

2. Collecte et organisation des données

Statistique Descriptive

Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

2026-2027

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

Pourquoi échantillonner ?

- Impossible ou trop coûteux de mesurer toute la population
- Exemples : analyse de sol, essai variétal, enquête paysanne

Deux conditions de validité :

1. Échantillon assez **grand** (précision)
2. Échantillon **représentatif** (absence de biais)

Attention : Principe fondamental

Un mauvais plan d'échantillonnage ne peut pas être corrigé par des statistiques sophistiquées.

Terme	Définition	Exemple
Population (N)	Toutes les unités d'intérêt	Champs de cacao de Côte d'Ivoire
Unité de sondage	Élément de base	Un champ de cacao
Échantillon (n)	Sous-ensemble observé	200 champs enquêtés
Base de sondage	Liste exhaustive	Registre foncier de la région
Paramètre	Valeur vraie (inconnue)	μ, σ
Estimateur	Valeur calculée sur n	$\bar{x}, s$

Qu'est-ce qu'un échantillon représentatif ?

Définition : Représentativité

Un échantillon est **représentatif** si sa composition reflète fidèlement la **diversité** de la population d'origine sur toutes les caractéristiques pertinentes pour l'étude.

Un échantillon représentatif doit :

- Inclure toutes les **sous-populations** présentes
- Respecter leurs **proportions relatives**
- Être sélectionné **sans favoritisme** (pas de biais de sélection)
- Couvrir la **variabilité** naturelle de la population

Contre-exemple classique

Estimer le rendement moyen en cacao d'une région en enquêtant uniquement les planteurs présents à une réunion coopérative.

⇒ Biais : les membres actifs d'une coopérative ont souvent des pratiques agricoles **au-dessus de la moyenne**.

1. Biais de sélection

Certains individus ont une probabilité nulle ou très faible d'être sélectionnés.

Exemples :

- Enquêter uniquement en bord de route
- Se limiter aux exploitations accessibles en saison sèche
- N'interroger que les hommes chefs de ménage

2. Taille insuffisante

Même sans biais, un échantillon trop petit laisse passer des sous-groupes entiers ou amplifie les fluctuations aléatoires.

Attention : Règle d'or

Un grand échantillon biaisé est **toujours moins fiable** qu'un petit échantillon bien construit.

La représentativité prime sur la taille.

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

Échantillonnage aléatoire simple

Principe : chaque unité a la même probabilité d'être tirée.

- + Simple, estimateurs non biaisés
- + Formules statistiques standard applicables
- Nécessite une base de sondage complète
- Peu efficace si population hétérogène

Mise en œuvre en R

```
sample(1:N, size=n, replace=FALSE)
```

```
set.seed(2024)  
# Tirer 40 parcelles parmi 500  
echantillon <- sample(1:500, size = 40, replace = FALSE)
```

Échantillonnage stratifié

Principe : diviser la population en strates homogènes, puis effectuer un EAS dans chaque strate.

Strate	N_h	Poids	n_h ($n = 60$)
Zone forêt	300	50 %	30
Zone savane	180	30 %	18
Zone côtière	120	20 %	12
Total	600	100 %	60

Quand stratifier ?

Quand les strates sont connues *a priori* et que la variable d'intérêt varie beaucoup entre strates.

Principe : tirer le premier élément au hasard, puis prélever tous les $k = \lfloor N/n \rfloor$ éléments.

```
N <- 500; n <- 40  
k <- floor(N / n) # pas de sondage = 12  
debut <- sample(1:k, 1) # premier element aleatoire  
ech <- seq(debut, by = k, length.out = n)
```

Transect de terrain

Marcher en ligne droite, prélever un échantillon de sol tous les 10 pas.

Attention : Périodicité

Si la population est périodique avec le même pas k , risque de biais.

Échantillonnage en grappes

Principe : la population est divisée en **grappes** naturelles (villages, écoles, champs). On tire aléatoirement des grappes entières, puis on observe **tous** les individus de la grappe sélectionnée.

Exemple agronomique

Estimer la production d'arachide dans une région de 80 villages.

Étape 1 : tirer au sort 8 villages (grappes).

Étape 2 : enquêter **tous** les planteurs de ces 8 villages.

- + Aucune liste individuelle requise
- + Économique (déplacements groupés)
- Moins précis si les grappes sont homogènes entre elles
- Risque de sur-représenter les caractéristiques locales

Attention : Grappes $\neq$ Strates

Dans la stratification on prélève **une partie** de chaque strate. Dans les grappes on prélève **tout** d'un petit nombre d'unités.

Définition : Définition

L'échantillonnage de convenance sélectionne les individus les plus **facilement accessibles**, sans procédure aléatoire.

Exemples courants :

- Interroger les premiers planteurs rencontrés au marché
- Analyser les parcelles situées en bord de route
- Enquêter uniquement dans son propre village

Attention : Biais systématique

Les individus facilement accessibles sont rarement représentatifs de la population entière. Les résultats **ne peuvent pas être généralisés**.

