fini ou infini.

Discussion

— Pourquoi ne pas tester tous les échantillons ? (Coût, temps, faisabilité)

— Comment choisir un échantillon représentatif ? (Aléatoire, stratification)

Application pratique

Soit une population de 100 pièces en céramique. On mesure deux variables :

— Taille des pores (en micromètres).

— Type de céramique (zircone, alumine, carbure de silicium).

Question : Identifiez les variables et leurs modalités.

2 Types de variables statistiques

2.1 Variables qualitatives

Variables qualitatives : Décrivent une qualité ou une catégorie, sans valeur numérique intrinsèque.

Elles peuvent être de deux types :

— *Nominales* : Pas d'ordre (ex. : type de matériau : acier, aluminium, titane).

— *Ordinales* : Avec un ordre (ex. : qualité d'un revêtement : faible, moyenne, élevée).

Exemple : Classification des alliages par type (ferreux, non ferreux).

2.2 Variables quantitatives

Variables quantitatives : Représentent une quantité mesurable.

Elles peuvent être de deux types :

— *Discrètes* : Valeurs dénombrables (ex. : nombre de défauts dans un échantillon).

— *Continues* : Valeurs dans un intervalle continu (ex. : résistance à la traction en MPa).

Exemple : Mesure de la densité d'un matériau (continue) ou nombre de cycles avant rupture (discrète).

2.3 Tableau récapitulatif

Type	Exemple	Caractéristique
Qualitative nominale	Type de matériau	Catégories sans ordre
Qualitative ordinale	Qualité du matériau	Catégories ordonnées
Quantitative discrète	Nombre de défauts	Valeurs dénombrables
Quantitative continue	Résistance (MPa)	Valeurs dans un intervalle

TABLE 1 – Types de variables statistiques

Application pratique

Pour chaque variable suivante, indiquez si elle est qualitative (nominale ou ordinale) ou quantitative (discrète ou continue) :

1. Dureté Vickers d'un alliage.
2. Type de polymère (ABS, PLA, PET).
3. Nombre de fissures observées.
4. Température de traitement thermique.
5. Classe de résistance (basse, moyenne, haute).

3 Effectif, Fréquence et Pourcentage

Effectif : Nombre d'observations d'une modalité ou d'une classe pour une variable donnée. On le note n_i. Et on note l'effectif total par N, c'est-à dire $N = \sum_i n_i$.

Exemple : Dans un échantillon de 100 pièces en acier, on observe 20 pièces avec des défauts de surface. L'effectif de la modalité « défaut » est 20.

Fréquence Rapport entre l'effectif d'une modalité (ou classe) et l'effectif total, exprimé sous forme décimale.
Formule : $f_i = \frac{n_i}{N}$, où n_i est l'effectif de la modalité i et N est l'effectif total.

Exemple : Si 20 pièces sur 100 ont un défaut, la fréquence des défauts est $\frac{20}{100} = 0,2$.
Remarque : On vérifie facilement que la somme des fréquences est 1 :
$\sum_i f_i = \sum_i \frac{n_i}{N} = \frac{1}{N} \sum_i n_i = \frac{N}{N} = 1$

Pourcentage : Fréquence exprimée en pourcentage ($f_i \times 100$) %.

Exemple : La fréquence de 0,2 correspond à 20 %.

3.1 Exemple concret

On mesure la qualité d'un revêtement sur 50 échantillons de céramique, avec les modalités suivantes : faible, moyenne, élevée. Résultats :

— Faible : 10 échantillons.

— Moyenne : 25 échantillons.

— Élevée : 15 échantillons.

Calculs :

— Effectif total : $10 + 25 + 15 = 50$.

— Fréquences : Faible $\frac{10}{50} = 0,2$, Moyenne $\frac{25}{50} = 0,5$, Élevée $\frac{15}{50} = 0,3$.

— Pourcentages : Faible 20 %, Moyenne 50 %, Élevée 30 %.

3.2 Tableau récapitulatif

Qualité	Effectif	Fréquence	Pourcentage
Faible	10	0,2	20 %
Moyenne	25	0,5	50 %
Élevée	15	0,3	30 %
Total	50	1,0	100 %

TABLE 2 – Série statistique pour la qualité du revêtement

4 Effectif cumulé et Fréquence cumulée

Effectif cumulé : Somme des effectifs des modalités ou classes jusqu'à une certaine valeur (souvent pour des variables ordinales ou quantitatives).

Exemple : Pour la qualité du revêtement (faible, moyenne, élevée), les effectifs cumulés sont :

— Faible : 10.

— Moyenne : $10 + 25 = 35$.

— Élevée : $35 + 15 = 50$.

Fréquence cumulée : Somme des fréquences jusqu'à une certaine modalité ou classe, exprimée sous forme décimale ou en pourcentage.
Formule : $F_i = \sum_{k=1}^{i} f_k$, où f_k est la fréquence de la modalité k.

Exemple : Fréquences cumulées pour l'exemple précédent :

— Faible : 0,2 (20 %).

— Moyenne : $0,2 + 0,5 = 0,7$ (70 %).

— Élevée : $0,7 + 0,3 = 1,0$ (100 %).

4.1 Exemple avec une variable quantitative

On mesure la résistance à la traction (en MPa) de 60 échantillons d'un alliage, regroupés en classes :

Classe (MPa)	Effectif	Fréquence	Effectif cumulé	Fréquence cumulée
$[200, 250[$	10	0,167	10	0,167
$[250, 300[$	20	0,333	30	0,500
$[300, 350[$	15	0,250	45	0,750
$[350, 400[$	15	0,250	60	1,000
Total	60	1,000		

TABLE 3 – Série statistique pour la résistance à la traction

Interprétation

— 50 % des échantillons ont une résistance inférieure à 300 MPa (fréquence cumulée à [250, 300[).

— Utile pour évaluer la proportion d'échantillons répondant à un seuil de performance (ex. : résistance minimale pour une application).

5 Exemple pratique

Défauts relevés sur une pièce de tissu

Un fabricant de tissu essaie une nouvelle machine ; il compte le nombre de défauts sur 75 échantillons de 10 mètres. Il a trouvé les résultats suivants :

Nombre d'individus : les 75 échantillons. $N = 75$

Effectif associé à la valeur k, le nombre n_k : par exemple, sur les 75 échantillons examinés, 11 présentent $k = 2$ défauts, donc si $k = 2$, $n_k = 11$.

Fréquence associée à la valeur k : le quotient n_k/n.

$11/75 = 0,146$ est la fréquence associée à la valeur $k = 2$.

Effectif cumulé associé à la valeur k : le nombre d'échantillons ayant au plus k défauts (k compris).

$38 + 15 + 11 = 64$ est la fréquence cumulée absolue associée à la valeur $k = 2$.

Fréquence cumulée associée à la valeur k, le nombre d'échantillons ayant au plus k défauts (k compris) divisé par n.

$64/75 = 0,853$ est la fréquence cumulée relative associée à la valeur $k = 2$.

6 Conclusion et discussion

— Les fréquences relatives et les fréquences cumulées relatives peuvent être utilisées pour comparer deux ou plusieurs populations.

