Fiche Méthodes - SM403

Un recueil de méthodes pour accompagner les révisions d'Analyse de Données

Doryan Denis

2025/2026

Table des matières

1 Statistique descriptive

Ce chapitre rappelle les outils de description d'un échantillon : paramètres résumant une variable, liens entre deux variables et régression linéaire, puis manipulation des matrices de données multivariées.

1.1 Calculer les paramètres de tendance centrale

Méthode.
Soit un échantillon de n observations $x_1, \dots, x_n$ d'une variable X.

1. **Moyenne empirique** :
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
2. **Médiane** : valeur m telle que la moitié des observations lui est inférieure ou égale. Pratiquement, on trie les données par ordre croissant.
   - Si n est impair, la médiane est la valeur de rang $\frac{n+1}{2}$.
   - Si n est pair, la médiane est la moyenne des valeurs de rangs $\frac{n}{2}$ et $\frac{n}{2}+1$.
3. **Mode** : valeur la plus fréquente dans l'échantillon (peut ne pas être unique).

Remarque. Données groupées par classes
Lorsque les données sont présentées sous forme de classes $[a_j, b_j[$ avec effectif f_j, on utilise le centre de classe $c_j = \frac{a_j+b_j}{2}$ et la formule :
$$\bar{x} = \frac{\sum_j f_j \, c_j}{\sum_j f_j}$$

Exemple. Application — CE 2026, Ex. 2
On dispose de quatre observations décrites par trois variables X_1, X_2, X_3 :
$$(15, 12, 17), \quad (10, 7, 13), \quad (8, 3, 10), \quad (15, 8, 15)$$
Calculer la médiane, le mode et l'étendue de X_1.

Solution.
On extrait les valeurs de X_1 : $15,\ 10,\ 8,\ 15$.
Moyenne : $\bar{x}_1 = \dfrac{15+10+8+15}{4} = \dfrac{48}{4} = \boxed{12}$
Médiane : on trie les valeurs : $8 < 10 < 15 = 15$. L'échantillon est de taille $n = 4$ (pair), donc :
$$\text{med}(X_1) = \frac{x_{(2)} + x_{(3)}}{2} = \frac{10+15}{2} = \boxed{12{,}5}$$
Mode : la valeur 15 apparaît deux fois, c'est la plus fréquente. $\boxed{\text{mode} = 15}$
Étendue : $e = 15 - 8 = \boxed{7}$

1.2 Calculer les paramètres de dispersion

Méthode.
Soit un échantillon $x_1, \ldots, x_n$ de moyenne $\bar{x}$ et d'écart-type σ.

1. **Variance empirique** :
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$
2. **Écart-type** : $\sigma = \sqrt{\sigma^2}$.
3. **Coefficient d'asymétrie** (Skewness) :
$$S_k = \frac{1}{n\,\sigma^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$
4. **Coefficient d'aplatissement** (Kurtosis) :
$$K = \frac{1}{n\,\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3$$

Remarque. Interprétation de S_k et K

- $S_k > 0$: distribution asymétrique à **droite** (queue longue à droite) ; $S_k < 0$: asymétrique à **gauche** ; $S_k \approx 0$: distribution symétrique.
- $K > 0$: distribution **leptocurtique** (pic plus aigu qu'une loi normale) ; $K < 0$: **platycurtique** (plus aplatie) ; $K = 0$: **mésocurtique** (comme une normale).

Exemple. Application — CE 2026, Ex. 2
Reprendre les données de l'exemple précédent (X_1 : $15, 10, 8, 15$, $\bar{x}_1 = 12$). Calculer la variance, l'écart-type et le coefficient d'asymétrie S_k de X_1, puis interpréter.

Solution.
Variance :
$$\sigma_1^2 = \frac{(15-12)^2 + (10-12)^2 + (8-12)^2 + (15-12)^2}{4} = \frac{9+4+16+9}{4} = \frac{38}{4} = \boxed{9{,}5}$$

Écart-type : $\sigma_1 = \sqrt{9{,}5} \approx \boxed{3{,}08}$
Coefficient d'asymétrie :
$$\begin{aligned} S_k &= \frac{(15-12)^3 + (10-12)^3 + (8-12)^3 + (15-12)^3}{4\,\sigma_1^3} \\ &= \frac{27 + (-8) + (-64) + 27}{4} \cdot \frac{1}{\sigma_1^3} = \frac{-18}{4} \cdot \frac{1}{\left(\frac{38}{4}\right)^{3/2}} \approx \frac{-4{,}5}{29{,}3} \approx \boxed{-0{,}15} \end{aligned}$$

Interprétation : $S_k \approx -0{,}15 < 0$, donc la distribution de X_1 est légèrement asymétrique à gauche. La valeur étant proche de zéro, l'asymétrie est faible.

1.3 Calculer la covariance et le coefficient de corrélation

Méthode.
Soient deux variables X et Y observées sur n individus. On définit :

1. **Covariance empirique** :
$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}$$
où $\overline{xy} = \frac{1}{n}\sum x_i y_i$.
2. **Coefficient de corrélation linéaire** :
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$

Remarque. Interprétation de ρ
$\rho \in [-1,\,1]$. Plus $|\rho|$ est proche de 1, plus le lien **linéaire** entre X et Y est fort.

- $\rho > 0$: lien positif (quand X croît, Y tend à croître).
- $\rho < 0$: lien négatif.
- $|\rho| \approx 0$: pas de lien linéaire apparent (mais un lien non linéaire peut exister).

Attention : corrélation $\neq$ causalité.

Exemple. Application — Polycopié, Chap. 4
On dispose de $n = 4$ individus décrits par deux variables X et Y : $(x_1, y_1) = (0,0)$, $(x_2, y_2) = (1,0)$, $(x_3, y_3) = (0,1)$, $(x_4, y_4) = (2,2)$. Calculer σ_{xy} et ρ_{xy}.

Solution.
Moyennes : $\bar{x} = \frac{0+1+0+2}{4} = \frac{3}{4}$, $\quad \bar{y} = \frac{0+0+1+2}{4} = \frac{3}{4}$
Variances : (le calcul est symétrique par rapport à X et Y)

$$\sigma_x^2 = \frac{0^2+1^2+0^2+2^2}{4} - \left(\frac{3}{4}\right)^2 = \frac{5}{4} - \frac{9}{16} = \frac{11}{16}, \qquad \sigma_y^2 = \frac{11}{16}$$

Covariance :

$$\overline{xy} = \frac{0\cdot 0 + 1\cdot 0 + 0\cdot 1 + 2\cdot 2}{4} = 1, \qquad \sigma_{xy} = 1 - \frac{3}{4}\cdot\frac{3}{4} = 1 - \frac{9}{16} = \boxed{\frac{7}{16}}$$

Coefficient de corrélation :

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} = \frac{7/16}{11/16} = \boxed{\frac{7}{11} \approx 0{,}636}$$

Le lien linéaire entre X et Y est modérément fort et positif.

1.4 Calculer la droite de régression par les moindres carrés

Propriété. Droite de régression des moindres carrés
Soit un échantillon $(x_i, y_i)_{i=1}^n$ de deux variables X et Y avec $\sigma_x \neq 0$. La **droite de régression de Y en X** est la droite $y = ax + b$ qui minimise la somme des erreurs quadratiques $E(a,b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$.
Ses coefficients sont :

$$\boxed{a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad b = \bar{y} - a\,\bar{x}}$$

La droite passe par le centre de gravité $G = (\bar{x}, \bar{y})$.

Méthode.
Pour calculer la droite de régression de Y en X :

1. Calculer $\bar{x}$, $\bar{y}$, $\overline{xy}$, $\overline{x^2}$ (ou σ_{xy} et σ_x^2 directement).
2. Calculer la pente : $a = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$
3. Calculer l'ordonnée à l'origine : $b = \bar{y} - a\,\bar{x}$.
4. Écrire la droite : $y = ax + b$.

Exemple. Application — CE 2026, Ex. 1 (Q3–Q4)
Après transformation des vitesses v_i (en m/s) mesurées aux instants t_i (en s), on pose $n_i = \operatorname{artanh}(v_i/9)$. Les données transformées sont :

t_i	0	1	2	3	4
n_i	0	0,42	0,75	1,02	1,22

Donner la droite de régression de n en fonction de t.

