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 $0,5$.

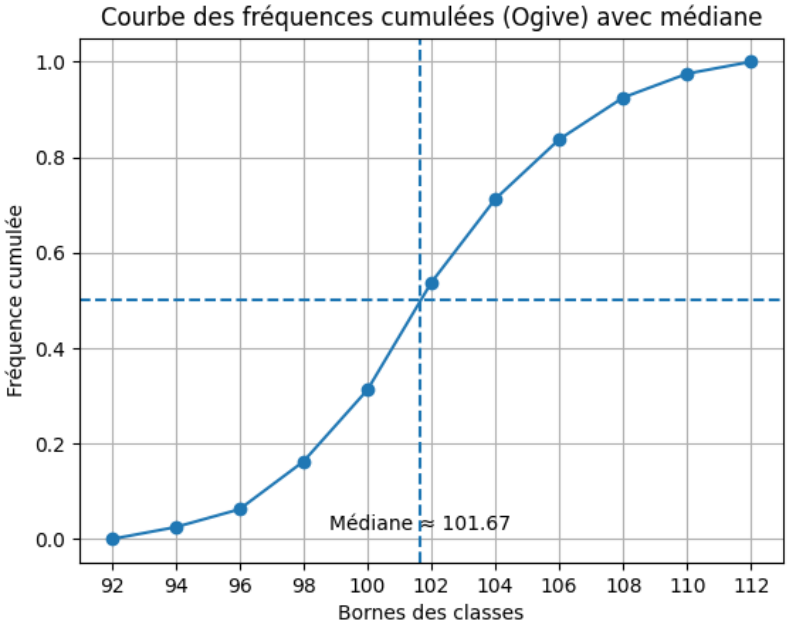


FIGURE 1 – Méthode graphique pour trouver la médiane.

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.

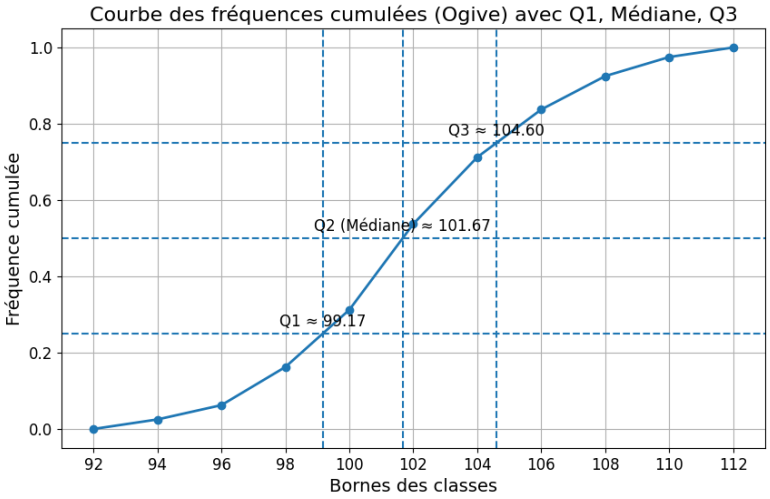


FIGURE 2 – Méthode graphique pour trouver les quartiles

1.5 Écart interquartile

L'**écart interquartile** (noté IQR ou $Q_3 - Q_1$) est la différence entre le troisième et le premier quartile. Il correspond à l'intervalle qui contient exactement 50 % des observations centrales de la série. Contrairement à l'étendue, qui est très sensible aux valeurs extrêmes, l'écart interquartile est une mesure de dispersion **robuste** : il n'est pas influencé par les valeurs aberrantes.

Dans l'exemple des résistances 250, 255, 260, 260, 265, on a $Q_1 = 252{,}5$ et $Q_3 = 262{,}5$, d'où un écart interquartile égal à 10 MPa. Cette valeur indique que la moitié centrale des mesures se concentre dans un intervalle de largeur 10 MPa autour de la médiane.

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.

2.1 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, $E = 270 - 250 = 20\,\text{MPa}$.

2.2 Variance et écart-type

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

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 : $s = 5.36$

2.3 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.

3 Caractéristiques de forme

Les caractéristiques de forme décrivent l'allure générale de la distribution. Elles permettent de quantifier l'asymétrie et le degré d'aplatissement (ou de concentration) des données autour de la moyenne. On utilise pour cela les moments centrés d'ordre 3 et 4.

3.1 Coefficient d'asymétrie (skewness)

Le moment centré d'ordre 3 est défini par :

$$\mu_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3$$

Le **coefficient d'asymétrie** (ou skewness) est alors :

$$\gamma_1 = \frac{\mu_3}{s^3}$$

Interprétation

— $\gamma_1 = 0$: distribution symétrique,

— $\gamma_1 > 0$: asymétrie à droite (queue longue vers les grandes valeurs),

— $\gamma_1 < 0$: asymétrie à gauche (queue longue vers les petites valeurs).

Exemple (données : 250, 255, 260, 260, 265)
Moyenne $\bar{x} = 260$, écart-type $s = \sqrt{50} \approx 7{,}071$.

$$\begin{aligned}\mu_3 &= \frac{1}{5}\Big[(250-260)^3 + (255-260)^3 + (260-260)^3 \\ &\quad + (260-260)^3 + (265-260)^3\Big] \\ &= \frac{1}{5}\big[(-10)^3 + (-5)^3 + 0 + 0 + 5^3\big] = \frac{-1000 - 125 + 125}{5} = -200\end{aligned}$$

$$\gamma_1 = \frac{-200}{(7{,}071)^3} \approx \frac{-200}{353{,}55} \approx -0{,}566$$

La distribution présente une légère asymétrie à gauche.

3.2 Coefficient d'aplatissement (kurtosis)

Le moment centré d'ordre 4 est défini par :

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4$$

Le **coefficient d'aplatissement** (ou kurtosis) est :

$$\gamma_2 = \frac{\mu_4}{s^4}$$

Interprétation

— $\gamma_2 = 3$: aplatissement identique à celui de la loi normale (distribution mésokurtique),

— $\gamma_2 > 3$: distribution plus pointue que la normale (leptokurtique),

— $\gamma_2 < 3$: distribution plus aplatie que la normale (platykurtique).

Exemple (suite)

$$\mu_4 = \frac{1}{5}\Big[(250-260)^4 + (255-260)^4 + 0 + 0 + (265-260)^4\Big]$$
$$= \frac{1}{5}\big[10^4 + 5^4 + 5^4\big] = \frac{10000 + 625 + 625}{5} = 2250$$
$$\gamma_2 = \frac{2250}{(7{,}071)^4} \approx \frac{2250}{2500} = 0{,}90$$

Comme $\gamma_2 \approx 0{,}90 < 3$, la distribution est nettement plus plate que la loi normale.

Remarque

Le calcul des moments d'ordre 3 et 4 devient rapidement fastidieux dès que N est grand. En pratique, on utilise un logiciel statistique. Ces coefficients restent néanmoins très utiles en génie des matériaux pour détecter des anomalies de distribution (asymétrie liée à un procédé, aplatissement trop faible ou trop fort indiquant une variabilité inhabituelle).

Conclusion

— 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.