— Dans le cas d'une distribution continue, les données sont en général regroupées en classes. Les fréquences absolues, relatives et cumulées sont définies par rapport aux classes et non par rapport aux valeurs de la variable.

— Questions ouvertes : Quelles variables dans vos projets pourraient être analysées avec ces outils ?

Cours de Statistiques pour Ingénieurs en Génie des Matériaux
Chapitre 2 : Représentations Graphiques de Séries numériques à une dimension

Objectifs de la séance

— Comprendre et savoir construire les diagrammes à bandes, circulaires, en bâtons, les polygones des effectifs/fréquences, les histogrammes et les courbes cumulatives.
— Interpréter ces représentations dans le contexte du génie des matériaux.
— Choisir la représentation graphique adaptée à une variable statistique.

1 Introduction

Les représentations graphiques permettent de visualiser et d'interpréter les données statistiques de manière intuitive. En génie des matériaux, elles aident à analyser les propriétés des matériaux (ex. : résistance, défauts) et à communiquer les résultats. Cette séance présente les principaux types de graphiques pour les séries statistiques à une variable, avec des exemples concrets et leurs représentations graphiques.

2 Différents modes de représentation graphique des données

Diagramme en feuilles

On décompose une donnée numérique en deux parties :
— la tige qui comprend le premier ou les deux premiers chiffres,
— la feuille qui comprend les autres chiffres.

On écrit les tiges les unes sous les autres et en regard de chaque tige, les feuilles correspondantes ; tiges et feuilles sont séparées par un trait vertical.

Exemple : Le tableau suivant donne le poids en grammes de 25 éprouvettes.

250	253	256	258	260	261	263	265	270
271	272	273	274	276	276	279	279	281
284	285	286	287	288	290	290		

Comme tige, on choisit les deux premiers chiffres de chaque mesure, c'est-à-

dire 25, 26, 27, 28 et 29. Les feuilles sont alors constituées du dernier chiffre de la mesure :

Tige	Feuilles								
25	0	3	6	8					
26	0	1	3	5					
27	0	1	2	3	4	6	6	9	9
28	1	4	5	6	7	8			
29	0	0							

Le diagramme indique que le poids moyen se situe entre 270 et 280 g et qu'il doit être voisin de 270 g.

Variables discrètes : Diagramme en bâtons

Définition : Représentation des effectifs ou fréquences par des rectangles de même largeur, espacés, pour des variables qualitatives ou discrètes.

Exemple : Classement de 100 familles en fonction du nombre d'enfants

On a relevé le nombre d'enfants de 100 familles choisies au hasard. Le tableau suivant donne les principales caractéristiques de cette étude. Construction : Axe horizontal (modalités), axe vertical (effectifs ou fréquences).

x_i	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
n_i	20	25	30	10	5	5	3	2	100
f_i	0,20	0,25	0,30	0,10	0,05	0,05	0,03	0,02	1

TABLE 1 – Statistique sur le nombre d'enfants de 100 familles.

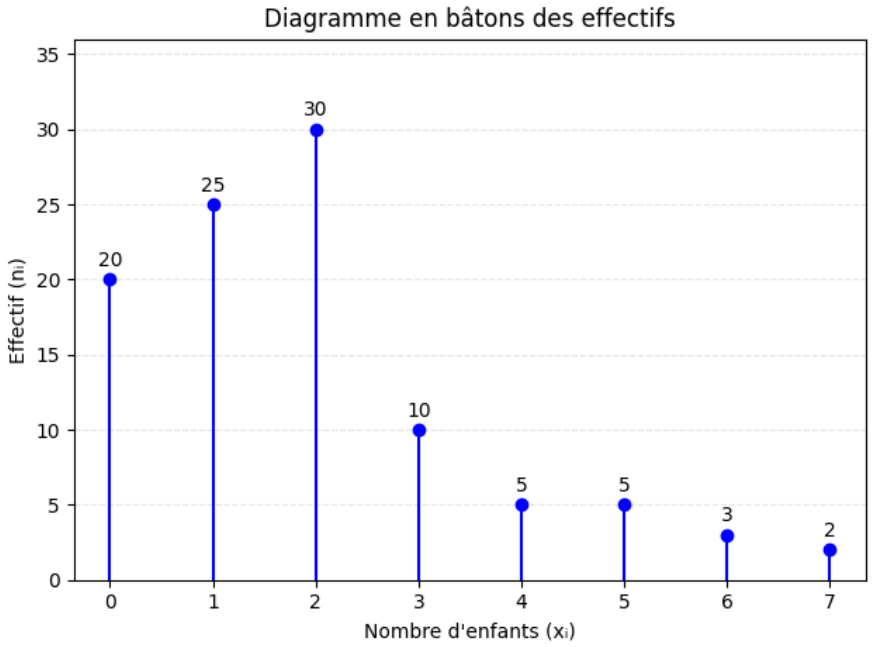

FIGURE 1 – Diagramme en bâtons des effectifs

Variables continues ou réparties en classes : Histogramme

Un histogramme est constitué de rectangles juxtaposés dont la base correspond à l'amplitude de chaque classe et dont la surface est proportionnelle à la fréquence absolue ou relative de cette classe.

L'histogramme est un outil statistique facile à utiliser, donnant rapidement une image du comportement d'un procédé industriel et l'allure globale de la distribution ; il montre l'étalement des données et apporte ainsi des renseignements sur la dispersion et sur les valeurs extrêmes ; il permet de déceler, éventuellement, des valeurs aberrantes.

Cas spécial de classes d'étendues différentes

Dans un histogramme statistique, c'est la surface des rectangles qui représentent les données, c'est-à-dire :

La hauteur de la barre multipliée par l'étendue de la classe est égale à l'effectif :

$$\text{hauteur de la barre} \times \text{étendue de la classe} = \text{effectif}$$

Donc, pour trouver la hauteur correct de chaque rectangle :

$$\text{hauteur} = \frac{\text{effectif}}{\text{largeur de la classe}}$$

Remarque : Cette hauteur calculée représente la densité de la classe.

Classe	Bornes	Largeur	Effectif	Hauteur
I	[0, 5[	5	10	2
II	[5, 15[	10	25	2.5
III	[15, 20[	5	30	6
IV	[20, 50[	30	20	0.66
V	[50, 55[	5	15	3

TABLE 2 – Tableau des classes avec largeurs inégales et effectifs.

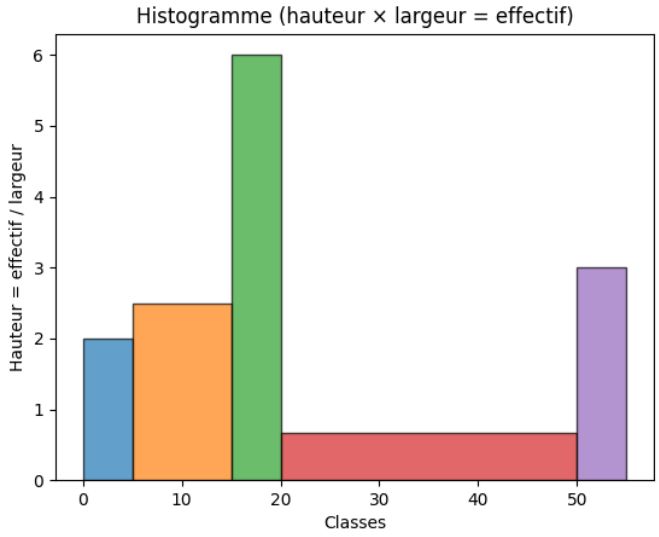

FIGURE 2 – Histogramme dans le cas de classes de différentes étendues

Il est toujours préférable d'avoir des classes de mêmes étendues, cela facilite grandement la représentation graphique.

