MATHÉMATIQUES

10

Statistique descriptive

BTS | Mathématiques | Tous groupements

Je vais apprendre à :

- Maîtriser le vocabulaire statistique : population, individu, caractère, effectif, fréquence
- Construire et exploiter des tableaux de fréquences (données discrètes et groupées en classes)
- Calculer les indicateurs de position : moyenne, médiane, mode, quartiles Q_1 et Q_3
- Calculer les indicateurs de dispersion : étendue, variance, écart-type, coefficient de variation
- Construire et interpréter une boîte à moustaches (boxplot)
- Étudier une série statistique à deux variables : nuage de points, covariance, coefficient de corrélation r, droite de régression
- Réaliser un ajustement linéaire d'une série chronologique et effectuer des prévisions

Situation professionnelle

Contrôle qualité et suivi énergétique

Situation 1 — Contrôle qualité (B1, B2, C1) : Un ingénieur qualité dans une usine d'usinage de pièces métalliques prélève quotidiennement un échantillon de 50 pièces et mesure leur diamètre au centième de millimètre. Il doit vérifier que la production respecte les tolérances, détecter toute dérive et produire un rapport statistique mensuel.

Situation 2 — Énergie et bâtiment (B3, D1, D2) : Un responsable technique suit chaque mois la consommation électrique (en kWh) de plusieurs bâtiments tertiaires. Il analyse la dispersion des consommations, identifie les bâtiments énergivores et modélise l'évolution de la consommation en fonction du temps pour anticiper les besoins futurs.

Problème central : Comment résumer, analyser et communiquer efficacement une série de mesures ? Les outils de la statistique descriptive répondent à cette question.

1. Vocabulaire statistique

DÉFINITION — POPULATION ET INDIVIDU

La **population** est l'ensemble des éléments étudiés. Chaque élément est un **individu**.

Exemple : La population est l'ensemble des 840 pièces usinées en janvier. Chaque pièce est un individu.

DÉFINITION — CARACTÈRE ET SÉRIE STATISTIQUE

Le **caractère** (ou variable statistique) est la grandeur observée sur chaque individu.

— Un caractère **qualitatif** prend des valeurs non numériques (couleur, référence, état).

— Un caractère **quantitatif discret** prend des valeurs numériques isolées (nombre de défauts, rang).

— Un caractère **quantitatif continu** peut prendre toute valeur dans un intervalle (diamètre, température).

La **série statistique** est la liste des valeurs observées.

DÉFINITION — EFFECTIF ET FRÉQUENCE

L'**effectif** n_i d'une valeur x_i est le nombre d'individus qui prennent cette valeur.

L'**effectif total** est $N = \sum_i n_i$.

La **fréquence relative** de x_i est $f_i = \dfrac{n_i}{N}$.

La **fréquence en pourcentage** est $f_i \times 100$.

La **fréquence cumulée croissante** (FCC) est la somme des fréquences jusqu'à la valeur x_i.

2. Statistique à une variable — Tableaux et représentations

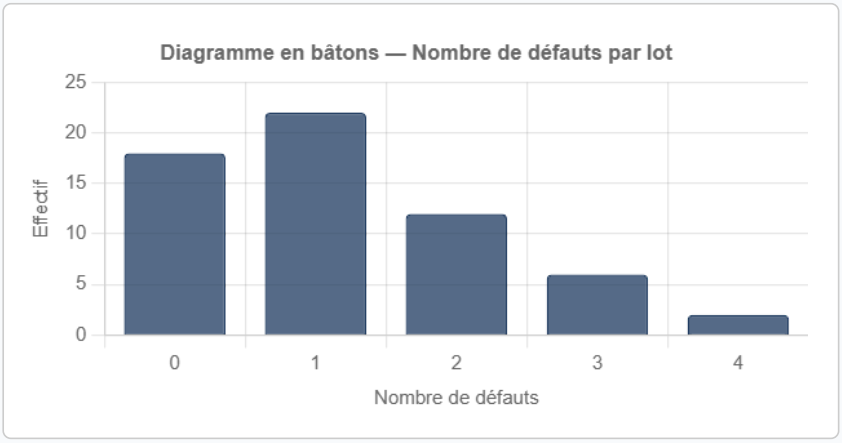
2.1 Série discrète — Tableau de fréquences

EXEMPLE — NOMBRE DE DÉFAUTS PAR LOT

Un technicien qualité relève le nombre de défauts observés sur 60 lots de production :

Nombre de défauts x_i	0	1	2	3	4	Total
Effectif n_i	18	22	12	6	2	60
Fréquence f_i	0,30	0,37	0,20	0,10	0,03	1,00
FCC	0,30	0,67	0,87	0,97	1,00	—

Le diagramme en bâtons représente les effectifs (ou fréquences) en fonction des valeurs.



2.2 Série continue — Tableau de classes

DÉFINITION — CLASSES ET AMPLITUDE

Lorsque le caractère est continu (ou que les valeurs sont très nombreuses), on regroupe les données en **classes** $[a_i\,;\,a_{i+1}[$.

L'**amplitude** de la classe est $a_{i+1} - a_i$.

