

5. Mesures de dispersion

Statistique Descriptive

Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

2026-2027

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Deux plantations, même moyenne

Plantation A (rendements en t/ha) :

3,8 4,0 4,1 4,0 3,9 4,2

Plantation B (rendements en t/ha) :

1,5 6,8 2,1 7,2 3,0 5,4

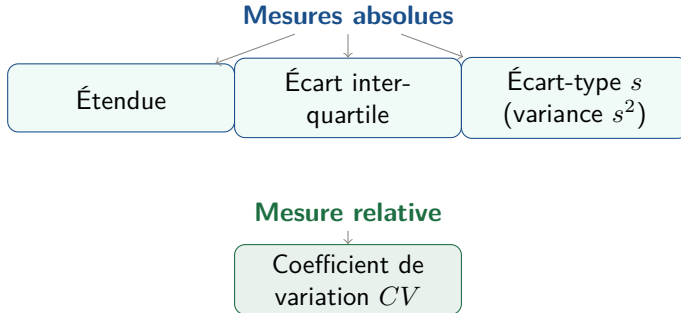
Dans les deux cas : $\bar{x} = 4,0$ t/ha

Ces deux plantations sont-elles identiques ?

- Plantation A : production **régulière**, risque faible
- Plantation B : production **erratique**, risque élevé

Leçon fondamentale

Deux séries avec la même moyenne peuvent être **totalelement différentes**. La dispersion est une information indispensable.



1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Étendue

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

Différence entre la valeur la plus grande et la plus petite.

Plantation A / Plantation B

$$R_A = 4,2 - 3,8 = 0,4 \text{ t/ha}$$

$$R_B = 7,2 - 1,5 = 5,7 \text{ t/ha}$$

Avantages

- ▶ Très simple à calculer
- ▶ Premier aperçu immédiat

Limites majeures

- ▶ N'utilise que **2 valeurs**
- ▶ Très sensible aux **valeurs extrêmes**
- ▶ Un seul arbre exceptionnel change tout le résultat

Quartiles

Les quartiles Q_1 , Q_2 , Q_3 partagent la série **ordonnée** en quatre parties égales :

- ▶ Q_1 : 25 % des valeurs sont en dessous
- ▶ Q_2 = médiane : 50 % des valeurs sont en dessous
- ▶ Q_3 : 75 % des valeurs sont en dessous



Calcul des quartiles : exemple pas à pas

Rendements de 10 parcelles (t/ha), triés

2,6 3,0 3,1 3,5 **3,8** **4,1** 4,4 4,8 5,0 5,5

Étape 1 : $n = 10$ valeurs triées

Étape 2 : Médiane Q_2

Position : entre rang 5 et rang 6

$$Q_2 = (3,8 + 4,1)/2 = \mathbf{3,95 \text{ t/ha}}$$

Étape 3 : Q_1 = médiane des 5 premières

$$Q_1 = \text{rang } 3 = \mathbf{3,1 \text{ t/ha}}$$

Étape 4 : Q_3 = médiane des 5 dernières

$$Q_3 = \text{rang } 8 = \mathbf{4,8 \text{ t/ha}}$$

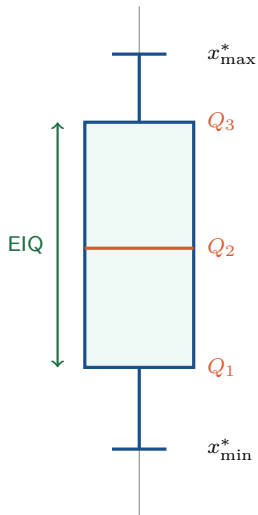
Écart interquartile (EIQ)

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

$$\text{EIQ} = 4,8 - 3,1 = \mathbf{1,7 \text{ t/ha}}$$

L'EIQ concentre les **50 % centraux** de la distribution.

Robuste : insensible aux valeurs extrêmes.