Solution.
On calcule les quatre quantités nécessaires (5 points) :

$$\bar{t} = \frac{0+1+2+3+4}{5} = 2$$

$$\bar{n} = \frac{0+0{,}42+0{,}75+1{,}02+1{,}22}{5} = \frac{3{,}41}{5} = 0{,}682$$

$$\overline{tn} = \frac{0\cdot 0 + 1\cdot 0{,}42 + 2\cdot 0{,}75 + 3\cdot 1{,}02 + 4\cdot 1{,}22}{5} = \frac{0+0{,}42+1{,}50+3{,}06+4{,}88}{5} = 1{,}972$$

$$\overline{t^2} = \frac{0^2+1^2+2^2+3^2+4^2}{5} = \frac{30}{5} = 6$$

Pente :

$$a = \frac{\overline{tn} - \bar{t}\,\bar{n}}{\overline{t^2} - \bar{t}^2} = \frac{1{,}972 - 2\times 0{,}682}{6-4} = \frac{1{,}972 - 1{,}364}{2} = \frac{0{,}608}{2} \approx 0{,}304$$

Ordonnée à l'origine :

$$b = \bar{n} - a\,\bar{t} = 0{,}682 - 0{,}304\times 2 = 0{,}682 - 0{,}608 = 0{,}074$$

La droite de régression est :

$$\boxed{n \approx 0{,}304\,t + 0{,}074}$$

1.5 Linéariser un modèle non linéaire et appliquer la régression

Méthode.
Lorsque le modèle théorique est de la forme $y = f(g(x))$ avec f bijective, on peut se ramener à une régression linéaire en posant $n = f^{-1}(y)$, ce qui donne une relation $n = a\,x + b$:

1. **Identifier** le changement de variable $n = f^{-1}(y)$ (la transformation qui linéarise le modèle).
2. **Calculer** les valeurs transformées $n_i = f^{-1}(y_i)$ pour chaque observation.
3. **Tracer** le nuage de points (x_i, n_i) pour vérifier visuellement l'alignement.
4. **Appliquer** la régression linéaire par les moindres carrés sur (x_i, n_i) (méthode 1.4) pour obtenir $n = ax + b$.
5. **Revenir** au modèle initial : $y = f(ax + b)$.

Remarque. Transformations linéarisantes usuelles

- Modèle $y = A\,e^{bx}$: poser $n = \ln y$ (régression de $\ln y$ en x).
- Modèle $y = A\,x^{b}$: poser $n = \ln y$, $u = \ln x$ (régression de $\ln y$ en $\ln x$).
- Modèle $y = v_\infty \tanh(at+b)$: poser $n = \operatorname{artanh}(y/v_\infty)$ (cf. exemple ci-dessous).

Exemple. Application — CE 2026, Ex. 1 (complet)
Une bille lâchée dans un fluide a une vitesse modélisée par $v(t) = v_\infty \tanh(at+b)$ avec $v_\infty > 0$. On sait que la vitesse limite est $v_\infty = 9\,\text{m/s}$. Les mesures expérimentales sont :

t_i (s)	0	1	2	3	4
v_i (m/s)	0	3,6	5,7	6,9	7,6

Déterminer une expression approchée de $v(t)$.

Solution.
Étape 1 — Identifier le changement de variable.
Le modèle s'écrit $\dfrac{v}{9} = \tanh(at+b)$. On applique artanh des deux membres :

$$\underbrace{\operatorname{artanh}\left(\frac{v}{9}\right)}_{=:\,n} = at + b$$

Le changement de variable $n = \operatorname{artanh}(v/9)$ linéarise le modèle.

Étape 2 — Calculer les valeurs n_i.
On utilise $\operatorname{artanh}(x) = \dfrac{1}{2}\ln\dfrac{1+x}{1-x}$:

t_i	0	1	2	3	4
$n_i = \operatorname{artanh}(v_i/9)$	0	0,42	0,75	1,02	1,22

Étape 3 — Vérification graphique.
Le nuage (t_i, n_i) est proche d'une droite passant par l'origine (voir nuage de points tracé en exam), confirmant la linéarisation.

Étape 4 — Régression linéaire de n en t.
D'après la méthode 1.4 (voir exemple précédent) :

$$n \approx 0{,}304\,t + 0{,}074$$

Étape 5 — Revenir au modèle.
On a $n = \operatorname{artanh}(v/9) \approx 0{,}304\,t + 0{,}074$, donc :

$$\frac{v(t)}{9} = \tanh(0{,}304\,t + 0{,}074)$$

$$\boxed{v(t) \approx 9\,\tanh(0{,}304\,t + 0{,}074)\ \text{m/s}}$$

Interprétation des paramètres : le 9 correspond à la vitesse limite v_∞ (saturation de $\tanh$ vers 1) ; $a \approx 0{,}304$ mesure la rapidité avec laquelle la vitesse atteint cette limite.

1.6 Construire les matrices M, M_c, M_s, Σ et R

Définition. Matrices de données multivariées
Soit un échantillon de n individus décrits par p variables $X_1, \ldots, X_p$ (moyennes $\bar{x}_j$, écarts-types σ_j).

- M $(n \times p)$: matrice des données **brutes**.
- M_c $(n \times p)$: matrice **centrée**, $(M_c)_{ij} = x_{ij} - \bar{x}_j$.
- M_s $(n \times p)$: matrice **centrée-réduite**, $(M_s)_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$.
- $\Sigma = \dfrac{1}{n} M_c^T M_c$ $(p \times p)$: matrice **variance-covariance** ; $\Sigma_{ij} = \sigma_{ij}$.
- $R = \dfrac{1}{n} M_s^T M_s$ $(p \times p)$: matrice des **corrélations** ; $R_{ii} = 1$, $R_{ij} = \rho_{ij}$.

Méthode.
Pour construire M_c, M_s, Σ et R à partir de M :

1. Calculer les moyennes $\bar{x}_j$ et les écarts-types σ_j de chaque colonne de M.
2. Centrer : soustraire $\bar{x}_j$ à chaque élément de la colonne $j \to M_c$.
3. Réduire : diviser chaque élément de la colonne j par $\sigma_j \to M_s$.
4. Calculer $\Sigma = \dfrac{1}{n} M_c^T M_c$ et $R = \dfrac{1}{n} M_s^T M_s$ par produit matriciel.

Vérification : les diagonales de Σ sont les variances σ_j^2 et les diagonales de R valent toutes 1.

Remarque.
Relation entre R et Σ : on a $R_{ij} = \sigma_{ij}/(\sigma_i \sigma_j)$, donc R se déduit de Σ en normalisant. Si toutes les variables sont déjà de même échelle, Σ suffit; sinon, on préfère R pour l'ACP.

Exemple. Application — Polycopié, Chap. 4
Soit un échantillon de $n = 4$ individus décrits par $p = 2$ variables X et Y :

$$(x_1, y_1) = (0, 0), \quad (x_2, y_2) = (1, 0), \quad (x_3, y_3) = (0, 1), \quad (x_4, y_4) = (2, 2)$$

Construire M, M_c, M_s, Σ et R.

Solution.
Matrices brute :

$$M = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 2 & 2 \end{pmatrix}$$

Étape 1 — Moyennes et écarts-types.
$\bar{x} = \bar{y} = \dfrac{3}{4}$. $\quad \sigma_x^2 = \sigma_y^2 = \dfrac{0^2 + 1^2 + 0^2 + 2^2}{4} - \left(\frac{3}{4}\right)^2 = \dfrac{5}{4} - \dfrac{9}{16} = \dfrac{11}{16}$. $\quad \sigma_x = \sigma_y = \dfrac{\sqrt{11}}{4}$.

