

7. Analyse bivariée

Statistique Descriptive

Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

2026-2027

1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

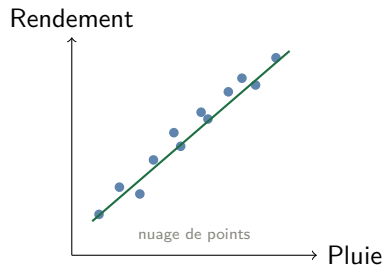
1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

Ce qu'on sait déjà (modules 4-6)

- ▶ Décrire *une* variable : centre, dispersion, forme
- ▶ Plantation A : rendement moyen = 3,8 t/ha, $s = 0,6$
- ▶ Plantation B : pluviométrie moyenne = 1 200 mm, $s = 180$

Ce qu'on ne sait pas encore

- ▶ Est-ce que plus de pluie $\Rightarrow$ meilleur rendement ?
- ▶ Comment quantifier ce lien ?
- ▶ Peut-on prédire le rendement à partir de la pluie ?



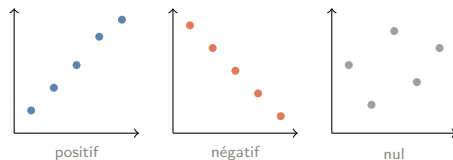
Selon la nature des variables

- ▶ Quantitative vs. Quantitative : Corrélation/Régression
- ▶ Qualitative vs. Qualitative : Tableau de contingence
- ▶ Qualitative vs. Quantitative : Comparaison de groupes

Ce module

On se concentre sur **deux variables quantitatives** : corrélation et régression linéaire simple.

Selon la forme du lien



- ▶ **Positif** : les deux variables croissent ensemble
- ▶ **Négatif** : l'une croît quand l'autre décroît
- ▶ **Nul** : pas de lien linéaire apparent

1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

Nuage de points

Représentation graphique de n paires (x_i, y_i) dans le plan cartésien.

Chaque individu = **un point**.

Ce qu'il faut observer

- **Direction** : tendance montante ou descendante ?
- **Force** : points serrés ou dispersés ?
- **Forme** : linéaire ? courbe ? groupes ?
- **Points aberrants** : isolés, loin des autres ?

Exemple : sol et rendement

Azote (kg/ha)	Rendement (t/ha)
20	1.8
40	2.3
60	3.1
80	3.8
100	4.5
120	4.9

Covariance empirique

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

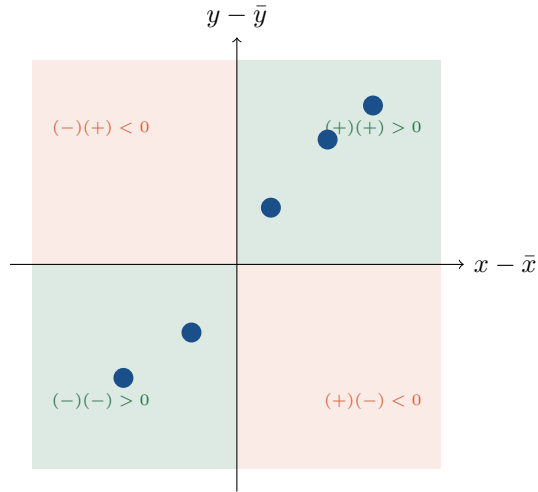
Interprétation du signe

- ▶ $\text{Cov} > 0$: quand X est grand, Y tend à l'être aussi
- ▶ $\text{Cov} < 0$: quand X est grand, Y tend à être petit
- ▶ $\text{Cov} = 0$: pas de lien linéaire

Limite de la covariance

Sa valeur dépend des **unités** de mesure.

On ne peut pas comparer deux covariances brutes.



Quadrants verts : contribution > 0

Quadrants rouges : contribution < 0

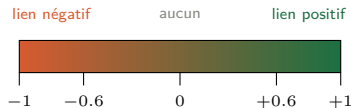
1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

Coefficient de corrélation de Pearson

$$r = \frac{\text{Cov}(X, Y)}{s_X \cdot s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Propriétés fondamentales

- ▶ Toujours : $-1 \leq r \leq 1$
- ▶ Sans unité (adimensionnel)
- ▶ $r = +1$: relation linéaire parfaite croissante



Règle pratique (agronomie)

$ r $	Interprétation
$< 0,3$	lien faible ou nul
$0,3-0,5$	lien modéré
$0,5-0,7$	lien assez fort
$> 0,7$	lien fort
$> 0,9$	lien très fort

Données : 6 parcelles - azote apporté (X) et rendement (Y)

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	20	1,8	-50	-1,7	85,0
2	40	2,3	-30	-1,2	36,0
3	60	3,1	-10	-0,4	4,0
4	80	3,8	+10	+0,3	3,0
5	100	4,5	+30	+1,0	30,0
6	120	4,9	+50	+1,4	70,0
$\bar{x} = 70$		$\bar{y} = 3,4$		$\sum =$	228,0

$$s_X = \sqrt{\frac{7000}{5}} = 37,4 \quad s_Y = \sqrt{\frac{5,3}{5}} = 1,03 \quad \text{Cov} = \frac{228,0}{5} = 45,6$$

$$r = \frac{45,6}{37,4 \times 1,03} \approx 1,18$$

Résultat impossible !