Lecture de la boîte :

- ▶ Trait central rouge : **médiane**
- ▶ Bords de la boîte : Q_1 et Q_3
- ▶ Longueur de la boîte : **EIQ**
- ▶ Moustaches : étendue "normale"
- ▶ Points isolés : **valeurs aberrantes**
(au-delà de $1,5 \times \text{EIQ}$)

```
boxplot(  
  rendement,  
  main = "Rendement",  
  ylab = "t/ha"  
)
```

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

L'idée derrière la variance

On veut mesurer à quel point chaque valeur s'éloigne de la **moyenne**.

Première idée : sommer les écarts $x_i - \bar{x}$

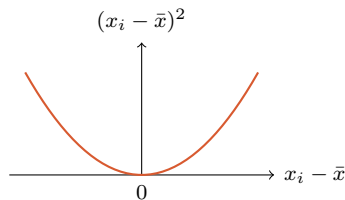
Problème

Les écarts positifs et négatifs s'annulent toujours :

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Solution : élever au carré avant de sommer

- ▶ $(x_i - \bar{x})^2 \geq 0$ toujours
- ▶ Les grands écarts sont **amplifiés**
- ▶ On obtient la **variance**



Variance de l'échantillon

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

On divise par $n - 1$ (et non n) pour obtenir un estimateur **non biaisé** de la variance de la population.

Écart-type

$$s = \sqrt{s^2}$$

Même unité que les données, c'est la mesure de dispersion la plus utilisée.

Pourquoi $n - 1$?

Avec n observations, on a **estimé** $\bar{x}$ à partir des mêmes données. Cela "consomme" un degré de liberté. Diviser par $n - 1$ corrige ce biais.

En R :

```
var(x)  # variance (n-1)
sd(x)   # ecart-type (n-1)
```

Attention : la "variance population"
 $\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ n'est utilisée que si l'on a le recensement complet.

Calcul pas à pas de la variance

Rendements de 5 parcelles (t/ha)

$$x_i : \quad 3,0 \quad 3,8 \quad 4,0 \quad 4,2 \quad 5,0 \quad \Rightarrow \quad \bar{x} = 4,0 \text{ t/ha}$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3,0	-1,0	1,00
3,8	-0,2	0,04
4,0	0,0	0,00
4,2	+0,2	0,04
5,0	+1,0	1,00
Total	0	2,08

$$s^2 = \frac{2,08}{5 - 1} = \frac{2,08}{4} = 0,52 \text{ (t/ha)}^2$$

$$s = \sqrt{0,52} \approx \mathbf{0,72 \text{ t/ha}}$$

```
rendement <- c(3.0, 3.8, 4.0, 4.2, 5.0)

# Moyenne
xbar <- mean(rendement)
xbar # 4.0

# Variance (methode manuelle)
ecarts2 <- (rendement - xbar)^2
sum(ecarts2) / (length(rendement) - 1)
# 0.52

# Fonction integree
var(rendement) # 0.52
sd(rendement) # 0.7211103
```

Unités

La variance s^2 s'exprime en **unités au carré** (ex. : $(\text{t/ha})^2$).

L'écart-type s retrouve l'unité d'origine (t/ha), d'où son usage plus intuitif.

Interprétation :

En moyenne, les rendements s'écartent de **0,72 t/ha** de la moyenne.

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
- 4. Coefficient de variation**
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Problème. Peut-on comparer directement l'écart-type du poids des noix de cajou (en grammes) avec l'écart-type de la hauteur des arbres (en mètres) ?

Non, les unités sont différentes, les moyennes aussi.

Coefficient de variation

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

Dispersion relative, **sans unité**.

Deux variables, une plantation

Hauteur des arbres

$$\bar{x} = 4,2 \text{ m} \quad s = 0,72 \text{ m}$$

$$CV = 0,72/4,2 \times 100 = \mathbf{17,1 \%}$$

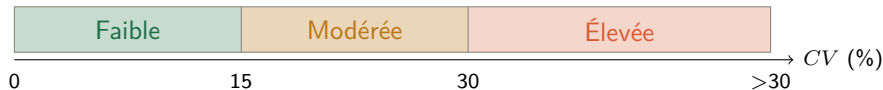
Poids des noix

$$\bar{x} = 6,8 \text{ g} \quad s = 2,3 \text{ g}$$

$$CV = 2,3/6,8 \times 100 = \mathbf{33,8 \%}$$

⇒ Le poids est **relativement** plus variable que la hauteur.