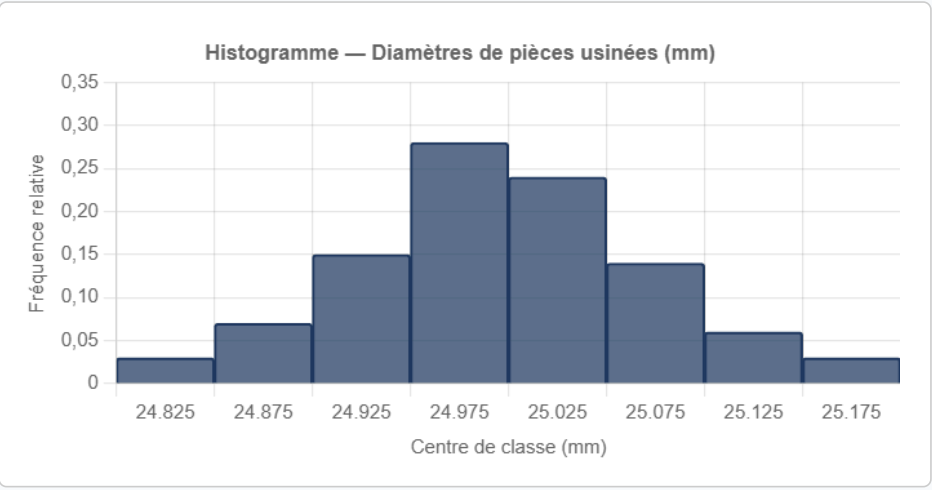
Le **centre de classe** est $c_i = \dfrac{a_i + a_{i+1}}{2}$.

EXEMPLE — DIAMÈTRE DE PIÈCES USINÉES (EN MM)

Un ingénieur qualité mesure le diamètre de 100 pièces. La cote nominale est 25,00 mm, tolérance ±0,15 mm.

Classe (mm)	Centre c_i	Effectif n_i	Fréquence f_i	FCC
[24,80 ; 24,85[	24,825	3	0,03	0,03
[24,85 ; 24,90[	24,875	7	0,07	0,10
[24,90 ; 24,95[	24,925	15	0,15	0,25
[24,95 ; 25,00[	24,975	28	0,28	0,53
[25,00 ; 25,05[	25,025	24	0,24	0,77
[25,05 ; 25,10[	25,075	14	0,14	0,91
[25,10 ; 25,15[	25,125	6	0,06	0,97
[25,15 ; 25,20[	25,175	3	0,03	1,00
Total	—	**100**	**1,00**	—

L'**histogramme** représente les fréquences (ou densités) en fonction des classes. La surface de chaque barre est proportionnelle à la fréquence.



3. Indicateurs de position

3.1 Moyenne

PROPRIÉTÉ — MOYENNE PONDÉRÉE

Pour une série de valeurs $x_1, x_2, \dots, x_k$ d'effectifs $n_1, n_2, \dots, n_k$:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \, x_i}{N}$$

où $N = \sum_{i=1}^{k} n_i$ est l'effectif total.

Pour des données groupées en classes, on remplace x_i par le centre de classe c_i.

EXEMPLE — CALCUL DE LA MOYENNE (DÉFAUTS PAR LOT)

Avec les données du tableau 2.1 :

$$\bar{x} = \frac{0 \times 18 + 1 \times 22 + 2 \times 12 + 3 \times 6 + 4 \times 2}{60} = \frac{0 + 22 + 24 + 18 + 8}{60} = \frac{72}{60} = 1{,}20$$

En moyenne, un lot présente **1,20 défaut**.

MINI-EXERCICE :

Un atelier relève le nombre de pièces rebutées par jour sur 5 jours : 4, 7, 5, 6, 8. Calculer la moyenne $\bar{x}$ du nombre de rebuts par jour.

EXEMPLE — CALCUL DE LA MOYENNE (CLASSES DE DIAMÈTRES)

En utilisant les centres de classe :

$$\bar{x} = \frac{24{,}825 \times 3 + 24{,}875 \times 7 + 24{,}925 \times 15 + 24{,}975 \times 28 + 25{,}025 \times 24 + 25{,}075 \times 14 + 25{,}125 \times 6 + 25{,}175 \times 3}{100}$$

$$\bar{x} = \frac{2499{,}70}{100} = 24{,}997 \text{ mm} \approx 25{,}00 \text{ mm}$$

3.2 Médiane

DÉFINITION — MÉDIANE ME

La **médiane** est la valeur qui partage la série en deux groupes d'effectifs égaux : 50 % des valeurs sont inférieures ou égales à Me et 50 % sont supérieures ou égales à Me.

— Sur données triées : si N est pair, Me est la moyenne des deux valeurs centrales ; si N est impair, Me est la valeur centrale.

— Sur tableau de fréquences cumulées : Me est la valeur pour laquelle FCC atteint 0,50.

MÉTHODE — MÉDIANE PAR INTERPOLATION LINÉAIRE (CLASSES)

On cherche la classe qui contient la médiane (FCC passe de moins de 0,50 à plus de 0,50). Si la classe médiane est $[a\,;\,b[$ et que la FCC avant cette classe est F_{avant}, l'effectif de la classe est n_c :

$$Me = a + \frac{0{,}50 - F_{\text{avant}}}{f_c} \times (b - a)$$

où $f_c = n_c / N$ est la fréquence de la classe médiane.