Étape 2 — Matrice centrée M_c.

$$M_c = M - \bar{x}\,\mathbf{1} = \begin{pmatrix} -3/4 & -3/4 \\ 1/4 & -3/4 \\ -3/4 & 1/4 \\ 5/4 & 5/4 \end{pmatrix}$$

Étape 3 — Matrice centrée-réduite M_s. (diviser chaque colonne par $\sigma = \sqrt{11}/4$)

$$M_s = \begin{pmatrix} -3/\sqrt{11} & -3/\sqrt{11} \\ 1/\sqrt{11} & -3/\sqrt{11} \\ -3/\sqrt{11} & 1/\sqrt{11} \\ 5/\sqrt{11} & 5/\sqrt{11} \end{pmatrix}$$

Étape 4 — Matrices Σ et R.
Par produit matriciel $\Sigma = \frac{1}{4} M_c^T M_c$:

$$M_c^T M_c = \begin{pmatrix} (-3/4)^2 + (1/4)^2 + (-3/4)^2 + (5/4)^2 & (-3/4)^2 + (1/4)(-3/4) + (-3/4)(1/4) + (5/4)^2 \\ \cdots & (-3/4)^2 + (-3/4)^2 + (1/4)^2 + (5/4)^2 \end{pmatrix} = \frac{1}{16} \begin{pmatrix} 44 & 28 \\ 28 & 44 \end{pmatrix}$$

$$\boxed{\Sigma = \frac{1}{4} \cdot \frac{1}{16} \begin{pmatrix} 44 & 28 \\ 28 & 44 \end{pmatrix} = \begin{pmatrix} 11/16 & 7/16 \\ 7/16 & 11/16 \end{pmatrix}}$$

De même, $R = \frac{1}{4} M_s^T M_s$:

$$M_s^T M_s = \frac{1}{11} \begin{pmatrix} 9+1+9+25 & 9-3-3+25 \\ 9-3-3+25 & 9+9+1+25 \end{pmatrix} = \frac{1}{11} \begin{pmatrix} 44 & 28 \\ 28 & 44 \end{pmatrix}$$

$$\boxed{R = \begin{pmatrix} 1 & 7/11 \\ 7/11 & 1 \end{pmatrix}}$$

Vérification : $\Sigma_{11} = 11/16 = \sigma_x^2$ ✓ ; $R_{11} = 1$ ✓ ; $R_{12} = 7/11 = \rho_{xy}$ (cohérent avec la méthode 1.3) ✓.

2 Estimation par intervalles

Ce chapitre distingue deux situations : le paramètre est **connu** (intervalle de pari, on encadre les réalisations futures) ou **inconnu** (intervalle de confiance, on encadre le paramètre à partir de l'échantillon).

Remarque. Notation adoptée dans ce cours
On note z_α la valeur de la loi normale centrée réduite telle que $P(-z_\alpha < Z < z_\alpha) = 1-\alpha$, c'est-à-dire $P(Z > z_\alpha) = \alpha/2$. Valeurs à connaître :

- Risque $\alpha = 10\%$: $z_\alpha = 1{,}645$
- Risque $\alpha = 5\%$: $z_\alpha = 1{,}96$
- Risque $\alpha = 1\%$: $z_\alpha = 2{,}576$

De même, $t_{n-1,\alpha}$ désigne le quantile bilatéral de la loi de Student à $n-1$ degrés de liberté au risque α.

2.1 Construire un intervalle de pari pour la proportion

Méthode.
La proportion p est **connue**. On cherche l'intervalle dans lequel se situera la fréquence empirique F observée sur un échantillon de taille n.

1. **Vérifier les conditions** : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ (approximation normale).
2. **Identifier le risque** α et lire z_α dans la table de la loi normale.
3. **Calculer l'intervalle de pari** :
$$\boxed{I = \left[p \pm z_\alpha \sqrt{\frac{p(1-p)}{n}} \right]}$$
4. **Conclure** : la fréquence empirique F se situera dans I avec une probabilité $1-\alpha$.

Exemple. Application — CE 2026, Ex. 3
On sait que 35% des clients d'une chaîne de cafés utilisent une application mobile. On interroge un échantillon aléatoire de 200 clients. Déterminer un intervalle de pari à 99% pour la proportion P de clients utilisant l'application dans cet échantillon.

Solution.
Conditions : $n = 200 \geq 30$, $np = 200 \times 0{,}35 = 70 \geq 5$, $n(1-p) = 130 \geq 5$. Conditions vérifiées.

Paramètres : $p = 0{,}35$, $\alpha = 1\%$, donc $z_\alpha = 2{,}576$.

Intervalle :
$$I = \left[0{,}35 \pm 2{,}576 \sqrt{\frac{0{,}35 \times 0{,}65}{200}} \right] = [0{,}35 \pm 2{,}576 \times 0{,}0337] = [0{,}35 \pm 0{,}087]$$

$$\boxed{P \in \left[\frac{53}{200} \,;\, \frac{87}{200} \right] \approx [0{,}26\,;\, 0{,}44]}$$

Conclusion : avec une probabilité de 99%, la proportion observée dans l'échantillon sera comprise entre 26% et 44%.

2.2 Construire un intervalle de pari pour la moyenne empirique

Méthode.
La loi de la population est connue : $\mathcal{N}(\mu, \sigma^2)$ avec μ et σ **connus**. On cherche l'intervalle dans lequel tombera la moyenne empirique $\overline{X}$ d'un échantillon de taille n.

1. **Identifier la loi de** $\overline{X}$: d'après le TCL, $\overline{X} \sim \mathcal{N}\left(\mu,\, \dfrac{\sigma^2}{n}\right)$ pour $n \geq 30$ ou si la population est normale.
2. **Calculer l'intervalle de pari** :
$$\boxed{I = \left[\mu \pm z_\alpha \frac{\sigma}{\sqrt{n}}\right]}$$
3. **Conclure** : $\overline{X}$ se situera dans I avec une probabilité $1 - \alpha$.

Remarque.
Plus n est grand, plus $\sigma/\sqrt{n}$ est petit, et plus l'intervalle est étroit : la moyenne empirique fluctue moins autour de μ quand l'échantillon est grand.

Exemple. Application — TD, Ex. 8
Dans une région, le nombre de touristes par journée suit une loi $\mathcal{N}(50\,000\,,\, 8\,000^2)$. On considère un échantillon de $n = 10$ journées. Quelle loi suit la moyenne journalière $\overline{X}$ sur ces 10 journées ? Donner un intervalle de pari à 95% pour $\overline{X}$.

Solution.
Loi de $\overline{X}$ **:**
$$\overline{X} \sim \mathcal{N}\left(\mu,\, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(50\,000,\, \frac{8\,000^2}{10}\right) = \mathcal{N}(50\,000\,,\, 6\,400\,000)$$
L'écart-type de $\overline{X}$ est $\sigma_{\overline{X}} = 8\,000/\sqrt{10} \approx 2\,530$.

Intervalle de pari à 95% ($\alpha = 5\%$, $z_\alpha = 1{,}96$) :
$$I = \left[50\,000 \pm 1{,}96 \times \frac{8\,000}{\sqrt{10}}\right] = [50\,000 \pm 1{,}96 \times 2\,530] = [50\,000 \pm 4\,959]$$
$$\boxed{\overline{X} \in [45\,041\,;\, 54\,959] \text{ touristes/jour}}$$
Conclusion : dans 95% des échantillons de 10 journées, la moyenne journalière observée sera comprise entre 45 041 et 54 959 touristes.

2.3 Calculer un intervalle de confiance pour la proportion

Méthode.
La proportion p est **inconnue**. On l'estime à partir d'une fréquence empirique p_0 observée sur un échantillon de taille n.

1. **Calculer** p_0 (proportion observée dans l'échantillon).
2. **Vérifier les conditions** : $n \geq 30$, $np_0 \geq 5$ et $n(1 - p_0) \geq 5$.
3. **Identifier le risque** α et lire z_α.
4. **Calculer l'intervalle de confiance** :
$$\boxed{I_C = \left[p_0 \pm z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right]}$$
5. **Conclure** : la proportion p de la population sera dans I_C avec une probabilité $1 - \alpha$.

Remarque. Différence avec l'intervalle de pari
L'**intervalle de pari** encadre la statistique (observation future) à partir d'un paramètre connu. L'**intervalle de confiance** encadre le paramètre inconnu à partir de l'observation. Les formules sont identiques, mais le sens est inversé : dans le pari, p est connu ; dans la confiance, p_0 l'estime.