Cas de classes de même étendue

Exemple : Étude de la dispersion d'un lot de 400 résistances

On a contrôlé 400 résistances dont la valeur nominale est égale à 100 kΩ et on a regroupé les résultats en classes d'amplitude 2 kΩ qui représente environ le dixième de la dispersion totale de l'échantillon contrôlé.

Classe	**Limites des classes**	n_i	N_i	f_i	F_i
I	[92, 94[	10	10	0,025	0,025
II	[94, 96[	15	25	0,0375	0,0625
III	[96, 98[	40	65	0,10	0,1625
IV	[98, 100[	60	125	0,15	0,3125
V	[100, 102[	90	215	0,225	0,5375
VI	[102, 104[	70	285	0,175	0,7125
VII	[104, 106[	50	335	0,125	0,8375
VIII	[106, 108[	35	370	0,0875	0,925
IX	[108, 110[	20	390	0,05	0,975
X	[110, 112[	10	400	0,025	1

TABLE 3 – Statistiques des données par classes.

Les classes étant toutes de même amplitude, l'histogramme est facile à tracer ; il suffit de construire des rectangles dont la hauteur est l'effectif des résistances de la classe correspondante.

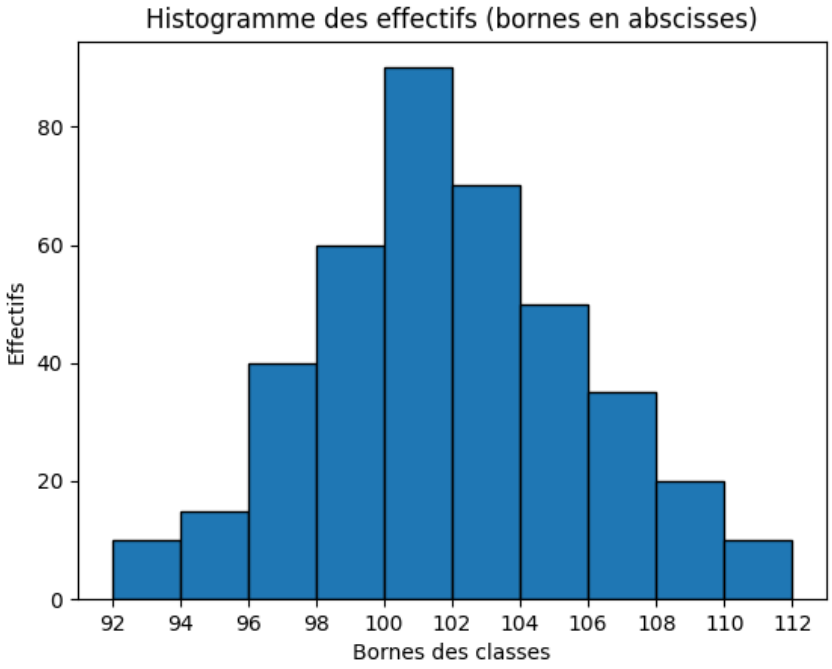

FIGURE 3 – Histogramme de la distribution

Le reste des graphiques peuvent être dessinés pour des variables discrètes et des variables continues. On a jusque-ici représenté les effectifs, on va voir comment représenter les fréquences.

Polygone de fréquences

Il permet de représenter sous forme de courbe, la distribution des effectifs ou fréquences. Il est obtenu en joignant, par des segments de droite, les milieux des côtés supérieurs de chaque rectangle de l'histogramme. (Pour fermer ce polygone, on ajoute à chaque extrémité une classe de fréquence nulle.)

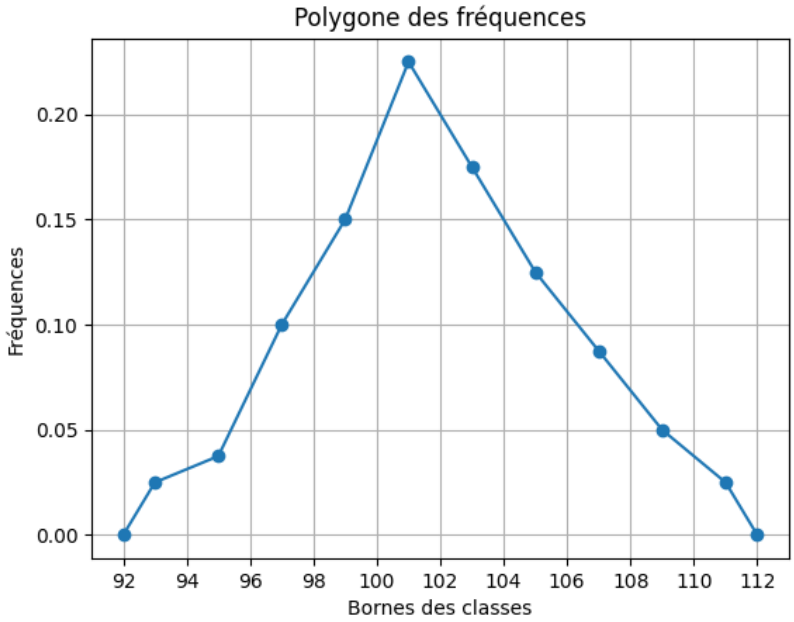

FIGURE 4 – Polygone des fréquences de l'exemple.

Remarque : On peut utiliser l'histogramme des effectifs pour dessiner le polygone des fréquences facilement, on rajoute un nouvel axe à droite pour les fréquences :

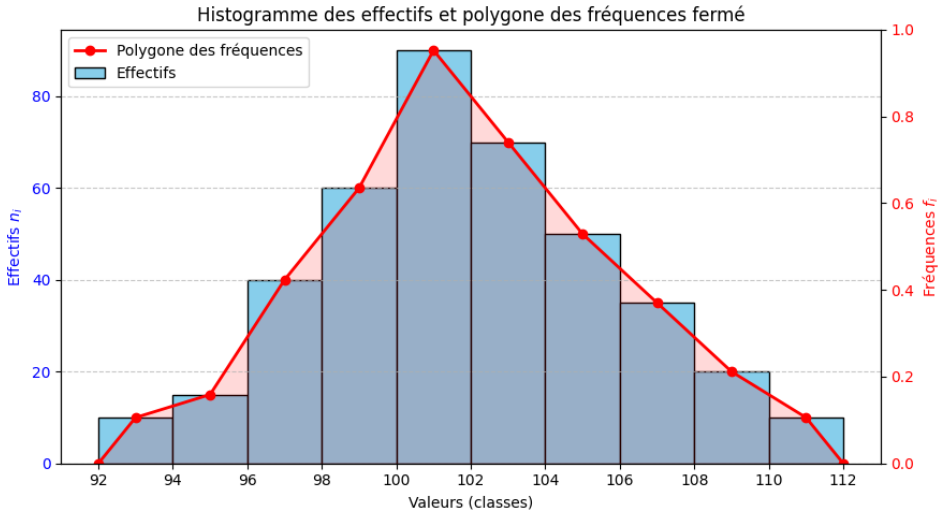

FIGURE 5 – Polygone des fréquences surimposé sur l'histogramme

Reste à représenter les fréquences cumulées.