EXEMPLE — MÉDIANE DES DIAMÈTRES

La FCC passe de 0,25 à 0,53 dans la classe [24,95 ; 25,00[. C'est la classe médiane.

$$Me = 24{,}95 + \frac{0{,}50 - 0{,}25}{0{,}28} \times 0{,}05 = 24{,}95 + \frac{0{,}25}{0{,}28} \times 0{,}05 = 24{,}95 + 0{,}0446 \approx 24{,}995 \text{ mm}$$

3.3 Mode

DÉFINITION — MODE

Le **mode** est la valeur (ou la classe) qui présente le plus grand effectif.

— Pour une série discrète : la valeur la plus fréquente.

— Pour une série groupée : la **classe modale** est la classe d'effectif maximal ; le mode est son centre.

EXEMPLE

Pour les défauts par lot : le mode est **1** (effectif 22, le plus élevé).

Pour les diamètres : la classe modale est [24,95 ; 25,00[(effectif 28), donc le mode ≈ **24,975 mm**.

3.4 Quartiles Q1 et Q3

DÉFINITION — QUARTILES

Les **quartiles** partagent la série ordonnée en quatre groupes d'effectifs égaux (25 % chacun) :
— Q_1 (premier quartile) : 25 % des valeurs lui sont inférieures ou égales.
— $Q_2 = Me$ (deuxième quartile) : la médiane.
— Q_3 (troisième quartile) : 75 % des valeurs lui sont inférieures ou égales.

MÉTHODE — QUARTILES PAR INTERPOLATION LINÉAIRE (CLASSES)

Même méthode que pour la médiane, en remplaçant 0,50 par 0,25 (pour Q_1) ou 0,75 (pour Q_3) :

$$Q_1 = a + \frac{0{,}25 - F_{\text{avant}}}{f_c} \times (b - a) \qquad Q_3 = a + \frac{0{,}75 - F_{\text{avant}}}{f_c} \times (b - a)$$

MINI-EXERCICE :

Les durées de vie (en milliers d'heures) de 9 ampoules LED, triées, sont : 32, 35, 38, 40, 42, 44, 47, 50, 55.
Déterminer la médiane Me, puis Q_1 et Q_3 (par lecture directe sur la série triée).

EXEMPLE — QUARTILES DES DIAMÈTRES

Q1 : La FCC passe de 0,10 à 0,25 dans [24,90 ; 24,95[.

$$Q_1 = 24{,}90 + \frac{0{,}25 - 0{,}10}{0{,}15} \times 0{,}05 = 24{,}90 + \frac{0{,}15}{0{,}15} \times 0{,}05 = 24{,}90 + 0{,}05 = 24{,}950 \text{ mm}$$

Q3 : La FCC passe de 0,77 à 0,91 dans [25,05 ; 25,10[.

$$Q_3 = 25{,}05 + \frac{0{,}75 - 0{,}77}{0{,}14} \times 0{,}05$$

Attention : FCC avant la classe est 0,77 > 0,75, donc on remonte dans la classe précédente [25,00 ; 25,05[, FCC avant = 0,53.

$$Q_3 = 25{,}00 + \frac{0{,}75 - 0{,}53}{0{,}24} \times 0{,}05 = 25{,}00 + \frac{0{,}22}{0{,}24} \times 0{,}05 = 25{,}00 + 0{,}0458 \approx 25{,}046 \text{ mm}$$

4. Indicateurs de dispersion

4.1 Étendue et intervalle interquartile

DÉFINITION — ÉTENDUE ET IIQ

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

$$e = x_{\max} - x_{\min}$$

L'**intervalle interquartile** (IIQ) mesure la dispersion du « cœur » de la distribution (50 % des données centrales) :

$$\text{IIQ} = Q_3 - Q_1$$

Ces indicateurs sont robustes face aux valeurs extrêmes (outliers).

4.2 Variance et écart-type

PROPRIÉTÉ — VARIANCE ET ÉCART-TYPE

La **variance** mesure la dispersion quadratique moyenne des valeurs autour de la moyenne :

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

Formule équivalente (calcul simplifié) :

$$V = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^{k} n_i \, x_i^2}{N} - \bar{x}^2$$

L'**écart-type** est $\sigma = \sqrt{V}$. Il s'exprime dans la même unité que les données.

ATTENTION

La variance est en unité *au carré* (mm^2, kWh^2…), ce qui la rend difficile à interpréter directement. On lui préfère l'écart-type σ qui s'exprime dans la même unité que les données.

EXEMPLE — VARIANCE DES DÉFAUTS PAR LOT

On calcule d'abord $\overline{x^2}$:

$$\overline{x^2} = \frac{0^2 \times 18 + 1^2 \times 22 + 2^2 \times 12 + 3^2 \times 6 + 4^2 \times 2}{60} = \frac{0 + 22 + 48 + 54 + 32}{60} = \frac{156}{60} = 2{,}60$$

$$V = \overline{x^2} - \bar{x}^2 = 2{,}60 - (1{,}20)^2 = 2{,}60 - 1{,}44 = 1{,}16$$

$$\sigma = \sqrt{1{,}16} \approx 1{,}077 \text{ défaut}$$

MINI-EXERCICE :

Cinq mesures de tension (en V) relevées sur une ligne sont : 228, 230, 231, 229, 232. Calculer la moyenne $\bar{x}$, la variance V puis l'écart-type σ à l'aide de la formule $V = \overline{x^2} - \bar{x}^2$.