$r > 1$ indique une erreur. Correction : $\sum(x_i - \bar{x})^2 = 7\,000$, $\sum(y_i - \bar{y})^2 \approx 7,06 \Rightarrow r \approx 0,997$ / 25

```
# Données
azote ← c(20, 40, 60, 80, 100, 120)
rendement ← c(1.8, 2.3, 3.1, 3.8, 4.5, 4.9)
# Nuage de points
plot(azote, rendement, xlab = "Azote apporté (kg/ha)",
      ylab = "Rendement (t/ha)", main = "Lien azote-rendement", pch = 16, col = "#1D6F42"
    )
# Coefficient de corrélation
cor(azote, rendement) # r de Pearson
cor(azote, rendement, method = "spearman") # Spearman (rang)
# Matrice de corrélation (plusieurs variables)
df ← data.frame(azote, rendement)
cor(df)
```

Résultat attendu

[1] 0.9972 lien très fort et positif entre azote et rendement

Piège classique

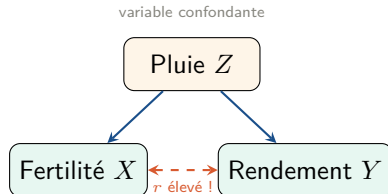
Un fort coefficient r ne prouve **pas** qu'une variable *cause* l'autre.

Trois situations possibles

1. X cause Y (azote $\rightarrow$ rendement)
2. Y cause X (peu probable ici)
3. Z cause les deux (**variable confondante**)

Exemples agronomiques

- Fertilisation et qualité du sol : confondant commun
- Taille des fruits et sucre : lien réel mais complexe



Règle d'or

La corrélation décrit une **association**.
La causalité se prouve par l'expérimentation.

1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

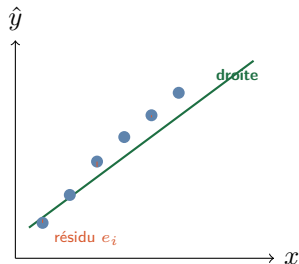
Droite de régression (moindres carrés)

$$\hat{y} = a + b x$$

$$b = \frac{\text{Cov}(X, Y)}{s_X^2} = r \cdot \frac{s_Y}{s_X} \quad a = \bar{y} - b \bar{x}$$

Interprétation des coefficients

- b (pente) : si X augmente de 1 unité, $\hat{Y}$ augmente de b unités
- a (ordonnée) : valeur prédite quand $X = 0$ (peut ne pas avoir de sens physique)



Le résidu $e_i = y_i - \hat{y}_i$ mesure l'écart entre observation et prédiction.

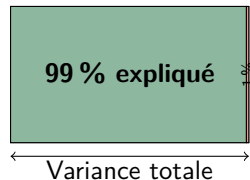
Application numérique

$$b = \frac{45,6}{1\,400} \approx 0,0326 \text{ t/ha par kg N/ha}, \quad a = 3,4 - 0,0326 \times 70 \approx 1,12$$

Coefficient de détermination R^2

$$R^2 = r^2 \in [0, 1]$$

$$R^2 = \frac{\text{Var expliquée}}{\text{Var totale}} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$



Lecture

- ▶ $R^2 = 0,99$: le modèle explique **99 %** de la variabilité du rendement
- ▶ $R^2 = 0,30$: seulement 30 %, mauvais ajustement

Attention aux artefacts

- ▶ R^2 élevé ne garantit pas que le *modèle est juste*
- ▶ Vérifier toujours le **graphe des résidus**

Notre exemple

$r = 0,997 \Rightarrow R^2 = 0,994$
99,4 % de la variation du rendement est expliquée par l'azote.

Régression linéaire en R

```
# Modèle de régression
mod ← lm(rendement ~ azote)
summary(mod)

# Résultats clés :
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.1286    0.0891    12.67  0.000 ***
# azote        0.0326    0.0009    36.22  0.000 ***
# Multiple R-squared:  0.9943

# Graphe avec droite ajustée
plot(azote, rendement, pch = 16, col = "#1D6F42", xlab = "Azote (kg/ha)",
     ylab = "Rendement (t/ha)")
abline(mod, col = "#1A4F8A", lwd = 2)

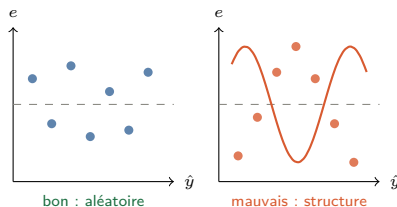
# Diagnostics du modèle
par(mfrow = c(2, 2))
plot(mod)
```

Hypothèses du modèle linéaire

1. **Linéarité** : relation $Y = a + bX + \varepsilon$
2. **Homoscédasticité** : variance des résidus constante
3. **Normalité** des résidus
4. **Indépendance** des observations

Graphes de diagnostic (`plot(mod)`)

- ▶ *Residuals vs Fitted* : linéarité + homoscédasticité
- ▶ *Normal Q-Q* : normalité des résidus
- ▶ *Scale-Location* : homoscédasticité
- ▶ *Cook's distance* : points influents



Si les hypothèses échouent

Transformer Y (log, racine), utiliser une régression polynomiale ou un modèle non linéaire.