Exemple. Application — TD, Ex. 9
Un échantillon de 50 entreprises d'un secteur donne les chiffres d'affaires (CA) suivants :

CA (en M€)	$[0;2[$	$[2;3[$	$[3;4[$	$[4;5[$	$[5;7[$
Effectif	6	12	17	10	5

Donner une estimation par intervalle de confiance au risque 1% de la proportion d'entreprises dont le CA dépasse $4{,}5$ M€.

Solution.
Calcul de p_0 : les entreprises avec CA $> 4{,}5$ M€ se trouvent dans les classes $[4;5[$ et $[5;7[$, soit $10+5=15$ entreprises.

$$p_0 = \frac{15}{50} = 0{,}3$$

Conditions : $n = 50 \geq 30$, $np_0 = 15 \geq 5$, $n(1-p_0) = 35 \geq 5$. Conditions vérifiées.

Paramètres : $\alpha = 1\%$, $z_\alpha = 2{,}576$.

Intervalle de confiance :

$$I_C = \left[0{,}3 \pm 2{,}576\sqrt{\frac{0{,}3\times 0{,}7}{50}}\right] = [0{,}3 \pm 2{,}576 \times 0{,}0648] = [0{,}3 \pm 0{,}167]$$

$$\boxed{I_C \approx [13{,}3\,\%\,;\ 46{,}7\,\%]}$$

Conclusion : on estime, avec un niveau de confiance de 99%, que la proportion d'entreprises du secteur dont le CA dépasse $4{,}5$ M€ est comprise entre $13{,}3\%$ et $46{,}7\%$.

2.4 Calculer un intervalle de confiance pour la moyenne

Définition. Écart-type corrigé σ_e
Lorsque la variance σ^2 de la population est inconnue, on l'estime à partir de l'écart-type empirique σ_0 observé sur l'échantillon de taille n. L'**écart-type corrigé** est :

$$\sigma_e = \sqrt{\frac{n}{n-1}}\,\sigma_0$$

Le facteur $\sqrt{n/(n-1)}$ corrige le biais de l'estimateur empirique.

Méthode.
La moyenne μ est inconnue. On l'estime par $\bar{x}$ (moyenne observée). On distingue deux cas selon que σ est connu ou non.

Cas 1 — σ connu (loi normale utilisée) :

$$\boxed{I_C = \left[\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}\right]}$$

Condition : $n \geq 30$ ou population normale.

Cas 2 — σ inconnu (loi de Student utilisée) :

1. Calculer $\sigma_e = \sqrt{\dfrac{n}{n-1}}\,\sigma_0$.
2. Lire $t_{n-1,\alpha}$ dans la table de Student à $n-1$ degrés de liberté, au risque bilatéral α.

3. Calculer :

$$\boxed{I_C = \left[\bar{x} \pm t_{n-1,\alpha}\,\frac{\sigma_e}{\sqrt{n}}\right]}$$

Condition : $n \geq 30$ ou population normale.

L'intervalle avec σ inconnu est **plus large** qu'avec σ connu (incertitude supplémentaire sur σ), mais tend vers le même résultat quand $n \to \infty$.

Remarque. Quand utiliser quelle loi ?

- σ **connu** $\Rightarrow$ loi $\mathcal{N}(0,1)$, on utilise z_α.
- σ **inconnu** $\Rightarrow$ loi $\mathcal{T}(n-1)$ (Student), on utilise $t_{n-1,\alpha}$. Pour n grand ($n > 100$), $t_{n-1,\alpha} \approx z_\alpha$: les deux intervalles coïncident pratiquement.

Exemple. Application — TD, Ex. 10
Dans l'atmosphère, le taux de gaz nocif (en ppm) est supposé de loi normale de moyenne μ et de variance σ^2 inconnues. Sur un échantillon de $n = 10$ prélèvements, on observe :

$$\bar{x} = 50 \text{ ppm}, \qquad \sigma_0^2 = 100 \text{ ppm}^2 \quad (\sigma_0 = 10)$$

1. Donner l'IC à 5% de μ (σ^2 inconnu).
2. Quel serait cet IC si $\sigma^2 = 100$ était connu ?

Solution.
Cas 1 — σ inconnu (Student, $n - 1 = 9$ ddl) :
On calcule l'écart-type corrigé :

$$\sigma_e = \sqrt{\frac{n}{n-1}}\,\sigma_0 = \sqrt{\frac{10}{9}} \times 10 \approx 1{,}054 \times 10 = 10{,}54 \text{ ppm}$$

Pour $\alpha = 5\%$ et 9 degrés de liberté : $t_{9;\,5\%} = 2{,}262$ (table de Student bilatérale).

$$I_C = \left[50 \pm 2{,}262 \times \frac{10{,}54}{\sqrt{10}}\right] = [50 \pm 2{,}262 \times 3{,}333] = [50 \pm 7{,}54]$$

$$\boxed{I_C \approx [42{,}46\,;\,57{,}54] \text{ ppm}}$$

Cas 2 — $\sigma = 10$ connu (loi normale) :
Pour $\alpha = 5\%$: $z_\alpha = 1{,}96$.

$$I_C = \left[50 \pm 1{,}96 \times \frac{10}{\sqrt{10}}\right] = [50 \pm 1{,}96 \times 3{,}162] = [50 \pm 6{,}20]$$

$$\boxed{I_C \approx [43{,}80\,;\,56{,}20] \text{ ppm}}$$

Comparaison : l'IC avec σ inconnu ($\pm 7{,}54$) est plus large que celui avec σ connu ($\pm 6{,}20$). L'incertitude sur σ élargit l'intervalle ; ici l'écart est notable car $n = 10$ est petit.

Récap.

Type	Formule	Loi	Conditions
IP proportion	$\left[p \pm z_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$	$\mathcal{N}$	$n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$
IP moyenne	$\left[\mu \pm z_\alpha \frac{\sigma}{\sqrt{n}}\right]$	$\mathcal{N}$	$n \geq 30$ ou pop. normale
IC proportion	$\left[p_0 \pm z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right]$	$\mathcal{N}$	$n \geq 30$, $np_0 \geq 5$, $n(1-p_0) \geq 5$
IC moyenne (σ connu)	$\left[\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}\right]$	$\mathcal{N}$	$n \geq 30$ ou pop. normale
IC moyenne (σ inconnu)	$\left[\bar{x} \pm t_{n-1,\alpha} \frac{\sigma_e}{\sqrt{n}}\right]$	$\mathcal{T}(n-1)$	$n \geq 30$ ou pop. normale

avec $\sigma_e = \sqrt{\frac{n}{n-1}}\,\sigma_0$.

3 Tests statistiques

Un test statistique est une procédure de décision fondée sur un échantillon : à partir d'une statistique calculée sur les données (le **critère**), on décide de rejeter ou non une hypothèse sur la population.

3.1 Formuler H_0 et H_1 ; choisir le type de test

Définition. Hypothèses et type de test

- **Hypothèse nulle** H_0 : hypothèse de départ, en général l'égalité entre le paramètre observé et le paramètre de référence. On ne la rejette que si l'échantillon apporte une preuve suffisante contre elle.
- **Hypothèse alternative** H_1 : ce que l'on cherche à mettre en évidence (différence, supériorité, infériorité).
- **Risque** α : probabilité de rejeter H_0 à tort (erreur de première espèce). On le fixe avant le test (souvent 5% ou 1%).

Méthode.
Pour formuler les hypothèses et choisir le type de test :

1. **Identifier le paramètre** testé (μ, π, ...) et la valeur de référence (μ_0, p, ...).
2. **Écrire** H_0 sous la forme d'une égalité : $H_0 : \{m = \mu_0\}$ ou $H_0 : \{\pi = p\}$.
3. **Choisir** H_1 selon ce que l'on cherche à montrer :
   - $H_1 : \{m \neq \mu_0\} \rightarrow$ **test bilatéral** (on ne sait pas dans quel sens) ;
   - $H_1 : \{m > \mu_0\}$ ou $H_1 : \{m < \mu_0\} \rightarrow$ **test unilatéral** (direction connue a priori).
4. **Fixer le risque** α.