4.3 Coefficient de variation

DÉFINITION — COEFFICIENT DE VARIATION (CV)

Le **coefficient de variation** est un indicateur de dispersion relative, sans unité :

$$CV = \frac{\sigma}{\bar{x}} \times 100 \quad (\%)$$

Il permet de **comparer la dispersion de deux séries** ayant des unités ou des moyennes différentes.
Règle pratique : $CV < 15\%$: faible dispersion ; $15\% \leq CV \leq 30\%$: dispersion modérée ; $CV > 30\%$: forte dispersion.

EXEMPLE — COMPARAISON DE DEUX POSTES DE PRODUCTION

Poste A : $\bar{x}_A = 50{,}2$ mm, $\sigma_A = 0{,}8$ mm $\rightarrow CV_A = \frac{0{,}8}{50{,}2} \times 100 \approx 1{,}6\%$

Poste B : $\bar{x}_B = 12{,}5$ mm, $\sigma_B = 0{,}4$ mm $\rightarrow CV_B = \frac{0{,}4}{12{,}5} \times 100 = 3{,}2\%$

Bien que $\sigma_A > \sigma_B$, le poste A est **relativement moins dispersé** que le poste B.

Bilan — Indicateurs de dispersion

Indicateur	Formule	Interprétation
Étendue	$e = x_{\max} - x_{\min}$	Amplitude totale
IIQ	$Q_3 - Q_1$	Dispersion des 50 % centraux
Variance	$V = \overline{x^2} - \bar{x}^2$	Dispersion quadratique
Écart-type	$\sigma = \sqrt{V}$	Dispersion dans l'unité des données
Coef. de variation	$CV = \frac{\sigma}{\bar{x}} \times 100$	Dispersion relative (%)

5. Boîte à moustaches (Boxplot)

DÉFINITION — BOÎTE À MOUSTACHES

La **boîte à moustaches** (ou boxplot) est une représentation graphique synthétisant cinq indicateurs statistiques :

- Le minimum $x_{\min}$
- Le premier quartile Q_1
- La médiane Me
- Le troisième quartile Q_3
- Le maximum $x_{\max}$

La boîte (rectangle) s'étend de Q_1 à Q_3. Les moustaches s'étendent jusqu'aux valeurs extrêmes (dans la limite de 1,5 × IIQ). Les points au-delà des moustaches sont des **valeurs aberrantes** (outliers).

MÉTHODE — CONSTRUCTION D'UNE BOÎTE À MOUSTACHES

1. Calculer Q_1, Me, Q_3, $x_{\min}$, $x_{\max}$.
2. Tracer un axe gradué.
3. Dessiner un rectangle de Q_1 à Q_3, avec un trait vertical à Me.
4. Tracer les moustaches depuis Q_1 jusqu'à la plus petite valeur $\geq Q_1 - 1{,}5 \times \text{IIQ}$, et depuis Q_3 jusqu'à la plus grande valeur $\leq Q_3 + 1{,}5 \times \text{IIQ}$.
5. Représenter les éventuels outliers par des points isolés.

EXEMPLE — CONSOMMATIONS ÉNERGÉTIQUES MENSUELLES (KWH)

Un technicien relève les consommations de 12 bâtiments (valeurs triées) :

$$1\,240 — 1\,380 — 1\,450 — 1\,510 — 1\,590 — 1\,640 — 1\,720 — 1\,800 — 1\,860 — 1\,950 — 2\,010 — 3\,200$$

$N = 12$, $x_{\min} = 1\,240$, $x_{\max} = 3\,200$

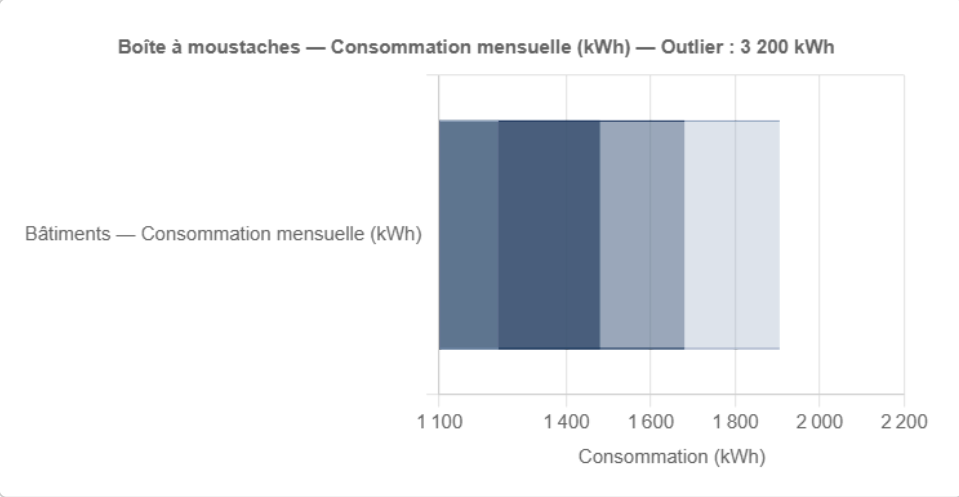
Quartiles :