1. Pourquoi étudier deux variables ensemble ?
2. Nuage de points et covariance
3. Coefficient de corrélation de Pearson
4. Régression linéaire simple
5. Corrélation de Spearman

Corrélation de Spearman ρ

On remplace les valeurs par leur **rang** puis on calcule le r de Pearson sur les rangs.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (\text{sans ex-aequo})$$

où $d_i = \text{rang}(x_i) - \text{rang}(y_i)$.

Quand utiliser Spearman ?

- ▶ Relation monotone mais non linéaire
- ▶ Variables ordinales (score de vigueur, qualité sanitaire)
- ▶ Présence de valeurs aberrantes influentes
- ▶ Distribution très asymétrique

Exemple : scores de qualité

Parcelle	Score sol	Rendement	d_i
A	5 (r :1)	4,9 (r :1)	0
B	4 (r :2)	4,2 (r :2)	0
C	3 (r :3)	3,5 (r :3)	0
D	2 (r :4)	2,8 (r :4)	0
E	1 (r :5)	1,5 (r :5)	0

$$\rho = 1 - \frac{6 \times 0}{5 \times 24} = 1,00$$

Ordre parfaitement conservé $\Rightarrow$
Spearman = 1.

Pearson vs Spearman en R

```
# Données : taux de germination et température
temp ← c(15, 18, 22, 25, 28, 32, 36)
germina ← c(40, 65, 82, 90, 88, 70, 45)
# Pearson (lien linéaire)
cor(temp, germina, method = "pearson") # faible car courbe en cloche
# Spearman (lien monotone)
cor(temp, germina, method = "spearman") # aussi faible !
# Visualisation
plot(temp, germina, pch = 16, col = "#1A4F8A", xlab = "Temperature (C)",
      ylab = "Germination (%)", main = "Relation non monotone")
lines(smooth.spline(temp, germina), col = "#1D6F42", lwd = 2)
# Conclusion : ici, ni Pearson ni Spearman ne capturent
# le lien en cloche → régression polynomiale
```

Corrélation $\approx 0 \neq$ pas de lien

Un r ou ρ proche de zéro signifie absence de lien *monotone*, pas absence de tout lien.

Étape	Outil	R	Ce qu'on lit
1. Explorer	Nuage de points	<code>plot(x, y)</code>	Direction, force, forme
2. Quantifier	Pearson r	<code>cor(x,y)</code>	-1 à +1, intensité du lien
	Spearman ρ	<code>cor(..., "spearman")</code>	Lien monotone, robuste
3. Modéliser	Régression	<code>lm(y ~ x)</code>	Pente, ordonnée, prédiction
4. Évaluer	R^2	<code>summary(mod)</code>	% de variance expliquée
5. Vérifier	Résidus	<code>plot(mod)</code>	Hypothèses du modèle

Questions à se poser

- ▶ Y a-t-il un lien ? (nuage, r)
- ▶ Dans quel sens ? (signe de r)
- ▶ À quelle force ? ($|r|$)
- ▶ Puis-je prédire ? (R^2 , résidus)

Contexte agronomique

L'analyse bivariée est le point de départ des **modèles de culture** : irrigation $\times$ rendement, fertilisation $\times$ qualité, densité $\times$ compétition.

Exercice : Que disent $\hat{a}$ et $\hat{b}$?

Consigne Pour chacune des droites $\hat{y} = \hat{a} + \hat{b}x$ ci-dessous, **formulez en une phrase** l'interprétation de la pente $\hat{b}$ et de l'ordonnée à l'origine $\hat{a}$ dans le contexte donné. Précisez les **unités** et indiquez si $\hat{a}$ est **agronomiquement interprétable**.

Cas	Équation	Variable x	Variable y
1	$\hat{y} = 1,2 + 0,035x$ $r = 0,96, x \in [0; 160]$	Dose d'azote (kg N/ha)	Rendement maïs (t/ha)
2	$\hat{y} = 12,4 + 0,18x$ $r = 0,88, x \in [200; 800]$	Ration journalière bovin (g MS/j)	Poids vif (kg)
3	$\hat{y} = -410 + 0,52x$ $r = 0,91, x \in [800; 1400]$	Pluviométrie annuelle (mm)	Biomasse herbacée (kg MS/ha)
4	$\hat{y} = -2,1 + 0,31x$ $r = 0,83, x \in [15; 40]$	Température de stockage (°C)	Pertes post-récolte (%)

Merci de votre attention

Statistiques Descriptives · Module 7

Equipe Bioinformatique et Biostatistique

Laboratoire de Microbiologie, Biotechnologie et Bioinformatique

Institut National Polytechnique Félix Houphouët-Boigny