Remarque. Bilatéral vs unilatéral et seuils
Pour un test bilatéral au risque α, on rejette H_0 si le critère dépasse le seuil z_α (bilatéral) en valeur absolue. Pour un test unilatéral au risque α, le seuil est $z_{2\alpha}$ (valeur unilatérale, soit la colonne « unilatéral α » de la table), ce qui correspond à $z_{\alpha/2}$ lu dans la colonne bilatérale :

Risque α	Seuil bilatéral z_α	Seuil unilatéral $z_{2\alpha}$
10%	$1{,}645$	$1{,}282$
5%	$1{,}96$	$1{,}645$
1%	$2{,}576$	$2{,}326$

Pour Student, même logique : lire $t_{n-1;\alpha}$ (bilatéral) ou $t_{n-1;2\alpha}$ (unilatéral) dans la table.

Exemple. Application — DE 2026, Ex. 1 (Q1)
Un laboratoire annonce que la résistance moyenne de ses composants est $\mu_0 = 100\ \Omega$. Sur un échantillon de $n = 40$ composants, on mesure $\bar{x} = 99{,}1\ \Omega$. On souhaite tester si la résistance moyenne de la population est **inférieure** à $100\ \Omega$, au risque $\alpha = 5\%$.
Écrire H_0 et H_1.

Solution.
Soit m la résistance moyenne de la population.

- $H_0 : \{m = 100\}$ (la résistance annoncée est conforme)
- $H_1 : \{m < 100\}$ (la résistance est inférieure)

Type de test : unilatéral (gauche), car H_1 précise une direction ($m < 100$). Le seuil à utiliser sera donc $z_{2\alpha}$ (valeur unilatérale à 5%) $= 1{,}645$.

3.2 Effectuer un test de conformité de la proportion

Méthode.
On teste si la proportion observée p_0 dans un échantillon de taille n est compatible avec la proportion théorique p d'une population de référence.

1. **Poser les hypothèses** : $H_0 : \{\pi = p\}$.
2. **Vérifier les conditions** : $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$.
3. **Calculer le critère** :
$$z = \frac{p_0 - p}{\sqrt{\dfrac{p(1-p)}{n}}}$$
4. **Lire le seuil** dans la table de la loi normale.
5. **Conclure** :
   - Bilatéral ($H_1 : \pi \neq p$) : rejeter H_0 si $|z| > z_\alpha$.
   - Unilatéral droit ($H_1 : \pi > p$) : rejeter H_0 si $z > z_{2\alpha}$.
   - Unilatéral gauche ($H_1 : \pi < p$) : rejeter H_0 si $z < -z_{2\alpha}$.

Exemple. Application — Polycopié, Chap. 3
Dans notre classe de P2, 40 étudiants ont passé leur année. 80% d'entre eux ont été admis en I1. Dans toute la promo P2 de l'Efrei, ce taux est de 70%. Peut-on dire que le taux de passage de notre classe est **significativement supérieur** à celui de la promo, au risque 5% ?

Solution.
Hypothèses : $H_0 : \{\pi = 0{,}7\}$, $H_1 : \{\pi > 0{,}7\}$ (test unilatéral droit).
Conditions : $n = 40 \geq 30$, $np = 28 \geq 5$, $n(1-p) = 12 \geq 5$. Vérifiées.
Critère :
$$z = \frac{0{,}8 - 0{,}7}{\sqrt{\dfrac{0{,}7 \times 0{,}3}{40}}} = \frac{0{,}1}{0{,}0725} \approx 1{,}38$$

Seuil unilatéral à 5% : $z_{2\alpha} = 1{,}645$.
Conclusion : $1{,}38 < 1{,}645$, le critère ne dépasse pas le seuil. On ne rejette pas H_0 au risque 5%. On ne peut pas conclure que le taux de passage de la classe est significativement supérieur à celui de la promo.

3.3 Effectuer un test de conformité de la moyenne

Méthode.
On teste si la moyenne observée $\bar{x}$ est compatible avec la moyenne théorique μ_0 d'une population de référence.

1. **Poser les hypothèses** : $H_0 : \{m = \mu_0\}$.
2. **Choisir la statistique** selon que σ est connu ou non (condition : $n \geq 30$ ou population normale) :
   - σ **connu** $\Rightarrow$ critère $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, seuil dans la table $\mathcal{N}(0,1)$.
   - σ **inconnu** $\Rightarrow$ calculer $\sigma_e = \sqrt{n/(n-1)}\,\sigma_0$, puis critère $t = \dfrac{\bar{x} - \mu_0}{\sigma_e/\sqrt{n}}$, seuil dans la table de Student à $n-1$ ddl.
3. **Lire le seuil** et **conclure** comme en 3.2 (bilatéral ou unilatéral selon H_1).

Exemple. Application — DE 2026, Ex. 1 (Cas A et B)
Reprendre le contexte de la méthode 3.1 : $\mu_0 = 100\ \Omega$, $n = 40$, $\bar{x} = 99{,}1\ \Omega$, $H_1 : \{m < 100\}$, $\alpha = 5\%$.
Cas A : $\sigma = 3\ \Omega$ connu.
Cas B : σ inconnu ; l'écart-type observé sur l'échantillon est $\sigma_o = 3\ \Omega$.

Solution.
Cas A — $\sigma = 3$ connu (loi normale) :

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{99{,}1 - 100}{3/\sqrt{40}} = \frac{-0{,}9}{0{,}474} \approx -1{,}9$$

Seuil unilatéral gauche à 5% : $z_{2\alpha} = 1{,}645$, soit un seuil négatif $-1{,}645$ (car $H_1 : m < 100$).
Comme $-1{,}9 < -1{,}645$, le critère dépasse le seuil. On rejette H_0 au risque 5%.

Cas B — σ inconnu (loi de Student, $n-1 = 39$ ddl) :

$$\sigma_e = \sqrt{\frac{40}{39}} \times 3 \approx 3{,}04\ \Omega$$

$$t = \frac{99{,}1 - 100}{3{,}04/\sqrt{40}} = \frac{-0{,}9}{0{,}481} \approx -1{,}87$$

Seuil unilatéral gauche à 5% avec 39 ddl : $t^{\text{uni}}_{39;\,5\%} = 1{,}684$, soit seuil $= -1{,}684$.
Comme $-1{,}87 < -1{,}684$, le critère dépasse le seuil. On rejette H_0 au risque 5%.

Comparaison : le seuil Student $(1{,}684)$ est légèrement plus grand que le seuil normal $(1{,}645)$: estimer σ à partir des données introduit de l'incertitude et durcit le critère.

3.4 Interpréter la p-valeur et calculer la taille critique

Définition. P-valeur et taille critique

- La **p-valeur** (ou p-value) est le plus petit risque α pour lequel on rejette H_0 avec les données observées. Elle s'obtient en localisant le critère calculé parmi les seuils de la table.
- La **taille critique** n_c est le plus petit n pour lequel on rejette H_0, à α et $\bar{x}$ fixés. Elle s'obtient en résolvant l'équation « critère = seuil » en n.

Méthode.
P-valeur :

1. Calculer le critère (valeur absolue pour un test unilatéral).
2. Localiser cette valeur dans la ligne correspondante de la table (Student ou Normale) : la p-valeur est l'intervalle de risque encadrant le critère.
3. Si p-valeur $< \alpha$: on rejette H_0 ; sinon, on ne rejette pas.

Taille critique n_c (test de la moyenne, σ connu ou estimé fixé) :

1. Écrire l'équation « critère = seuil » :
$$\frac{|\bar{x} - \mu_0|}{\sigma_e/\sqrt{n}} = t_{\text{seuil}}$$
2. Isoler n :
$$\sqrt{n} = \frac{t_{\text{seuil}} \times \sigma_e}{|\bar{x} - \mu_0|} \implies n_c = \left\lceil \left(\frac{t_{\text{seuil}}\, \sigma_e}{|\bar{x} - \mu_0|} \right)^2 \right\rceil$$
3. **Prendre le premier entier** $\geq n_c$ car n est discret.

Remarque.
La p-valeur ne se lit qu'en lecture d'intervalle sur la table (pas de valeur exacte sans logiciel). Elle correspond à la probabilité, sous H_0, d'observer un résultat *au moins aussi extrême* que celui observé.