- Q_1 = moyenne des valeurs de rang 3 et 4 = $\dfrac{1\,450 + 1\,510}{2} = 1\,480$ kWh
- Me = moyenne des valeurs de rang 6 et 7 = $\dfrac{1\,640 + 1\,720}{2} = 1\,680$ kWh
- Q_3 = moyenne des valeurs de rang 9 et 10 = $\dfrac{1\,860 + 1\,950}{2} = 1\,905$ kWh

$\text{IIQ} = 1\,905 - 1\,480 = 425$ kWh

Borne supérieure moustache : $Q_3 + 1{,}5 \times 425 = 1\,905 + 637{,}5 = 2\,542{,}5$ kWh

La valeur 3 200 kWh dépasse 2 542,5 kWh : c'est un **outlier**. Ce bâtiment est anormalement énergivore.



Interprétation : La majorité des bâtiments consomment entre 1 480 et 1 905 kWh par mois. Le bâtiment à 3 200 kWh devra faire l'objet d'un audit énergétique.

À retenir — Comparaison avec la boîte à moustaches

La boîte à moustaches est particulièrement utile pour **comparer plusieurs groupes** (plusieurs machines, plusieurs bâtiments, plusieurs périodes). Elle visualise instantanément les différences de position, de dispersion et la présence d'outliers.

6. Statistique à deux variables

6.1 Nuage de points et point moyen

DÉFINITION — SÉRIE À DEUX VARIABLES

Une **série statistique à deux variables** est un ensemble de n couples $(x_i\,;\,y_i)$, où x_i et y_i sont deux caractères quantitatifs mesurés sur le même individu i.
Exemples : (température extérieure, consommation de chauffage) ; (section du câble, résistance électrique) ; (année, chiffre d'affaires).

PROPRIÉTÉ — POINT MOYEN

Le **point moyen** $G(\bar{x}\,;\,\bar{y})$ a pour coordonnées les moyennes des deux séries :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Toute droite de régression passe par le point moyen G.

EXEMPLE — CONSOMMATION D'ÉNERGIE ET SURFACE DE VITRAGE

Un bureau d'études thermique analyse 8 bâtiments de taille comparable. Pour chacun, on note la surface de vitrage x (en m^2) et la déperdition thermique annuelle y (en MWh) :

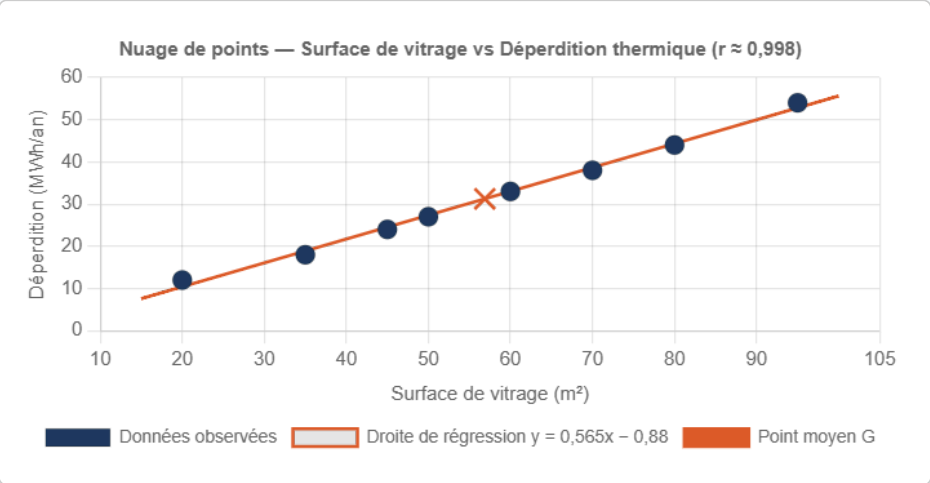
Bâtiment	1	2	3	4	5	6	7	8
Surface vitrage x (m^2)	20	35	45	50	60	70	80	95
Déperdition y (MWh)	12	18	24	27	33	38	44	54

Calcul du point moyen :

$$\bar{x} = \frac{20+35+45+50+60+70+80+95}{8} = \frac{455}{8} = 56{,}875 \text{ m}^2$$

$$\bar{y} = \frac{12+18+24+27+33+38+44+54}{8} = \frac{250}{8} = 31{,}25 \text{ MWh}$$

Le point moyen est $G(56{,}875\,;\,31{,}25)$.



6.2 Covariance et coefficient de corrélation linéaire

DÉFINITION — COVARIANCE

La **covariance** mesure la tendance des deux variables à varier simultanément dans le même sens ou en sens opposé :

$$\text{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y}$$

— $\text{Cov}(x,y) > 0$: les deux variables tendent à croître ensemble.
— $\text{Cov}(x,y) < 0$: quand l'une croît, l'autre décroît.
— $\text{Cov}(x,y) = 0$: pas de relation linéaire apparente.

