

3. Représentation graphique des données

Statistique Descriptive

Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

2026–2027

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Un graphique n'est pas une décoration : c'est un **outil de communication scientifique**.

Cinq principes fondamentaux :

1. **Vérité** : ne pas déformer les données (axe Y à zéro pour les effectifs)
2. **Clarté** : une seule idée principale par graphique
3. **Efficacité** : maximiser le ratio information / encre
4. **Contexte** : titre, étiquettes d'axes, unités, source
5. **Accessibilité** : lisible en noir et blanc, polices suffisamment grandes

Attention : Anti-exemples classiques

Axe Y tronqué, camembert 3D, dégradés inutiles, couleurs sans signification.

Objectif	Type de variable	Graphique recommandé
Comparer des catégories	Qualitative	Diagramme en barres
Montrer des proportions	Qualitative (≤ 5 classes)	Camembert
Distribuer une variable	Continue	Histogramme, densité
Comparer des distributions	Continue $\times$ catégorie	Boîte à moustaches, violon
Explorer une relation	Quant. $\times$ Quant.	Nuage de points
Suivre dans le temps	Quant. $\times$ temps	Courbe

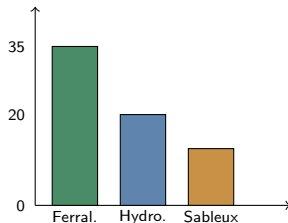
1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Règles clés :

- Barres **ne se touchent pas** (variable discrète / catégorielle)
- Axe Y commence à **zéro**
- Trier par effectif décroissant (sauf ordre naturel)
- Pas de 3D, pas d'effets décoratifs

Variantes :

- **Groupées** : comparer les effectifs
- **Empilées à 100 %** : comparer les proportions



```
library(ggplot2)
library(dplyr)

essai |>
  count(type_sol) |>
  mutate(type_sol = reorder(type_sol, n)) |>
  ggplot(aes(x = type_sol, y = n)) +
  geom_col(fill = "#1D9E75", width = 0.6) +
  geom_text(aes(label = n), hjust = -0.2) +
  coord_flip() +
  labs(x = "Type de sol", y = "Nombre de parcelles",
       caption = "Essai maïs - INPHB 2026")
```

```
# Groupées : comparer les effectifs
```

```
ggplot(essai, aes(x = zone, y = n, fill = type_sol)) +  
  geom_col(position = "dodge", width = 0.7)
```

```
# Empilées a 100 % : comparer les proportions
```

```
essai |>
```

```
  count(zone, type_sol) |>
```

```
  group_by(zone) |>
```

```
  mutate(pct = n / sum(n)) |>
```

```
  ggplot(aes(x = zone, y = pct, fill = type_sol)) +
```

```
  geom_col(position = "stack") +
```

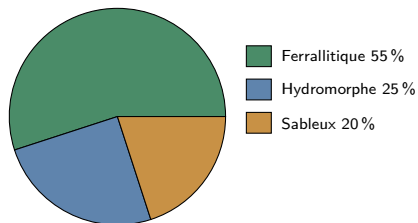
```
  scale_y_continuous(labels = scales::percent_format())
```

Attention : Quand éviter le camembert

- Plus de 5 classes
- Parts proches (ex. 28 % vs 32 %)
- Jamais en 3D

Quand l'accepter

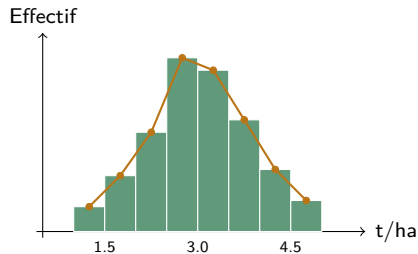
- ≤ 5 classes seulement
- Une part domine nettement ($> 50\%$)



1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Choix de l'axe Y :

- **Effectifs** : lecture directe du tableau
- **Fréquences relatives** : comparaison entre groupes
- **Densité** (f_i/h) : superposer une loi théorique



Attention : Rappel

Les barres **se touchent** (variable continue).

Nombre de classes :

règle de Sturges $k = 1 + \log_2 n$.