Courbe de fréquences cumulées

On joint les points ayant pour abscisses la limite supérieure des classes et pour ordonnées les fréquences cumulées correspondant à la classe considérée (pour le premier point, on porte la valeur 0). Elle donne le nombre d'observations inférieures à une valeur quelconque de la série.

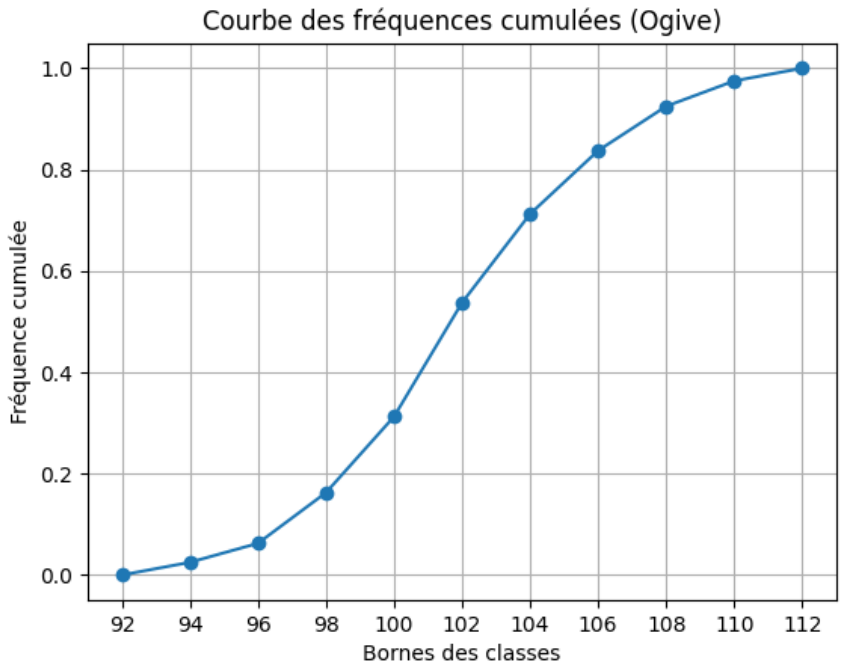

FIGURE 6 – Courbe cumulative croissante de la distribution de l'exemple.

Autres modes de représentations graphiques

On définit des diagrammes à secteurs circulaires et des diagrammes à rectangles horizontaux.

Le diagramme à secteurs circulaires consiste en un cercle découpé en secteurs circulaires ; l'aire de chaque secteur, représentant la proportion des différentes composantes d'un tout, est proportionnelle aux fréquences, relatives ou absolues.

3 Conclusion

— Résumé : Chaque représentation graphique (diagrammes à bandes, circulaire, en bâtons, polygone, histogramme, courbe cumulative) a un usage spécifique selon le type de variable.

— Applications en génie des matériaux : Visualisation des propriétés (dureté, résistance), contrôle qualité, comparaison de matériaux.

Cours de Statistiques pour Ingénieurs en Génie des Matériaux
Chapitre 3 : Représentation numérique des données

Objectifs de la séance

- Comprendre les caractéristiques de position : moyenne, médiane, mode, quartiles.
- Maîtriser les caractéristiques de dispersion : étendue, variance, écart-type, coefficient de variation.
- Explorer les caractéristiques de forme : asymétrie et aplatissement.
- Appliquer ces notions à des données en génie des matériaux.

Introduction

Une série de données peut être résumée par quelques valeurs numériques appelées caractéristiques des séries statistiques, classées en quatre grandes catégories :

- les caractéristiques de tendance centrale,
- les caractéristiques de dispersion,
- les caractéristiques de forme,
- les caractéristiques de concentration.

On verra des exemples des deux premières catégories dans ce cours.

1 Caractéristiques de tendance centrale

Elles donnent une idée de l'ordre de grandeur des valeurs constituant la série ainsi que la position où semblent se concentrer les valeurs de cette série. Les principales caractéristiques de tendance centrale sont la moyenne arithmétique, la médiane, le mode et les quantiles.

1.1 Moyenne arithmétique $\bar{x}$

La **moyenne arithmétique** est la somme des valeurs divisée par le nombre d'observations.

La moyenne n'est définie que pour les variables quantitatives.

Formule : Pour calculer la moyenne arithmétique, deux cas sont à distinguer selon la façon dont les données ont été recueillies.

Cas 1 : N données non réparties en classes, avec n_i l'effectif de chaque modalité x_i :

$$\bar{x} = \frac{1}{N}\sum_{i=1} n_i x_i = \sum_{i=1} f_i x_i$$

Cas 2 : N données réparties en k classes, la classe i étant d'effectif n_i et de fréquence f_i :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

avec c_i le centre de la classe i.

Exemple : Résistance à la traction (MPa) de 5 échantillons : 250, 260, 270, 255, 265. Moyenne : $\bar{x} = \frac{250+260+270+255+265}{5} = 260\,\text{MPa}$.

Exemple : On reprend l'exemple des résistances du chapitre précédent de l'étude de résistances

Classe	**Limites des classes**	**Centres** c_i	**Fréquences** f_i
I	[92, 94[	93	0,025
II	[94, 96[	95	0,0375
III	[96, 98[	97	0,10
IV	[98, 100[	99	0,15
V	[100, 102[	101	0,225
VI	[102, 104[	103	0,175
VII	[104, 106[	105	0,125
VIII	[106, 108[	107	0,0875
IX	[108, 110[	109	0,05
X	[110, 112[	111	0,025

TABLE 1 – Statistiques des données par classes.

On obtient alors la moyenne : $\bar{x} = 93 \times 0.025 + 95 \times 0.0375 + 97 \times 0.10 + 99 \times 0.15 + 101 \times 0.225 + 103 \times 0.175 + 105 \times 0.125 + 107 \times 0.875 + 109 \times 0.05 + 111 \times 0.025 =$ **186.1625 k**Ω

1.2 Médiane M_e

La **médiane** est la valeur, observée ou possible, dans la série des données classées par ordre croissant (ou décroissant) qui partage cette série en deux parties comprenant exactement le même nombre de données de part et d'autre de M_e.

Comme pour la moyenne arithmétique, on distingue deux cas.

Cas 1 : N données non réparties en classes :

— pour une série ayant un nombre impair de données, la médiane est une valeur observée de la série ;

— pour une série ayant un nombre pair de données, on peut prendre pour valeur médiane, indifféremment l'une ou l'autre des valeurs centrales ou n'importe quelle valeur intermédiaire entre ces deux valeurs, par exemple, la moyenne arithmétique de ces deux valeurs, mais, dans ces conditions, ce n'est pas une valeur observée.