DÉFINITION — COEFFICIENT DE CORRÉLATION LINÉAIRE R

Le **coefficient de corrélation linéaire** de Pearson est :

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \cdot \sigma_y}$$

où σ_x et σ_y sont les écarts-types de chaque série.

Propriété : $-1 \leq r \leq 1$.

— r proche de 1 : forte corrélation linéaire positive.

— r proche de –1 : forte corrélation linéaire négative.

— $|r|$ proche de 0 : pas de corrélation linéaire.

En pratique : $|r| \geq 0{,}9$ est considéré comme une bonne corrélation linéaire.

EXEMPLE — CALCUL DE LA COVARIANCE ET DE R (VITRAGE/DÉPERDITION)

Tableau de calcul :

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	20	12	400	144	240
2	35	18	1 225	324	630
3	45	24	2 025	576	1 080
4	50	27	2 500	729	1 350
5	60	33	3 600	1 089	1 980
6	70	38	4 900	1 444	2 660
7	80	44	6 400	1 936	3 520
8	95	54	9 025	2 916	5 130
Total	**455**	**250**	**30 075**	**9 158**	**16 590**

$$\overline{x^2} = \frac{30\,075}{8} = 3\,759{,}375 \; ; \; \overline{y^2} = \frac{9\,158}{8} = 1\,144{,}75 \; ; \; \overline{xy} = \frac{16\,590}{8} = 2\,073{,}75$$

Covariance :

$$\mathrm{Cov}(x,y) = \overline{xy} - \bar{x}\,\bar{y} = 2\,073{,}75 - 56{,}875 \times 31{,}25 = 2\,073{,}75 - 1\,777{,}34 = 296{,}41$$

Écarts-types :

$$\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{3\,759{,}375 - (56{,}875)^2} = \sqrt{3\,759{,}375 - 3\,234{,}77} = \sqrt{524{,}61} \approx 22{,}90 \text{ m}^2$$

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{1\,144{,}75 - (31{,}25)^2} = \sqrt{1\,144{,}75 - 976{,}56} = \sqrt{168{,}19} \approx 12{,}97 \text{ MWh}$$

Coefficient de corrélation :

$$r = \frac{296{,}41}{22{,}90 \times 12{,}97} = \frac{296{,}41}{297{,}02} \approx 0{,}998$$

Le coefficient $r \approx 0{,}998$ est très proche de 1 : il y a une **très forte corrélation linéaire positive** entre la surface de vitrage et les déperditions thermiques.

MINI-EXERCICE :

Pour 5 chantiers, on relève la température extérieure x (°C) et la consommation de chauffage y (kWh) : $(2\,;\,90)$, $(4\,;\,80)$, $(6\,;\,70)$, $(8\,;\,60)$, $(10\,;\,50)$. Calculer $\bar{x}$, $\bar{y}$, la covariance $\mathrm{Cov}(x,y)$, puis la pente a de la droite de régression de y en x.

6.3 Droite de régression par la méthode des moindres carrés

PROPRIÉTÉ — DROITE DE RÉGRESSION DE Y EN X

La droite de régression de y en x, obtenue par la méthode des **moindres carrés**, minimise la somme des carrés des écarts entre les valeurs observées et les valeurs prédites.

Son équation est $y = ax + b$ avec :

$$a = \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{\text{Cov}(x,y)}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - a\,\bar{x}$$

Cette droite passe obligatoirement par le point moyen $G(\bar{x}\,;\,\bar{y})$.

MÉTHODE — CALCUL DE LA DROITE DE RÉGRESSION

1. Calculer $\bar{x}$, $\bar{y}$, $\overline{xy}$, $\overline{x^2}$.
2. Calculer $\text{Cov}(x,y) = \overline{xy} - \bar{x}\,\bar{y}$ et $\sigma_x^2 = \overline{x^2} - \bar{x}^2$.
3. Calculer $a = \dfrac{\text{Cov}(x,y)}{\sigma_x^2}$.
4. Calculer $b = \bar{y} - a\,\bar{x}$.
5. Écrire l'équation $y = ax + b$ et l'utiliser pour faire des prévisions.

EXEMPLE — DROITE DE RÉGRESSION (VITRAGE/DÉPERDITION)

On reprend les calculs précédents :

$$a = \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{296{,}41}{524{,}61} \approx 0{,}565 \text{ MWh/m}^2$$

$$b = \bar{y} - a\,\bar{x} = 31{,}25 - 0{,}565 \times 56{,}875 = 31{,}25 - 32{,}13 = -0{,}88 \text{ MWh}$$

Équation de la droite de régression :

$$\boxed{y = 0{,}565\,x - 0{,}88}$$

Vérification : pour $x = \bar{x} = 56{,}875$: $y = 0{,}565 \times 56{,}875 - 0{,}88 = 32{,}13 - 0{,}88 = 31{,}25 = \bar{y}$ ✓

Prévision : Pour un bâtiment avec 65 m² de vitrage :

$$y = 0{,}565 \times 65 - 0{,}88 = 36{,}73 - 0{,}88 \approx 35{,}8 \text{ MWh/an}$$

ATTENTION

La droite de régression de y en x est différente de la droite de régression de x en y. On utilise :

— $y = ax + b$ pour **prédire y à partir de x** (connaissant x).