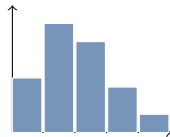
Histogramme avec ggplot2

```
# Histogramme (effectifs)
ggplot(essai, aes(rdt_tha)) +
  geom_histogram(bins = 12,
    fill = "#1D9E75", colour = "white") +
  labs(x = "Rendement (t/ha)", y = "Effectif")

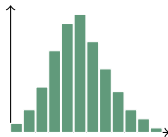
# Histogramme + courbe de densite
ggplot(essai, aes(rdt_tha)) +
  geom_histogram(
    aes(y = after_stat(density)),
    bins = 12, fill = "#1D9E75",
    colour = "white", alpha = 0.8) +
  geom_density(colour = "#BA7517", linewidth = 1)
```

Impact du choix de bins sur la lecture :

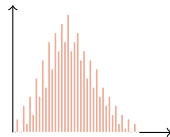
Trop peu (5)



Optimal (12)



Trop (40)



Règle de Sturges

Pour $n = 80$: $k = 1 + \log_2(80) \approx 7.3 \rightarrow 8$ classes. Ajuster visuellement si nécessaire.

Histogrammes par groupe : facettes

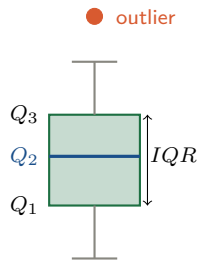
```
ggplot(essai, aes(rdt_tha, fill = variete)) +  
  geom_histogram(bins = 10,  
    colour = "white", show.legend = FALSE) +  
  facet_wrap(~ variete, nrow = 2) +  
  geom_vline(  
    data = essai |>  
      group_by(variete) |>  
      summarise(m = mean(rdt_tha)),  
    aes(xintercept = m),  
    colour = "#D85A30",  
    linetype = "dashed") +  
  labs(x = "Rendement (t/ha)",  
    y = "Effectif",  
    caption = "Ligne rouge = moyenne")
```

Quand utiliser les facettes ?

Comparer la distribution d'une variable continue entre plusieurs groupes (≤ 6).

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
- 4. Boîte à moustaches**
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Élément	Valeur
Bord inférieur	Q_1 (25 ^e pct.)
Ligne centrale	Médiane Q_2
Bord supérieur	Q_3 (75 ^e pct.)
Hauteur	$IQR = Q_3 - Q_1$
Moustache basse	$Q_1 - 1.5 \times IQR$
Moustache haute	$Q_3 + 1.5 \times IQR$
Points isolés	Outliers potentiels



```
ggplot(essai,  
  aes(x = reorder(variete, rdt_tha, median),  
    y = rdt_tha, fill = zone)) +  
  geom_boxplot(  
    position = position_dodge(0.8),  
    width = 0.6,  
    outlier.size = 1.5) +  
  stat_summary(  
    aes(group = zone), fun = mean,  
    geom = "point", shape = 4, size = 2,  
    position = position_dodge(0.8),  
    colour = "white") +  
  labs(x = "Variete (triee par mediane)",  
    y = "Rendement (t/ha)",  
    caption = "Croix = moyenne")
```

Boîte à moustaches

- + Lisible même sur $n < 30$
- + Q_1 , Q_2 , Q_3 bien visibles
- + Outliers identifiés clairement
- Masque bimodalité et asymétrie

Violon

- + Montre la forme complète
- + Révèle bimodalité, plateaux
- Peu fiable si $n < 30$
- Plus difficile à lire

Pratique recommandée

Combiner violon + boîte intérieure : `geom_violin()` + `geom_boxplot(width = 0.15)`

Attention : Limite du violon

Déconseillé si $n < 30$ par groupe, la densité estimée est peu fiable.

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Un point = une observation.

x = première variable, y = deuxième variable.

Enrichissements possibles :

Canal	Utilisation
Couleur	3 ^e variable catégorielle (ex. zone agro-écologique)
Forme	4 ^e variable catégorielle (ex. niveau de fumure)
Taille	Variable quantitative (<i>bubble chart</i>)
<code>geom_smooth</code>	Tendance générale (LOESS ou droite de régression)
<code>alpha</code>	Densité des points (superposition)

```
r_val <- cor(essai$pluvio_mm, essai$rdt_tha)

ggplot(essai, aes(pluvio_mm, rdt_tha,
                  colour = zone,
                  shape = fumure)) +
  geom_point(size = 2.5, alpha = 0.7) +
  geom_smooth(method = "lm",
             se = FALSE, linewidth = 0.8,
             show.legend = FALSE) +
  annotate("text", x = 780, y = 6.0,
         label = paste0("r = ", round(r_val, 3)),
         hjust = 0, size = 4,
         colour = "#D85A30",
         fontface = "bold") +
  labs(x = "Pluviometrie (mm)", y = "Rendement (t/ha)")
```