Exemple. Application — DE 2026, Ex. 1 Cas B (Q c et d)
Suite du Cas B : $n = 40$, $t \approx -1{,}87$, seuil $-1{,}684$, $\sigma_e = 3{,}04\ \Omega$, $\bar{x} = 99{,}1\ \Omega$, $\mu_0 = 100$.
(c) Déterminer la p-valeur de ce test.
(d) Pour $\alpha = 5\%$, à partir de quelle taille n_c d'échantillon la conclusion bascule-t-elle en rejet ?

Solution.
(c) P-valeur :
On compare $|t| = 1{,}87$ aux seuils unilatéraux de la table de Student à 39 ddl (assimilé à 40 ddl) :

$$t^{\text{uni}}_{39;\,5\%} = 1{,}684 \quad < \quad 1{,}87 \quad < \quad t^{\text{uni}}_{39;\,2{,}5\%} = 2{,}021$$

Le critère se situe entre les colonnes 5% et $2{,}5\%$ (unilatéral), donc :

$$\boxed{\text{p-valeur} \in [\,2{,}5\,\%\,;\,5\,\%\,]}$$

Interprétation : si on avait pris $\alpha = 3\%$ par exemple, on aurait aussi rejeté H_0. Si on avait pris $\alpha = 2\%$, on ne l'aurait pas rejeté.

(d) Taille critique n_c :
On résout « critère = seuil » en n (avec $\sigma_e \approx 3{,}04$ et seuil unilatéral $5\% = 1{,}684$) :

$$\frac{0{,}9}{3{,}04/\sqrt{n}} = 1{,}684 \implies \sqrt{n} = \frac{1{,}684 \times 3{,}04}{0{,}9} \approx \frac{5{,}119}{0{,}9} \approx 5{,}69$$

$$n_c = \lceil 5{,}69^2 \rceil = \lceil 32{,}4 \rceil = \boxed{33}$$

Interprétation : avec seulement 33 composants dans l'échantillon (au lieu de 40), on aurait déjà rejeté H_0 à 5%. Avec 32 composants, on n'aurait pas rejeté.

3.5 Effectuer un test du χ^2 d'indépendance

Méthode.
On teste si deux variables qualitatives (à k_1 et k_2 modalités) sont **indépendantes** dans la population, à partir d'un tableau de contingence $(k_1 \times k_2)$ d'effectifs observés o_{ij}.

1. **Poser les hypothèses** : H_0 : les deux variables sont indépendantes ; H_1 : elles sont dépendantes.
2. **Calculer les effectifs marginaux** : $o_{i\bullet} = \sum_j o_{ij}$ (totaux lignes) et $o_{\bullet j} = \sum_i o_{ij}$ (totaux colonnes).
3. **Calculer les effectifs théoriques** sous H_0 (hypothèse d'indépendance) :
$$n_{ij} = \frac{o_{i\bullet} \times o_{\bullet j}}{n}$$
4. **Vérifier les conditions** : $n \geq 50$ et $n_{ij} \geq 5$ pour toutes les cases.
5. **Calculer le critère** du test :
$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - n_{ij})^2}{n_{ij}}$$
6. **Calculer le nombre de degrés de liberté** :
$$\text{ddl} = (k_1 - 1)(k_2 - 1)$$
7. **Lire le seuil** $\chi^2_{\alpha,\text{ddl}}$ dans la table du χ^2.
8. **Conclure** : si $\chi^2 > \chi^2_{\alpha,\text{ddl}}$, on rejette H_0 (les variables sont dépendantes).

Remarque.
Le test du χ^2 est toujours **unilatéral droit** : une grande valeur du critère signifie un grand écart entre effectifs observés et théoriques, donc une dépendance entre les variables. On ne rejette H_0 que si χ^2 est trop grand.

Exemple. Application — DE 2026, Ex. 3
Une école d'ingénieurs souhaite savoir si la méthode de révision (Présentiel / Distanciel / Hybride) influence la réussite à l'examen (Réussi / Échoué). Un échantillon de $n = 200$ étudiants donne le tableau de contingence suivant :

Formation\Résultat	Réussi	Échoué	Total
Présentiel	48	32	80
Distanciel	21	39	60
Hybride	41	19	60
Total	110	90	200

Tester l'indépendance au risque 5%.

Solution.
1. **Hypothèses :**
   - H_0 : le type de formation et la réussite sont **indépendants** (la méthode de révision n'a pas d'influence).
   - H_1 : le type de formation et la réussite sont **dépendants**.
2. **Effectifs théoriques** $(n_{ij} = o_{i\bullet} \times o_{\bullet j} \,/\, 200)$:

	Réussi	Échoué
Présentiel	$\frac{80\times110}{200} = 44$	$\frac{80\times90}{200} = 36$
Distanciel	$\frac{60\times110}{200} = 33$	$\frac{60\times90}{200} = 27$
Hybride	$\frac{60\times110}{200} = 33$	$\frac{60\times90}{200} = 27$

3. **Condition de validité :** $n = 200 \geq 50$ et $n_{ij} \geq 5$ pour toutes les cases. Vérifiée.
4. **Degrés de liberté :**
$$\text{ddl} = (3 - 1)(2 - 1) = 2$$

5. Critère :

$$\begin{aligned}\chi^2 &= \frac{(48-44)^2}{44} + \frac{(32-36)^2}{36} + \frac{(21-33)^2}{33} + \frac{(39-27)^2}{27} + \frac{(41-33)^2}{33} + \frac{(19-27)^2}{27} \\ &= \frac{16}{44} + \frac{16}{36} + \frac{144}{33} + \frac{144}{27} + \frac{64}{33} + \frac{64}{27} \\ &\approx 0{,}36 + 0{,}44 + 4{,}36 + 5{,}33 + 1{,}94 + 2{,}37 \approx \mathbf{14{,}80}\end{aligned}$$

6. Seuil : table du χ^2 à 2 ddl, risque 5% : $\chi^2_{5\%;\,2} = 5{,}99$.
7. Conclusion : $14{,}80 > 5{,}99$, le critère dépasse le seuil. On rejette H_0 au risque 5%.
La méthode de révision choisie a un impact significatif sur la réussite à l'examen d'Analyse de Données.

Récap.

Test	Loi	Critère	Seuil	Conditions
Proportion	$\mathcal{N}$	$z = \dfrac{p_0 - p}{\sqrt{p(1-p)/n}}$	z_α (bil.)	$n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$
Moyenne (σ connu)	$\mathcal{N}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	z_α (bil.)	$n \geq 30$ ou pop. normale
Moyenne (σ inconnu)	$\mathcal{T}(n-1)$	$t = \dfrac{\bar{x} - \mu_0}{\sigma_e/\sqrt{n}}$	$t_{n-1,\alpha}$	$n \geq 30$ ou pop. normale
χ^2 indépendance	χ^2(ddl)	$\chi^2 = \sum_{i,j} \dfrac{(o_{ij} - n_{ij})^2}{n_{ij}}$	$\chi^2_{\alpha,\text{ddl}}$	$n \geq 50$, $n_{ij} \geq 5$

ddl $= (k_1 - 1)(k_2 - 1)$ pour le χ^2 d'indépendance. $\sigma_e = \sqrt{n/(n-1)}\,\sigma_0$ pour la moyenne en σ inconnu. Bilatéral $\leftrightarrow$ rejeter si |critère| $>$ seuil ; unilatéral $\leftrightarrow$ remplacer z_α par $z_{2\alpha}$.

4 Analyse en Composantes Principales (ACP)

L'ACP cherche de nouvelles variables (axes propres) qui résument le mieux possible l'information contenue dans p variables corrélées, en les remplaçant par un petit nombre d'axes non corrélés portant la majorité de la variance.

Remarque. Pré-requis : matrices M, M_c, M_s, Σ, R
La construction de ces matrices est détaillée en **section 1.6**. En pratique, l'ACP s'applique sur la matrice des corrélations R (lorsque les variables n'ont pas la même unité) ou sur la matrice variance-covariance Σ (unités comparables). On rappelle :

$$R = \frac{1}{n} M_s^T M_s, \qquad \Sigma = \frac{1}{n} M_c^T M_c$$

Les p variables et n individus de M_s ont chacun moyenne 0 et écart-type 1.