— $x = a'y + b'$ pour **prédire x à partir de y**, avec $a' = \dfrac{\text{Cov}(x,y)}{\sigma_y^2}$.

À retenir — Condition d'utilisation de la régression

On n'utilise le modèle de régression linéaire que si $|r| \geq 0{,}9$ (corrélation forte). Sinon, le modèle linéaire est inapproprié et peut conduire à des prévisions très erronées.

7. Série chronologique — Ajustement linéaire

DÉFINITION — SÉRIE CHRONOLOGIQUE

Une **série chronologique** est une série statistique à deux variables dans laquelle la variable x représente le **temps** (mois, trimestre, année, rang…) et y la grandeur observée à chaque instant.

On distingue :

— La **tendance** (trend) : évolution à long terme.

— Les **variations saisonnières** : oscillations périodiques.

— Les **variations résiduelles** : perturbations aléatoires.

MÉTHODE — AJUSTEMENT LINÉAIRE D'UNE SÉRIE CHRONOLOGIQUE

On code le temps : on attribue le rang 1 à la première observation, 2 à la deuxième, etc. Puis on applique la méthode des moindres carrés pour trouver $y = at + b$ (tendance linéaire).

EXEMPLE — CHIFFRE D'AFFAIRES D'UNE ENTREPRISE DE MENUISERIE (K€)

Un dirigeant d'entreprise spécialisée dans l'agencement intérieur relève le chiffre d'affaires annuel sur 7 ans :

Année	2018	2019	2020	2021	2022	2023	2024
Rang t	1	2	3	4	5	6	7
CA y (k€)	312	338	295	371	408	445	482

Calculs préliminaires :

t	y	t^2	ty
1	312	1	312
2	338	4	676
3	295	9	885
4	371	16	1 484
5	408	25	2 040
6	445	36	2 670
7	482	49	3 374
28	**2 651**	**140**	**11 441**

$$\bar{t} = \frac{28}{7} = 4 \qquad \bar{y} = \frac{2\,651}{7} = 378{,}71 \text{ k€}$$

$$\overline{t^2} = \frac{140}{7} = 20 \qquad \overline{ty} = \frac{11\,441}{7} = 1\,634{,}43$$

$$\mathrm{Cov}(t,y) = \overline{ty} - \bar{t}\,\bar{y} = 1\,634{,}43 - 4 \times 378{,}71 = 1\,634{,}43 - 1\,514{,}84 = 119{,}59$$

$$\sigma_t^2 = \overline{t^2} - \bar{t}^2 = 20 - 16 = 4$$

$$a = \frac{119{,}59}{4} = 29{,}90 \text{ k€/an}$$

$$b = \bar{y} - a\,\bar{t} = 378{,}71 - 29{,}90 \times 4 = 378{,}71 - 119{,}60 = 259{,}11 \text{ k€}$$

Tendance linéaire :

$$\boxed{y = 29{,}90\,t + 259{,}11}$$

Vérification du coefficient de corrélation :

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2}$$

$$\overline{y^2} = \frac{312^2 + 338^2 + 295^2 + 371^2 + 408^2 + 445^2 + 482^2}{7} = \frac{1\,033\,067}{7} \approx 147\,581$$

$$\sigma_y = \sqrt{152\,833{,}86 - (378{,}71)^2} = \sqrt{147\,581 - 143\,421{,}46} = \sqrt{4\,159{,}54} \approx 64{,}49$$

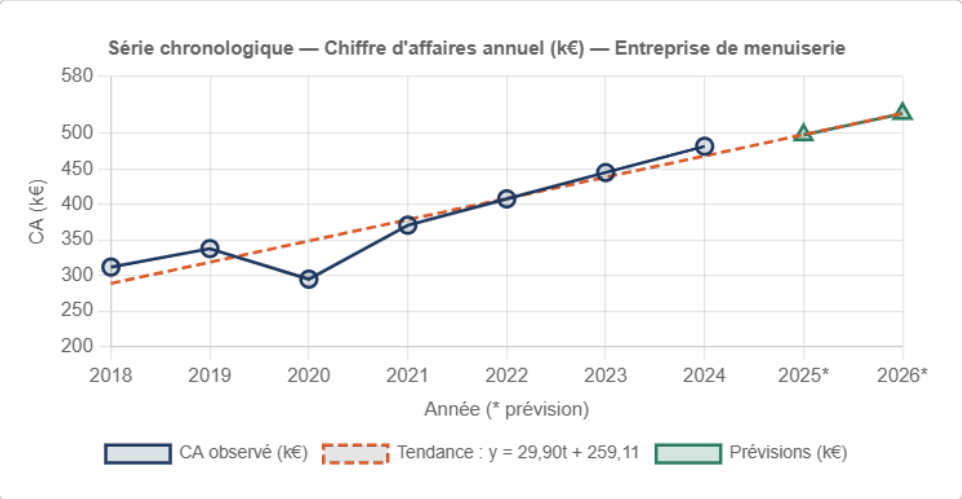
$$r = \frac{119{,}59}{\sigma_t \times \sigma_y} = \frac{119{,}59}{2 \times 64{,}49} = \frac{119{,}59}{128{,}98} \approx 0{,}927$$

$r \approx 0{,}986 \geq 0{,}9$: l'ajustement linéaire est **très bien adapté**.