Nuage de points avec taille (*bubble chart*)

```
essai |>
  group_by(variete, zone) |>
  summarise(
    rdt_moy = mean(rdt_tha),
    epis_moy = mean(nb_epis),
    n       = n(), .groups = "drop") |>
  ggplot(aes(x = epis_moy,
             y = rdt_moy,
             colour = variete,
             size   = n,
             label  = variete)) +
  geom_point(alpha = 0.8) +
  ggrepel::geom_text_repel(
    size = 3, show.legend = FALSE) +
  scale_size_continuous(range = c(4,12))
```

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Éléments obligatoires :

- Titre ou légende de figure
- Étiquettes des axes avec **unités** (`labs(x = "Rendement (t/ha)")`)
- Légende si couleur ou forme encode une variable

Éléments recommandés :

- Sous-titre : contexte et taille d'échantillon
- Source dans la caption (`labs(caption = ...)`)
- Lignes de référence agronomique (`geom_hline`)

Annotation de seuil

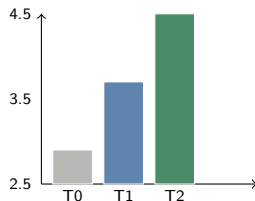
```
geom_hline(yintercept = 3.0, linetype = "dashed")  
+ annotate("text", ...)
```

Indique le seuil de rentabilité directement sur le graphique.

L'anti-exemple : l'axe Y tronqué

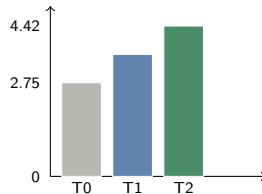
À éviter

Axe Y commence à 2.5 t/ha



Correct

Axe Y commence à 0



Attention : Règle absolue

L'axe Y d'un diagramme en barres commence **toujours à zéro**. Un axe tronqué gonfle artificiellement les différences visuelles.

1. Principes d'une bonne visualisation
2. Variables qualitatives : barres et camembert
3. Variables continues : histogramme et densité
4. Boîte à moustaches
5. Nuage de points
6. Bonnes pratiques de visualisation
7. Résumé et exercices

Objectifs atteints :

1. Sélectionner le bon graphique selon variable et objectif
2. Barres simples, groupées, empilées avec ggplot2
3. Histogramme (continu) vs diagramme en bâtons (discret)
4. Lire une boîte à moustaches : Q_1 , Q_2 , Q_3 , IQR , moustaches, outliers
5. Comparer des distributions : boîtes et violons
6. Nuage de points enrichi (couleur, forme, taille)
7. Bonnes pratiques : axe à 0, étiquettes, source, seuils agronomiques

Prochain module

M4 : Tendance centrale :

moyenne arithmétique, géométrique, harmonique, médiane, mode.

Exercice 3.1 : Choisir et critiquer un graphique

Quel graphique utiliser dans chaque situation ?

1. Comparer la fréquence de 4 types de ravageurs observés sur 200 parcelles de maïs.
2. Montrer la distribution des teneurs en phosphore (mg/kg) de 80 échantillons de sol.
3. Explorer la relation entre dose d'azote (kg/ha) et rendement (t/ha) sur 60 parcelles.
4. Comparer les distributions de rendement de 5 variétés de sorgho.
5. Montrer que CMS8704 représente 45 % des semences certifiées distribuées (TZPB 30 %, local 25 %).



Exercice 3.2 : Produire un graphique annoté

À partir du jeu de données arachide (60 parcelles, 3 niveaux de fumure phosphatée) :

```
set.seed(99)
arachide <- data.frame(
  fumure = rep(c("P0", "P30", "P60"),
              each = 20),
  rdt_tha = c(rnorm(20, 1.2, 0.25),
              rnorm(20, 1.8, 0.30),
              rnorm(20, 2.3, 0.35)),
  pluvio = round(rnorm(60, 900, 120))
)
```

1. Histogramme des rendements + courbe de densité.
2. Boîtes à moustaches par fumure + seuil de rentabilité 1.5 t/ha.
3. Nuage rendement $\times$ pluviométrie, coloré par fumure, droite de régression par groupe.

Merci de votre attention

Statistiques Descriptives · Module 3

Equipe Bioinformatique et Biostatistique

Laboratoire de Microbiologie, Biotechnologie et Bioinformatique

Institut National Polytechnique Félix Houphouët-Boigny