Vidéo explicative - ACP

En complément de cette fiche, une vidéo explicative sur l'ACP que j'ai réalisée est disponible pour accompagner vos révisions et la compréhension de cette partie :

https://youtu.be/dMULTOzn770

4.1 Diagonaliser R : valeurs propres et matrice de passage P

Méthode.
On cherche les p valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ et les vecteurs propres unitaires $u_1, u_2, \ldots, u_p$ de la matrice R (symétrique réelle, donc diagonalisable).

1. **Polynôme caractéristique** : résoudre $\det(R - \lambda I) = 0$ pour obtenir les valeurs propres, ordonnées de la plus grande à la plus petite.
2. **Vecteurs propres** : pour chaque λ_i, résoudre $(R - \lambda_i I)\, u = 0$. Normaliser : $u_i \leftarrow u_i / \|u_i\|$.
3. **Matrice de passage** : concaténer les vecteurs propres en colonnes :

$$P = \left(u_1 \,\middle|\, u_2 \,\middle|\, \cdots \,\middle|\, u_p \right)$$

P est **orthogonale** : $P^T = P^{-1}$, donc $P^T P = I$.

Remarque. Vérification rapide
On peut contrôler que $R = PDP^T$ avec $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, ou que $Ru_i = \lambda_i u_i$ terme à terme. De plus, $\mathrm{Tr}(R) = \sum_{i=1}^{p} \lambda_i = p$ (la trace est invariante par changement de base) : la somme des valeurs propres vaut toujours p.

Exemple. Application — Polycopié, Chap. 4 (exemple filé)
On reprend les données de la **section 1.6** ($n = 4$ individus, $p = 2$ variables X et Y). On a calculé :

$$R = \begin{pmatrix} 1 & \frac{7}{11} \\ \frac{7}{11} & 1 \end{pmatrix}$$

Calculer les valeurs propres λ_1, λ_2 et la matrice de passage P.

Solution.
1. Valeurs propres :

$$\det(R - \lambda I) = (1-\lambda)^2 - \left(\tfrac{7}{11}\right)^2 = \left(1 - \tfrac{7}{11} - \lambda\right)\left(1 + \tfrac{7}{11} - \lambda\right) = 0$$

$$\boxed{\lambda_1 = 1 + \tfrac{7}{11} = \tfrac{18}{11}, \quad \lambda_2 = 1 - \tfrac{7}{11} = \tfrac{4}{11}}$$

Vérification : $\lambda_1 + \lambda_2 = \frac{18+4}{11} = 2 = p$ ✓

2. Vecteurs propres :

Pour $\lambda_1 = 18/11$: $(R - \lambda_1 I)\, u = 0 \Rightarrow \begin{pmatrix} -7/11 & 7/11 \\ 7/11 & -7/11 \end{pmatrix} u = 0 \Rightarrow a = b$

$$\Rightarrow u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Pour $\lambda_2 = 4/11$: $(R - \lambda_2 I)\, u = 0 \Rightarrow \begin{pmatrix} 7/11 & 7/11 \\ 7/11 & 7/11 \end{pmatrix} u = 0 \Rightarrow a = -b$

$$\Rightarrow u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

3. Matrice de passage :

$$\boxed{P = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}}$$

P est orthogonale et représente une rotation d'angle $\pi/4 = 45$.

4.2 Calculer les qualités globales d'explication (*qge*)

Définition. *qge* : qualité globale d'explication
La *qge* du i-ème axe propre mesure la part d'information (variance totale) portée par cet axe :

$$\mathrm{qge}(i) = \frac{\lambda_i}{p}$$

On a $\sum_{i=1}^{p} \mathrm{qge}(i) = 1$. En pratique, on retient les premiers axes dont la somme des qge dépasse 80–90%.

Méthode.

1. Calculer $\mathrm{qge}(i) = \lambda_i / p$ pour chaque axe i.
2. **Choisir le nombre d'axes** à retenir : garder les k premiers axes tels que $\sum_{i=1}^{k} \mathrm{qge}(i) \geq 80\%$ (ou selon le critère métier).
3. **Interpréter** : qge(1) $\to 1$ quand les données sont parfaitement alignées ; qge $= 1/p$ pour chaque axe si les variables sont toutes indépendantes.

Exemple. Application — Polycopié, Chap. 4 (suite)
Calculer les qge des deux axes propres et les interpréter.

Solution.

$$\mathrm{qge}(1) = \frac{\lambda_1}{p} = \frac{18/11}{2} = \frac{9}{11} \approx \boxed{81{,}8\%}$$

$$\mathrm{qge}(2) = \frac{\lambda_2}{p} = \frac{4/11}{2} = \frac{2}{11} \approx \boxed{18{,}2\%}$$

Le premier axe propre porte $81{,}8\%$ de l'information. Un seul axe suffit pour une bonne synthèse (on est au-dessus du seuil 80%). Le second axe n'apporte que $18{,}2\%$ d'information marginale.
Remarque : si $\rho_{xy} \to 1$ (corrélation parfaite), alors $\lambda_1 \to 2$ et $\mathrm{qge}(1) \to 100\%$: tout l'information se concentre sur le premier axe (le nuage est parfaitement aligné).

4.3 Calculer la matrice F des individus et reconstruire M_s

Définition. Matrice F des coordonnées factorielles
La matrice F $(n \times p)$ contient les coordonnées des n individus dans la **nouvelle base propre** (axes $u_1, \ldots, u_p$) :

$$F = M_s\, P$$

La ligne i de F donne les « nouvelles notes » de l'individu i sur chacun des axes propres. **Reconstruction inverse** : puisque P est orthogonale ($P^{-1} = P^T$) :

$$M_s = F\, P^T$$

Méthode.
Calcul direct (de M_s vers F) :

1. Multiplier $F = M_s \times P$ (produit $n \times p$ par $p \times p$).
2. La colonne j de F contient les projections des individus sur l'axe u_j.

Reconstruction (de F vers M_s, si F et P sont donnés) :

1. Calculer $M_s = F \times P^T$ (puisque $P^{-1} = P^T$).

Exemple. Application — DE 2026, Ex. 2 (Q3) : reconstruction de M_s
On dispose de la matrice F des coordonnées de $n = 3$ individus dans la base propre et de la matrice de passage :

$$F = \begin{pmatrix} 2 & 0 \\ -1 & 1 \\ 0 & -2 \end{pmatrix}, \qquad P = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Reconstruire la matrice des données centrées-réduites M_s.

Solution.
Puisque P est orthogonale, $P^{-1} = P^T$:

$$P^T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

$$M_s = F\, P^T = \begin{pmatrix} 2 & 0 \\ -1 & 1 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

Individu 1 : $\left(\frac{2}{\sqrt{2}} + 0,\ \frac{2}{\sqrt{2}} + 0\right) = (\sqrt{2},\ \sqrt{2})$
Individu 2 : $\left(\frac{-1}{\sqrt{2}} - \frac{1}{\sqrt{2}},\ \frac{-1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\right) = (-\sqrt{2},\ 0)$
Individu 3 : $\left(0 + \frac{2}{\sqrt{2}},\ 0 - \frac{2}{\sqrt{2}}\right) = (\sqrt{2},\ -\sqrt{2})$

$$\boxed{M_s = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} \end{pmatrix}}$$

Les données initiales centrées-réduites des 3 individus étaient bien $(\sqrt{2}, \sqrt{2})$, $(-\sqrt{2}, 0)$ et $(\sqrt{2}, -\sqrt{2})$.

Exemple. Application — Polycopié, Chap. 4 (suite) : calcul de F
Calculer la matrice F pour les 4 individus de l'exemple filé, en utilisant M_s (section 1.6) et P (section 4.1).