Cas 2 : N données réparties en k classes. La médiane est obtenue :

— soit par interpolation linéaire à l'intérieur de la classe centrale, si le nombre de classes est impair,

— soit en prenant la moyenne des deux classes « centrales », si le nombre de classes est pair.

Exemple 1 : Données triées : 250, 255, 260, 265, 270. Médiane : 260 MPa.

Exemple 2 : Étude de deux séries d'observations
On considère les séries d'observations suivantes.
Série I : 5 observations classées par ordre croissant : 2, 5, 8, 11, 14
Moyenne arithmétique 8, médiane 8
Série II : 6 observations classées par ordre croissant : 6, 6, 14, 16, 18, 18
Moyenne arithmétique 13, médiane 15
Série III : On réunit les deux séries précédentes :
2, 5, 6, 6, 8, 11, 14, 14, 16, 18, 18
Moyenne arithmétique 10,72, médiane 11

Exemple 3 : Dans l'exemple des résistances, on peut trouver la valeur médiane par le graphique des fréquences cumulées, en projetant la valeur qui donne la la fréquence cumulée égale à $0,5$.

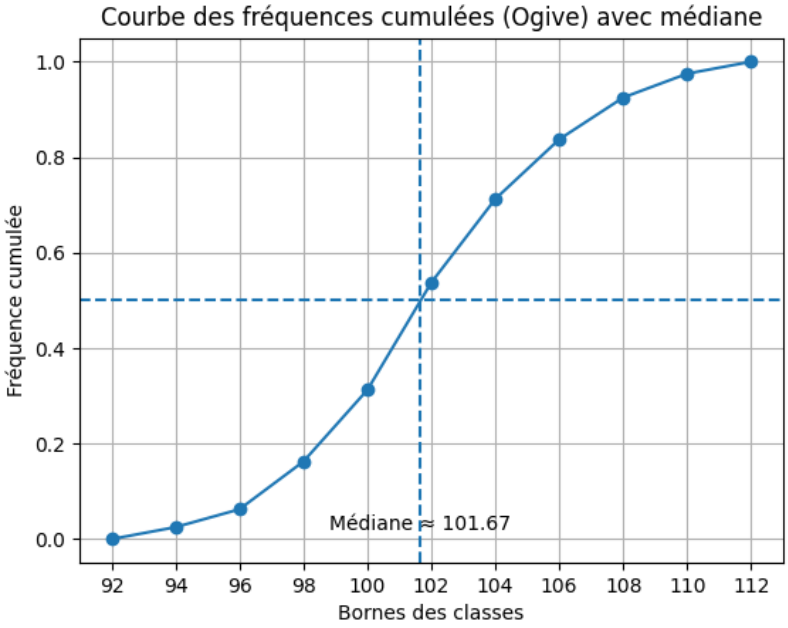

FIGURE 1 – Méthode graphique pour trouver la médiane.

On peut lire la médiane dans les abscisses : $M_e = 101,67$.
Formule : La médiane d'une distribution x est donnée par :

— pour les variables ordinales ou discrètes :
  — si la fréquence cumulée en $x_{i-1} < 0.5$ et celle en $x_i > 0.5$, alors la médiane vaut x_i.
  — si la fréquence cumulée en $x_{i-1} = 0.5$, alors la médiane vaut x_i.
— pour les variables réparties en classes $[a_{i-1}, a_i[$: si $F(a_{i-1} < 0.5$ et $F(a_i) > 0.5$, alors la classe médiane est $[a_{i-1}, a_i[$, et on calcule la valeur médiane par interpolation linéaire sur l'intervalle $[a_{i-1}, a_i[$:

$$M_e = a_{i-1} + (a_i - a_{i-1})\frac{0.5 - F(a_{i-1})}{F(a_i) - F(a_{i-1})}$$

1.3 Mode ou valeur dominante M_0

Le **mode** est la valeur de la variable statistique la plus fréquente que l'on observe dans une série d'observations.

Si la variable est une variable discrète, le mode s'obtient facilement. Si la variable est une variable continue, on définit une classe modale.

Exemple : Pour les données 250, 255, 260, 260, 265, le mode est 260 MPa.

Remarque : Le mode n'existe pas toujours et quand il existe, il n'est pas toujours unique.

Exemple : Suite de l'exemple 2
Série I : pas de mode.
Série II : deux modes 6 et 18.
Série III : les deux séries réunies, trois modes 6, 14 et 18.

Exemple : Dispersion d'un lot de 400 résistances (suite)
On ne peut pas définir une valeur modale en ne connaissant pas la distribution à l'intérieur de chaque classe.
On définit une classe modale, c'est la classe V.

1.4 Quantiles

Cette notion est très utilisée dans les sciences humaines.

Les **quantiles** sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.

Pour $k = 4$, les quantiles, appelés *quartiles*, sont trois nombres Q_1, Q_2, Q_3 tels que :

— 25 % des valeurs prises par la série sont inférieures à Q_1,

— 25 % des valeurs prises par la série sont supérieures à Q_3,

— Q_2 est la médiane M_e ,

— $Q_3 - Q_1$ est l'intervalle interquartile, il contient 50 % des valeurs de la série.

Pour $k = 10$, les quantiles sont appelés *déciles*, il y a neuf déciles D_1, D_2... 10 % des valeurs de la série sont inférieures à D_1...

Exemple : Pour les données triées,

$$250, 255, 260, 260, 265$$

Le premier quantile est $Q_1 = 252,5$ (moyenne de 250 et 255) ; Le deuxième quantile est $Q_2 = 260$ (médiane) ; et enfin, le troisième quantile est $Q_3 = 262,5$ (moyenne de 260 et 265).

Pour $k = 100$, les quantiles sont appelés centiles, il y a 99 centiles, chacun correspondant à 1 % de la population.

Exemple graphique : Pour trouver graphiquement les quartiles dans l'exemple de dispersion des résistance, on trouve l'abscisse qui a pour ordonnée 0.25, 0.5 et 0.75 en fréquence cumulée.

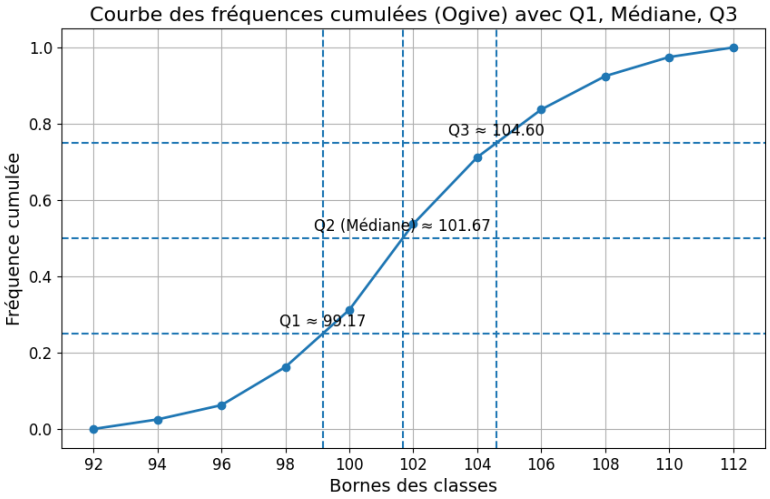

FIGURE 2 – Méthode graphique pour trouver les quartiles

2 Caractéristiques de dispersion

Ces caractéristiques quantifient les fluctuations des valeurs observées autour de la valeur centrale et permettent d'apprécier l'étalement de la série. Les principales sont : l'écart-type ou son carré appelé variance, le coefficient de variation et l'étendue.