Interpréter le coefficient de variation



Usages en agronomie

- ▶ Comparer la variabilité de **deux cultures**
- ▶ Juger l'homogénéité d'un **essai expérimental**
- ▶ Évaluer la **fiabilité** d'une mesure répétée

Limite du CV

Le CV n'a de sens que si $\bar{x} > 0$ et si l'échelle a un **zéro absolu**. Il est inutilisable sur les températures en °C ou les pH.

CV en R et comparaison entre groupes

```
# R n'a pas de fonction cv() native
cv ← function(x) sd(x) / mean(x) * 100

# Comparer deux variables
hauteur ← c(3.2, 4.1, 2.8, 5.0, 3.7, 4.4, 6.2, 4.8, 3.5, 4.0)
poids ← c(5.1, 8.3, 4.2, 9.8, 6.5, 7.7, 11.2, 8.9, 5.8, 7.0)

cv(hauteur) # 18.4 %
cv(poids) m# 24.7 %

# Comparer deux groupes sur la meme variable
rendA ← c(3.8, 4.0, 4.1, 4.0, 3.9, 4.2)
rendB ← c(1.5, 6.8, 2.1, 7.2, 3.0, 5.4)

cv(rendA) # 3.5 %, tres homogene
cv(rendB) # 57.8 %, tres variable
```



1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Résumé à cinq chiffres

$(x_{\min}, Q_1, Q_2, Q_3, x_{\max})$

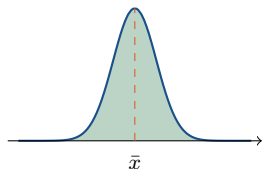
Ces cinq valeurs capturent **position** et **dispersion** sans aucune hypothèse sur la forme de la distribution.

En R :

```
summary(rendement)
# Min. 1st Qu. Median Mean
# 1.5 2.8 3.95 4.0
# 3rd Qu. Max.
# 5.4 7.2
fivenum(rendement)
```

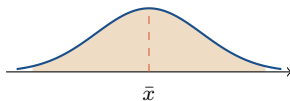
Plantation A vs Plantation B

	Plant. A	Plant. B
$x_{\min}$	3,8	1,5
Q_1	3,9	2,5
Q_2	4,0	4,0
Q_3	4,1	6,3
$x_{\max}$	4,2	7,2
$\bar{x}$	4,0	4,0
s	0,14	2,21
CV (%)	3,5	55,3
ElQ	0,2	3,8



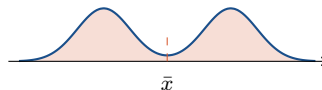
Faible dispersion

s petit : courbe étroite



Dispersion modérée

s moyen : courbe étalée



Forte dispersion

s grand : possible bimodalité

Un grand s peut signaler une distribution bimodale (deux groupes mélangés), d'où l'importance de visualiser toujours ses données.

Quelle mesure choisir ?

Situation	Mesure	Pourquoi	Exemple
Première exploration rapide	Étendue R	Immédiat	Vérifier les saisies
Distribution symétrique, pas de valeurs aberrantes	Écart-type s	Le plus complet	Essai variétal
Distribution asymétrique ou valeurs aberrantes	ElQ	Robuste	Revenus, surfaces
Comparer deux variables d'unités différentes	CV (%)	Sans unité	Poids vs hauteur
Rapport à un seuil réglementaire	s ou CV	Norme FAO/ISO	Teneur en eau