Solution.
On rappelle :

$$M_s = \begin{pmatrix} -3/\sqrt{11} & -3/\sqrt{11} \\ 1/\sqrt{11} & -3/\sqrt{11} \\ -3/\sqrt{11} & 1/\sqrt{11} \\ 5/\sqrt{11} & 5/\sqrt{11} \end{pmatrix}, \qquad P = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

On calcule $F_{i1} = \frac{m_{i1}+m_{i2}}{\sqrt{2}}$ et $F_{i2} = \frac{-m_{i1}+m_{i2}}{\sqrt{2}}$:

Individu	F_{i1}	F_{i2}
1	$(-3-3)/(\sqrt{11}\sqrt{2}) = -6/\sqrt{22}$	$(3-3)/(\sqrt{11}\sqrt{2}) = 0$
2	$(1-3)/(\sqrt{11}\sqrt{2}) = -2/\sqrt{22}$	$(-1-3)/(\sqrt{11}\sqrt{2}) = -4/\sqrt{22}$
3	$(-3+1)/(\sqrt{11}\sqrt{2}) = -2/\sqrt{22}$	$(3+1)/(\sqrt{11}\sqrt{2}) = 4/\sqrt{22}$
4	$(5+5)/(\sqrt{11}\sqrt{2}) = 10/\sqrt{22}$	$(-5+5)/(\sqrt{11}\sqrt{2}) = 0$

$$\boxed{F = \frac{1}{\sqrt{22}} \begin{pmatrix} -6 & 0 \\ -2 & -4 \\ -2 & 4 \\ 10 & 0 \end{pmatrix}}$$

4.4 Calculer la matrice des saturations S

Définition. Matrice des saturations S
La matrice S ($p \times p$) contient les **coefficients de corrélation** entre les p anciennes variables (axes initiaux) et les p nouveaux axes propres :

$$S = P\, D^{1/2} \qquad \text{avec } D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$$

Ainsi $s_{ij} = \mathrm{corr}(\text{variable } i, \text{ axe propre } j)$. Les propriétés :

$$\sum_{i=1}^{p} s_{ij}^2 = \lambda_j \qquad \text{et} \qquad \sum_{j=1}^{p} s_{ij}^2 = 1$$

indiquent que chaque variable est globalement bien représentée sur l'ensemble des axes (somme = 1).

Méthode.

1. Former $D^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$.
2. Calculer $S = P \times D^{1/2}$.
3. **Interpréter** : $|s_{ij}|$ proche de 1 signifie que la variable i est fortement liée au j-ème axe propre. $|s_{ij}|$ proche de 0 signifie qu'elle est peu corrélée à cet axe.

Exemple. Application — Polycopié, Chap. 4 (suite)
Calculer et interpréter la matrice S pour l'exemple filé ($\lambda_1 = 18/11$, $\lambda_2 = 4/11$, P donné en section 4.1).

Solution.
On forme :

$$D^{1/2} = \begin{pmatrix} \sqrt{18/11} & 0 \\ 0 & \sqrt{4/11} \end{pmatrix} = \begin{pmatrix} 3\sqrt{2}/\sqrt{11} & 0 \\ 0 & 2/\sqrt{11} \end{pmatrix}$$

$$S = P\,D^{1/2} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} 3\sqrt{2}/\sqrt{11} & 0 \\ 0 & 2/\sqrt{11} \end{pmatrix}$$

$$\boxed{S = \begin{pmatrix} 3/\sqrt{11} & -\sqrt{2}/\sqrt{11} \\ 3/\sqrt{11} & \sqrt{2}/\sqrt{11} \end{pmatrix} \approx \begin{pmatrix} 0{,}905 & -0{,}426 \\ 0{,}905 & 0{,}426 \end{pmatrix}}$$

Interprétation :

- Les deux variables X et Y sont fortement corrélées à l'axe 1 ($s \approx 0{,}905$: corrélation positive forte). L'axe 1 représente une sorte de « note moyenne ».
- X est corrélée négativement à l'axe 2 ($-0{,}426$) et Y positivement ($+0{,}426$) : l'axe 2 représente la différence $Y - X$.

Vérification : $s_{11}^2 + s_{21}^2 = 9/11 + 9/11 = 18/11 = \lambda_1$ ✓ et $s_{11}^2 + s_{12}^2 = 9/11 + 2/11 = 1$ ✓

4.5 Calculer les qualités de représentation (*qlt*) des individus

Définition. *qlt* : qualité de représentation d'un individu
La *qlt* de l'individu i sur l'axe j mesure la part de sa « distance au centre » expliquée par cet axe :

$$\mathrm{qlt}(i,j) = \frac{F_{ij}^2}{\sum_{k=1}^{p} F_{ik}^2}$$

La somme sur tous les axes vaut 1 : $\sum_{j=1}^{p} \mathrm{qlt}(i,j) = 1$. Un individu avec $\mathrm{qlt}(i,1) \approx 1$ est très bien représenté par le premier axe ; $\mathrm{qlt}(i,1) \approx 0$ signifie qu'il est mal représenté et que sa position sur cet axe n'est pas fiable.

Méthode.
Soit la matrice F connue (calculée en section 4.3).

1. Pour l'individu i, calculer la norme carrée de sa ligne : $d_i^2 = \sum_{k=1}^{p} F_{ik}^2$.
2. Calculer $\mathrm{qlt}(i,j) = F_{ij}^2 / d_i^2$ pour chaque axe j.
3. **Interpréter** : qlt proche de 1 sur un axe → individu bien représenté sur cet axe (on peut se fier à sa position) ; qlt proche de 0 → position peu fiable sur cet axe.

Exemple. Application — Polycopié, Chap. 4 (suite)
Calculer les qlt des 4 individus de l'exemple filé sur chaque axe propre, en utilisant la matrice F calculée en section 4.3.

Solution.
On rappelle :

$$F = \frac{1}{\sqrt{22}} \begin{pmatrix} -6 & 0 \\ -2 & -4 \\ -2 & 4 \\ 10 & 0 \end{pmatrix}$$

Pour chaque individu i, on calcule $d_i^2 = (F_{i1}^2 + F_{i2}^2)$, puis $\mathrm{qlt}(i,j) = F_{ij}^2 / d_i^2$:

Ind.	F_{i1}	F_{i2}	$\mathrm{qlt}(i,1)$	$\mathrm{qlt}(i,2)$
1	$-6/\sqrt{22}$	0	$36/(36+0) = \mathbf{1}$	0
2	$-2/\sqrt{22}$	$-4/\sqrt{22}$	$4/(4+16) = \mathbf{0{,}2}$	$16/20 = \mathbf{0{,}8}$
3	$-2/\sqrt{22}$	$4/\sqrt{22}$	$4/20 = \mathbf{0{,}2}$	$16/20 = \mathbf{0{,}8}$
4	$10/\sqrt{22}$	0	$100/(100+0) = \mathbf{1}$	0

Interprétation :

- Les individus 1 et 4 (qui se trouvent exactement sur l'axe u_1) sont parfaitement représentés par le premier axe (qlt $= 1$) et ignorés par le second.
- Les individus 2 et 3 sont plutôt représentés par le deuxième axe (qlt $= 0{,}8$) : leur position sur l'axe 1 ne reflète que 20% de leur distance au centre, donc leurs coordonnées sur cet axe sont à interpréter avec prudence.

Récap. Pipeline complet de l'ACP

1. **Construire** M_c et M_s à partir de M (section 1.6).
2. **Calculer** $R = \frac{1}{n} M_s^T M_s$ (section 1.6).
3. **Diagonaliser** R : trouver $\lambda_1 \geq \cdots \geq \lambda_p$ et les vecteurs propres unitaires $u_i \rightarrow$ matrice P (section 4.1).
4. **Calculer** les qge$(i) = \lambda_i / p$ et choisir le nombre d'axes à retenir (section 4.2).
5. **Calculer** $F = M_s P$ (coordonnées des individus dans la nouvelle base) ou **reconstruire** $M_s = F P^T$ (section 4.3).
6. **Calculer** $S = P D^{1/2}$ (saturations = corrélations variables–axes) (section 4.4).
7. **Calculer** les qlt$(i,j) = F_{ij}^2 / \sum_k F_{ik}^2$ (qualité de représentation des individus) (section 4.5).

Rappel des formules clés :

$$F = M_s P, \quad M_s = F P^T, \quad S = P D^{1/2}, \quad \mathrm{qge}(i) = \frac{\lambda_i}{p}, \quad \mathrm{qlt}(i,j) = \frac{F_{ij}^2}{\sum_k F_{ik}^2}$$