Prévision pour 2025 (rang $t = 8$) et 2026 (rang $t = 9$) :

$$y_{2025} = 29{,}90 \times 8 + 259{,}11 = 239{,}20 + 259{,}11 = 498{,}31 \text{ k€}$$

$$y_{2026} = 29{,}90 \times 9 + 259{,}11 = 269{,}10 + 259{,}11 = 528{,}21 \text{ k€}$$



ATTENTION — EXTRAPOLATION

Toute prévision hors de la plage des données observées (extrapolation) doit être interprétée avec prudence. Le modèle linéaire suppose que la tendance observée se poursuit, ce qui peut ne pas être vérifié à long terme (retournement de conjoncture, saturation du marché, etc.).

8. Récapitulatif — Formules essentielles

Statistique à une variable

Indicateur	Formule
Moyenne	$\bar{x} = \dfrac{\sum n_i x_i}{N}$
Variance	$V = \dfrac{\sum n_i x_i^2}{N} - \bar{x}^2$
Écart-type	$\sigma = \sqrt{V}$
Coefficient de variation	$CV = \dfrac{\sigma}{\bar{x}} \times 100$
Intervalle interquartile	$\text{IIQ} = Q_3 - Q_1$

Statistique à deux variables

Indicateur	Formule
Covariance	$\mathrm{Cov}(x,y) = \overline{xy} - \bar{x}\,\bar{y}$
Coefficient de corrélation	$r = \dfrac{\mathrm{Cov}(x,y)}{\sigma_x \cdot \sigma_y}$
Pente de régression	$a = \dfrac{\mathrm{Cov}(x,y)}{\sigma_x^2}$
Ordonnée à l'origine	$b = \bar{y} - a\,\bar{x}$

À retenir — Démarche statistique complète

1. **Décrire la population** et le caractère étudié.
2. **Construire le tableau** de fréquences (et calculer les FCC).
3. **Calculer les indicateurs de position** : $\bar{x}$, Me, mode, Q_1, Q_3.
4. **Calculer les indicateurs de dispersion** : σ, CV, IIQ.
5. **Représenter graphiquement** (histogramme, boîte à moustaches).
6. **Interpréter** dans le contexte professionnel.
7. Si étude à 2 variables : calculer r, et si $|r| \geq 0{,}9$, déterminer la droite de régression.

MATHÉMATIQUES

10

Statistique descriptive

BTS | Mathématiques | Durée : 40 min | /20

Nom : _____ Prénom : _____ Date : _____

Exercice 1 — Moyenne d'une série discrète (3 pts)

Un atelier relève le nombre de pièces rebutées par jour sur 5 jours : 4, 6, 5, 7, 8.

a. Calculer la moyenne $\bar{x}$. *(1,5 pt)*
b. Calculer $\overline{x^2}$. *(1,5 pt)*

Exercice 2 — Médiane et quartiles (4 pts)

Les durées de vie (en milliers d'heures) de 11 modules LED, triées, sont :

Rang	1	2	3	4	5	6	7	8	9	10	11
Valeur	12	15	18	20	23	25	28	30	34	38	42

a. Déterminer la médiane Me. *(1,5 pt)*
b. Déterminer Q_1 et Q_3 (lecture directe). *(1,5 pt)*
c. Calculer l'intervalle interquartile (IIQ). *(1 pt)*

Exercice 3 — Variance, écart-type, coefficient de variation (4 pts)

Cinq mesures de tension (en V) relevées sur une ligne sont : 196, 198, 200, 202, 204.

a. Calculer la moyenne $\bar{x}$. *(1 pt)*
b. Calculer la variance V et l'écart-type σ. *(2 pts)*
c. Calculer le coefficient de variation CV. Qualifier la dispersion. *(1 pt)*

Exercice 4 — Boîte à moustaches et outlier (4 pts)

Les consommations électriques mensuelles (en kWh) de 9 bâtiments, triées, sont :

Rang	1	2	3	4	5	6	7	8	9
kWh	120	135	142	150	160	168	175	182	260

a. Déterminer Q_1, Me, Q_3 (lecture directe). *(1,5 pt)*

b. Calculer l'IIQ, puis la borne supérieure de moustache $Q_3 + 1{,}5 \times \text{IIQ}$. *(1,5 pt)*

c. La valeur 260 kWh est-elle une valeur aberrante (outlier) ? Justifier. *(1 pt)*

Exercice 5 — Droite de régression (5 pts)

Un bureau d'études suit le chiffre d'affaires y (en k€) d'une entreprise d'agencement sur 5 ans, le temps étant codé par son rang t :

Rang t	1	2	3	4	5
CA y (k€)	30	34	36	42	48

a. Calculer $\bar{t}$, $\bar{y}$, $\overline{ty}$ et $\overline{t^2}$. *(2 pts)*

b. Calculer $\text{Cov}(t, y)$ et σ_t^2, puis la pente a et l'ordonnée à l'origine b. *(2 pts)*

c. Donner l'équation $y = at + b$ et prévoir le CA pour $t = 6$. *(1 pt)*