Étendue

C'est la plus simple caractéristique de dispersion qu'on puisse avoir.

L'**étendue** est la différence entre la valeur maximale et minimale.

Formule : $E = x_{\text{max}} - x_{\text{min}}$.

Exemple : Pour 250, 255, 260, 265, 270, $E = 270 - 250 = 20\,\text{MPa}$.

Variance et écart-type

La **variance** d'un échantillon, notée s^2, est appelée aussi écart quadratique moyen ou variance empirique. La racine carrée de la variance est appelée *écart-type*.

Remarque : La variance est la moyenne de la somme des carrés des écarts par rapport à la moyenne arithmétique.

La moyenne arithmétique $\bar{x}$ et l'écart-type s s'expriment avec la même unité que les valeurs observées x_i. (La variance est alors mesurée en carré de l'unité des valeurs observées.)

— *Cas 1* : N données non réparties en classes, n_i l'effectif de chaque modalité x_i :

$$s^2 = \frac{1}{N}\sum_{i=1} n_i(x_i - \bar{x})^2$$

Formule simplifiée ne faisant apparaître que les données :

$$s^2 = \left(\frac{1}{N}\sum_{i=1} n_i x_i^2\right) - \bar{x}^2$$

La variance est donc égale à la moyenne des carrés moins le carré de la moyenne.

Exemple : Pour les données, $\bar{x} = 260$. Calcul de la variance :

$$s^2 = \frac{250^2 + 255^2 + 260^2 + 265^2 + 270^2}{5} - 260^2 = 50\,\text{MPa}^2.$$

Et l'écart-type est $s = \sqrt{50} \approx 7,07\,\text{MPa}$

— *Cas 2* : N données réparties en k classes, la classe i étant d'effectif n_i. On prend c_i le centre de la classe i.
Dans ces conditions, on obtient :

$$s^2 = \frac{1}{N}\sum_{i=1}^{k} n_i (c_i - \bar{x})^2$$

$$s^2 = \left(\frac{1}{N}\sum_{i=1}^{k} n_i c_i^2\right) - \bar{x}^2$$

Exemple

On considère la répartition suivante des notes d'un groupe de 20 élèves :

Classe de notes	Milieu de classe (x_i)	Effectif (n_i)
[0 ; 5[	2.5	2
[5 ; 10[	7.5	4
[10 ; 15[	12.5	6
[15 ; 20]	17.5	8

1. Calcul de la moyenne

$$\bar{x} = \frac{1}{N}\sum n_i x_i$$

$$\sum n_i x_i = (2 \times 2.5) + (4 \times 7.5) + (6 \times 12.5) + (8 \times 17.5) = 5 + 30 + 75 + 140 = 250$$

$$\boxed{\bar{x} = \frac{250}{20} = 12.5}$$

2. Calcul de la variance

$$s^2 = \frac{1}{N}\sum n_i (x_i - \bar{x})^2$$

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$
2.5	2	-10	100	200
7.5	4	-5	25	100
12.5	6	0	0	0
17.5	8	$+5$	25	200

$$\sum n_i(x_i - \bar{x})^2 = 2 \times 200 + 4 \times 100 + 6 \times 0 + 8 \times 200 = 2400$$

$$\boxed{s^2 = \frac{2400}{20} = 120}$$

L'écart-type est $s = \sqrt{120} = 10.95$

Remarque : L'écart-type s caractérise la dispersion d'une série de valeurs. Plus s est petit, plus les données sont regroupées autour de la moyenne arithmétique $\bar{x}$ et plus la population est homogène ; cependant avant de conclure, il faut faire attention à l'ordre de grandeur des données.

Exemple : Séries d'observations de l'exemple
Série I
Variance : $s^2 = \frac{1}{5}(2^2 + 5^2 + 8^2 + 11^2 + 14^2) - 8^2 = 18$
Ecart-type : $s = \sqrt{18} = 4.24$
Série II
Variance : $s^2 = 26.33$
Ecart-type : $s = 5.13$
Série III
Variance : $s^2 = 28.74$
Ecart-type : $s = 5.36$

Coefficient de variation

Coefficient de variation est le rapport de l'écart-type à la moyenne, exprimé en pourcentage.

Formule : $CV = \dfrac{\sigma}{\bar{x}} \times 100$.

Exemple : $CV = \frac{7{,}07}{260} \times 100 \approx 2,72\,\%$.

— Le coefficient de variation ne dépend pas des unités choisies.

— Il permet d'apprécier l'homogénéité de la distribution, une valeur du coefficient de variation inférieure à 15 % traduit une bonne homogénéité de la distribution.

Conclusion et discussion

— Résumé : Les caractéristiques de position, dispersion et forme permettent de décrire une distribution statistique de manière complète.

— Applications en génie des matériaux : Évaluation de la variabilité des propriétés (ex. : dureté, résistance), optimisation des procédés, contrôle qualité.

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

Objectifs de la séance

- Comprendre les séries statistiques à deux variables et leurs représentations.
- Maîtriser les distributions d'effectifs et de fréquences marginales et conditionnelles.
- Calculer et interpréter la covariance pour évaluer la relation entre deux variables.
- Appliquer ces notions à des données en génie des matériaux.

Durée : 1h30

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 $i = 1, \ldots, n$.

2 Tableau d'effectifs

On regroupe les valeurs possibles de X et de Y dans un tableau à double entrée.

	y_1	y_2	$\cdots$	y_m	Total ligne
x_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
Total colonne	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	n

où :

$$n_{i\cdot} = \sum_{j=1}^{m} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij}, \quad n = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}.$$

3 Fréquences

On définit les fréquences par :

$$f_{ij} = \frac{n_{ij}}{n},$$

Fréquences marginales

Elles représentent la répartition d'une variable indépendamment de l'autre :

$$f_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad f_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

$f_{i\cdot}$ = fréquence marginale de $X = x_i$, $f_{\cdot j}$ = fréquence marginale de $Y = y_j$.

Fréquences conditionnelles

Elles expriment la fréquence d'une variable en fonction d'une valeur de l'autre variable :

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

Exemple : probabilité que la dureté soit élevée ($Y = y_j$) sachant que la température est $X = x_i$.

4 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.