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Étude de cas : deux variétés de maïs

```
# Rendements en t/ha
# 12 parcelles/variete
variete_A <-c(4.2, 4.5, 3.8, 4.1, 4.6,
4.0, 3.9, 4.4, 4.3, 4.7, 4.1, 4.2)

variete_B <-c(2.8, 6.1, 3.5, 5.9, 4.0,
7.2, 3.1, 5.5, 4.8, 2.2, 6.8, 4.1)

# Resume complet
tapply(c(variete_A, variete_B),
  rep(c("A","B"), each=12),
  function(x) c(moy = mean(x),
    med = median(x), sd = sd(x),
    cv = sd(x)/mean(x)*100,
    iqr = IQR(x)
  ))
```

Résultats

	Variété A	Variété B
Moyenne	4,23	4,67
Médiane	4,20	4,45
s	0,26	1,61
CV (%)	6,1	34,5
EIQ	0,35	2,80

Question : quelle variété recommanderiez-vous à un petit producteur ? À une exploitation industrielle ?

Variété A

- ▶ Rendement moyen légèrement inférieur (4,23 t/ha)
- ▶ $CV = 6,1\%$: très **homogène**
- ▶ Risque faible d'une mauvaise récolte
- ⇒ Recommandée pour les **petits producteurs** (sécurité alimentaire)

Variété B

- ▶ Rendement moyen supérieur (4,67 t/ha)
- ▶ $CV = 34,5\%$: très **variable**
- ▶ Bonnes années excellentes, mauvaises années catastrophiques
- ⇒ Envisageable pour les **exploitations industrielles** avec gestion du risque

Message clé

En agronomie, la **stabilité** d'un rendement est souvent aussi importante que son niveau moyen. Un agriculteur ne choisit pas seulement la variété la plus productive, mais celle qui **échoue le moins souvent**.

```
# Boxplots compares
boxplot(variete_A, variete_B,
  names = c("Variete A", "Variete B"),
  col    = c("lightgreen", "lightyellow"),
  ylab   = "Rendement (t/ha)",
  main   = "Comparaison des varietes",
  notch  = FALSE)

# Ajouter les moyennes
points(
  1, mean(variete_A),
  pch = 18, col = "red", cex = 1.5
)
points(
  2, mean(variete_B),
  pch = 18, col = "red", cex = 1.5
)
```

À lire sur le graphique :

- ▶ La boîte de A est **courte** : faible EIQ
- ▶ La boîte de B est **longue** : fort EIQ
- ▶ Médiane (trait noir) $\approx$ moyenne (losange rouge) pour A
 $\Rightarrow$ distribution symétrique
- ▶ Pour B, vérifiez si des valeurs aberrantes apparaissent

Le losange rouge (moyenne) rajouté manuellement permet de voir l'asymétrie quand il s'écarte du trait médian.

1. Pourquoi mesurer la dispersion ?
2. Étendue et écart interquartile
3. Variance et écart-type
4. Coefficient de variation
5. Tableau de bord complet d'une série
6. Application agronomique complète
7. Récapitulatif

Les mesures

- ▶ **Étendue** R : simple, fragile
- ▶ **EIQ** $= Q_3 - Q_1$: robuste, résiste aux extrêmes
- ▶ **Variance** s^2 : base théorique
- ▶ **Écart-type** s : interprétable, même unité que x
- ▶ **CV** $= s/\bar{x} \times 100$: compare des variables différentes

Les réflexes

1. Toujours **visualiser** (boxplot, histogramme)
2. Rapporter **centre + dispersion** ensemble
3. Si valeurs aberrantes $\rightarrow$ préférer **médiane + EIQ**
4. Comparer des unités différentes $\rightarrow$ utiliser **CV**
5. Un grand CV indique **hétérogénéité**, chercher pourquoi

**Moyenne seule = information incomplète.
Toujours accompagner d'une mesure de dispersion.**

Merci de votre attention

Statistiques Descriptives · Module 5

Equipe Bioinformatique et Biostatistique

Laboratoire de Microbiologie, Biotechnologie et Bioinformatique

Institut National Polytechnique Félix Houphouët-Boigny