5 Vers le 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 Exemple numérique concret

Nous considérons l'étude de la relation entre la *température de traitement* X (en °C) et la *dureté* Y (en HV). On regroupe les observations dans le tableau d'effectifs suivant (total $n = 30$) :

$X \backslash Y$	150	200	250	Total ligne
600	2	3	0	5
700	1	6	3	10
800	0	2	13	15
Total colonne	3	11	16	30

Fréquences marginales

Les fréquences marginales (sur $n = 30$) :

$$f_{1\cdot} = \frac{5}{30} = \frac{1}{6}, \qquad f_{2\cdot} = \frac{10}{30} = \frac{1}{3}, \qquad f_{3\cdot} = \frac{15}{30} = \frac{1}{2},$$
$$f_{\cdot 1} = \frac{3}{30} = 0.1, \qquad f_{\cdot 2} = \frac{11}{30} \approx 0.3667, \qquad f_{\cdot 3} = \frac{16}{30} \approx 0.5333.$$

Moyennes

On pose $x_1 = 600,\ x_2 = 700,\ x_3 = 800$ et $y_1 = 150,\ y_2 = 200,\ y_3 = 250$.

$$\bar{x} = \sum_{i=1}^{3} x_i f_{i\cdot} = 600 \cdot \frac{1}{6} + 700 \cdot \frac{1}{3} + 800 \cdot \frac{1}{2} = \frac{2200}{3} \approx 733.3333 \text{ °C},$$

$$\bar{y} = \sum_{j=1}^{3} y_j f_{\cdot j} = 150 \cdot 0.1 + 200 \cdot \frac{11}{30} + 250 \cdot \frac{16}{30} = \frac{665}{3} \approx 221.6667 \text{ HV}.$$

Tableau des fréquences

$X \backslash Y$	150	200	250	Fréquence marginale $X = x_i$
600	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{0}{30}$	$\frac{5}{30}$
700	$\frac{1}{30}$	$\frac{6}{30}$	$\frac{3}{30}$	$\frac{10}{30}$
800	$\frac{0}{30}$	$\frac{2}{30}$	$\frac{13}{30}$	$\frac{15}{30}$
Fréquence marginale $Y = y_i$	$\frac{3}{30}$	$\frac{11}{30}$	$\frac{16}{30}$	1

Covariance

La covariance se calcule par

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{3} \sum_{j=1}^{3} (x_i - \bar{x})(y_j - \bar{y}) f_{ij},$$

où $f_{ij} = n_{ij}/30$. En remplaçant par les valeurs on obtient :

$$\mathrm{Cov}(X,Y) \;=\; \frac{16000}{9} \;\approx\; 1777.7777.$$

Variances et coefficient de corrélation

On calcule également les variances (marginales) :

$$s_X^2 = \sum_{i=1}^{3} (x_i - \bar{x})^2 f_{i\cdot} = \frac{50000}{9} \approx 5555.5556,$$

$$s_Y^2 = \sum_{j=1}^{3} (y_j - \bar{y})^2 f_{\cdot j} = \frac{10025}{9} \approx 1113.8889.$$

Les écarts-types sont donc

$$s_X \approx 74.5356, \qquad s_Y \approx 33.3750.$$

Le coefficient de corrélation linéaire de Pearson est

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y} \approx \frac{1777.7778}{74.5356 \times 33.3750} \approx 0.71465.$$

Interprétation

Le coefficient $r \approx 0.715$ indique une **corrélation positive assez forte** entre la température de traitement et la dureté : dans cet exemple, des températures plus élevées sont en moyenne associées à une dureté plus haute. Cela justifie, selon le cas concret, d'examiner des modèles linéaires ou des analyses plus fines (régression, test d'hypothèse) pour quantifier l'effet.

Remarque : les valeurs numériques sont arrondies pour la lisibilité ; dans un calcul formel on conservera les fractions exactes lorsque cela est souhaitable.

7 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.

Cours de Statistiques pour Ingénieurs en Génie des Matériaux
Chapitre 5 : Séries Statistiques à Deux Variables: Nuage de points et droite de tendance

Introduction

Ce document explique de façon générale comment construire un *nuage de points* et obtenir la *droite de tendance* (régression linéaire simple) pour une série statistique bidimensionnelle $(x_i, y_i)_{i=1,\dots,n}$. Une partie théorique est suivie d'un exemple numérique complet (taille et poids de 19 adolescents) avec figure.

1 Cas général : définitions et objectifs

Soit un échantillon de n individus, pour chaque individu on observe un couple (x_i, y_i) où X est la variable explicative (axe horizontal) et Y la variable expliquée (axe vertical).

1.1 Nuage de points

Le *nuage de points* est l'ensemble des points (x_i, y_i) tracés dans un repère orthonormé. Il permet :

— de visualiser la forme de la relation (croissante, décroissante, aucune relation) ;
— d'identifier la dispersion et les éventuelles valeurs aberrantes (outliers) ;
— de repérer si un modèle linéaire semble adapté.

1.2 Objectif de la droite de tendance

On cherche à approcher la relation entre X et Y par un modèle linéaire

$$\widehat{Y} = \alpha X + \beta,$$

où β (pente) et α (ordonnée à l'origine) sont choisis selon un critère (classiquement la méthode des moindres carrés, qui minimise la somme des carrés des résidus).

2 Formules générales (méthode des moindres carrés)

On définit les moyennes :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

La covariance (populationnelle, pondérée par $1/n$) :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}).$$

Les variances populationnelles :

$$\mathrm{V}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2, \qquad \mathrm{V}(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2.$$

La pente α et l'ordonnée à l'origine β s'obtiennent comme :

$$\alpha = \frac{\mathrm{Cov}(X,Y)}{\mathrm{V}(X)}, \qquad \beta = \bar{Y} - \alpha\bar{X}.$$

Le coefficient de corrélation :

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

3 Procédure pratique pour tracer le nuage et la droite

1. Tracer le nuage de points (x_i, y_i).
2. Calculer $\bar{X}$ et $\bar{Y}$.
3. Calculer $\mathrm{V}(X)$ et $\mathrm{C}(X,Y)$ (somme des produits centrés).
4. Déterminer α et β.
5. Tracer la droite en calculant $\widehat{Y}$ pour deux X extrêmes (ou en traçant la fonction $\alpha X + \beta$ sur l'intervalle visualisé).
6. Vérifier r pour juger la qualité de l'ajustement.

4 Exemple complet : taille et poids de 19 adolescents

Nous utilisons l'échantillon suivant (taille en cm ; poids en kg) :

$$\{(140, 38.2), (161, 44.3), (155, 46.1), (148, 38.2), (155, 50.5), (123, 22.4), (160, 40.4), (140, 34.7),$$
$$(165, 50.5), (172, 50.5), (155, 38.1), (160, 57.3), (142, 39.3), (157, 46.1), (142, 37.1), (148, 45.9),$$
$$(180, 66.3), (167, 60.0), (165, 50.5)\}.$$

4.1 Tableau des données (index, taille, poids)

i	Taille x_i (cm)	Poids y_i (kg)
1	140	38.2
2	161	44.3
3	155	46.1
4	148	38.2
5	155	50.5
6	123	22.4
7	160	40.4
8	140	34.7
9	165	50.5
10	172	50.5
11	155	38.1
12	160	57.3
13	142	39.3
14	157	46.1
15	142	37.1
16	148	45.9
17	180	66.3
18	167	60.0
19	165	50.5

4.2 Calculs (résultats numériques)

Les calculs (effectués pas à pas selon la procédure ci-dessus) donnent les valeurs numériques suivantes :

$$n = 19, \qquad \bar{X} \approx 154.4737 \text{ cm}, \qquad \bar{Y} \approx 45.0737 \text{ kg}.$$

$$\mathrm{Cov}(X, Y) \approx 113.2125 \text{ (cm·kg)},$$
$$\mathrm{V}(X) \approx 171.1967 \text{ cm}^2, \qquad \mathrm{V}(Y) \approx 96.1893 \text{ kg}^2.$$

Les écarts-types :

$$s_X = \sqrt{V(X)} \approx 13.0842 \text{ cm}, \qquad s_Y = \sqrt{V(Y)} \approx 9.8076 \text{ kg}.$$

Les paramètres de la droite :

$$\alpha = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)} \approx 0.6613006, \qquad \beta = \bar{Y} - \alpha \bar{X} \approx -57.0798566.$$

Donc la droite de tendance (prédiction de Y en fonction de X) s'écrit :

$$\widehat{Y} \approx 0.6613\, X - 57.0799.$$

Le coefficient de corrélation :

$$r \approx 0.88223$$

4.3 Interprétation

— La pente $\alpha \approx 0.6613$ signifie qu'une augmentation de 1 cm de taille est associée, en moyenne, à une augmentation d'environ 0.66 kg du poids.
— Le $r \approx 0.78$ indique que près de 78 % de la variance du poids est expliquée par la taille via ce modèle linéaire simple (dans cet échantillon).
— Le nuage de points et la droite aident à visualiser la force et la direction de la relation ; cependant, la corrélation n'implique pas causalité et il faut vérifier la présence éventuelle d'outliers ou d'effets non-linéaires.

5 Figure : nuage de points et droite de régression

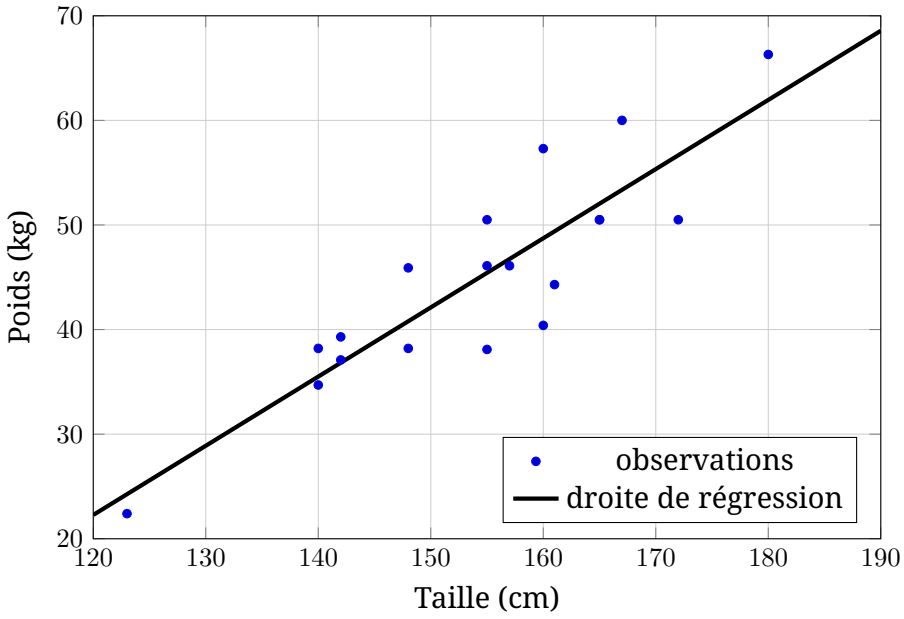

FIGURE 1 – Nuage de points (taille vs poids) et droite de tendance estimée par moindres carrés

6 Liaison linéaire, liaison non linéaire et absence de liaison

Cette section distingue trois situations fréquemment rencontrées lorsqu'on examine la relation entre deux variables : liaison linéaire, liaison non linéaire, et absence de liaison. L'objectif est d'aider à l'interprétation visuelle et statistique du nuage de points.

6.1 Liaison linéaire

Une *liaison linéaire* se reconnaît lorsque les points du nuage sont environ alignés (forme oblongue). Dans ce cas :

— la pente α est significative (valeur non nulle),
— le coefficient de corrélation r est proche de ± 1 pour une forte liaison,
— la droite de régression est un bon modèle descriptif (si les résidus ne montrent pas de structure).

6.2 Liaison non linéaire

Une *liaison non linéaire* existe lorsque la relation entre X et Y suit une forme courbe (parabole, exponentielle, puissance, etc.). Dans ce cas :

— la corrélation linéaire r peut être faible même si une relation forte existe ;
— il convient d'envisager des transformations (log, racine) ou des modèles non linéaires (polynômes, régression non linéaire) ;
— l'analyse des résidus révèle une structure systématique si l'on force une droite.

6.3 Absence de liaison

L'*absence de liaison* se traduit par un nuage très dispersé sans tendance apparente. Quelques remarques :

— $r \approx 0$ indique peu (ou pas) de liaison linéaire ;
— attention : $r \approx 0$ n'implique pas toujours indépendance — une structure non linéaire peut exister ;
— il faut examiner visuellement le nuage et vérifier les résidus.

6.4 Illustrations : trois mini-nuages

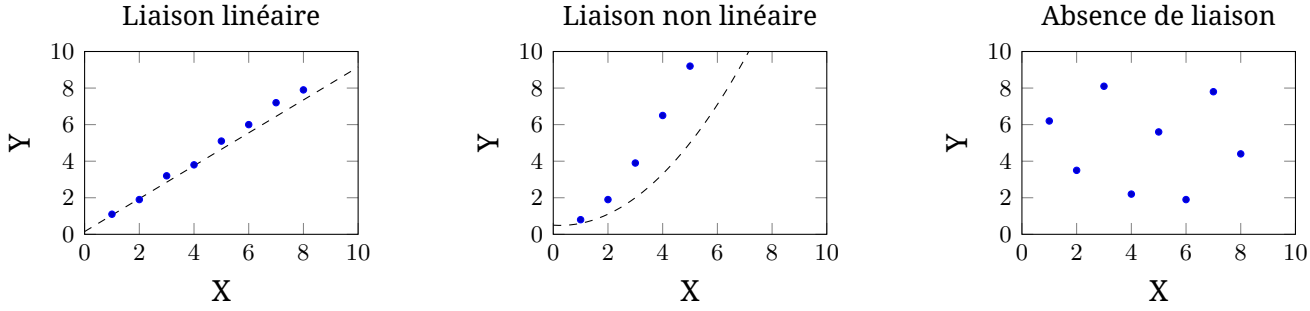

FIGURE 2 – Exemples visuels : liaison linéaire, liaison non linéaire, absence de liaison

7 Conclusion — bonnes pratiques

— Toujours vérifier la linéarité visuelle avant d'utiliser une droite de régression.
— Examiner les résidus : non-corrélation, homoscédasticité, indépendance.
— En ingénierie des matériaux, interpréter les résultats dans le contexte physique (mécanismes, contraintes de procédé, variabilité expérimentale).
— Utiliser R^2 avec prudence : une valeur élevée n'assure pas la validité du modèle au-delà de l'échantillon.