Attention : usage acceptable

Acceptable en phase **exploratoire** (piloter un questionnaire, tester un protocole) mais **jamais** pour produire une estimation officielle ou publier un résultat scientifique.

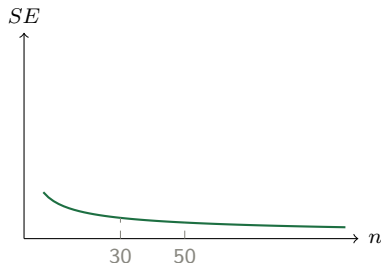
Comparaison des méthodes d'échantillonnage

Méthode	Avantages	Inconvénients	Usage agronomique
Aléatoire simple	Simple, estimateurs non biaisés	Nécessite liste complète	Essais homogènes
Stratifié	Précis si hétérogène	Nécessite connaître les strates	Enquêtes multi-zones
Systématique	Rapide sur le terrain	Risque si périodicité = k	Transects, inventaires
En grappes	Pas de liste individuelle	Moins précis	Zones sans registre
Convenance	Rapide, peu coûteux	Biais systématique	Pilotage seulement

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
- 3. Taille d'échantillon**
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

L'erreur standard de la moyenne :

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$



Rendements décroissants

Diviser SE par 2

$\Rightarrow$ quadrupler n .

Au-delà de $n \approx 30-50$, les gains deviennent marginaux.

Formule pratique pour calculer n

Estimer une moyenne avec marge d'erreur e (IC 95 %) :

$$n \approx \left(\frac{1.96 \cdot s}{e} \right)^2$$

Rendement maïs

$s = 0.8$ t/ha $e = \pm 0.3$ t/ha, $n \approx (1.96 \times 0.8/0.3)^2 = \lceil 27.3 \rceil = \mathbf{28}$ parcelles

Attention : D'où vient s ?

s est inconnu avant l'enquête. On l'estime à partir de :

- une **étude pilote** (15–30 observations préliminaires) ;
- une **étude antérieure** sur la même culture ou région ;
- l'**étendue** : $s \approx (x_{\max} - x_{\min})/4$ (provenant de la loi normale).

```
s <- 0.8; e <- 0.3; z <- 1.96  
n_min <- ceiling((z * s / e)^2) # n_min = 28
```

Taille d'échantillon pour estimer une proportion

Estimer une proportion p avec marge d'erreur e (IC 95 %) :

$$n \approx \frac{z^2 p (1 - p)}{e^2}$$

- $z = 1,96$ pour un niveau de confiance de 95 %
- Si p est inconnu, toujours utiliser $p = 0,5$

Proportion d'exploitations utilisant des engrais

On veut estimer cette proportion avec $e = \pm 1 \%$, p inconnu.

$$n \approx \frac{1,96^2 \times 0,25}{0,01^2} = \frac{3,8416 \times 0,25}{0,0001} = \frac{0,9604}{0,0001} = \mathbf{9604} \text{ exploitations}$$

```
z <- 1.96; p <- 0.5; e <- 0.01  
n_prop <- ceiling(z^2 * p * (1 - p) / e^2) # n_prop = 9604
```

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

Règles *tidy data* (Wickham, 2014) :

1. 1 variable = 1 colonne
2. 1 observation = 1 ligne
3. 1 valeur = 1 cellule

Non-tidy (format large) :

Parcelle	T0	T1	T2
P1	2.8	3.5	4.1
P2	3.0	3.8	4.3

Tidy (format long) :

Parcelle	Traitement	Rendement
P1	T0	2.8
P1	T1	3.5
P1	T2	4.1
P2	T0	3.0

Transformer large → long avec R

```
library(tidyr)

messy <- data.frame(
  parcelle = paste0("P", 1:5),
  T0 = c(2.8, 3.0, 2.6, 2.9, 3.1),
  T1 = c(3.5, 3.8, 3.3, 3.6, 3.9),
  T2 = c(4.1, 4.3, 3.9, 4.2, 4.5)
)

tidy_df <- messy |>
  pivot_longer(
    cols = starts_with("T"),
    names_to = "traitement",
    values_to = "rendement_tha"
  )
```

À faire	À éviter
<code>rdt_tha</code>	Rendement (t/ha), Rdt maïs
NA pour valeur manquante	Cellule vide, tiret, "nd"
Dates ISO : 2024-05-14	14/05/24, 14 mai 2024
Modalités cohérentes	"Forêt" / "forêt" / "Foret"
Séparateur décimal : point .	Mélanger , et .

Validation rapide

`table(df$variete)` révèle les fautes de frappe et incohérences de casse.

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

Variable discrète : tableau de fréquences

Pour chaque modalité i :

$$f_i = \frac{n_i}{n} \quad F_i = \sum_{j \leq i} f_j$$

Nb épis	Effectif n_i	Fréq. f_i	Fréq. cum. F_i
1	3	5.0 %	5.0 %
2	24	40.0 %	45.0 %
3	27	45.0 %	90.0 %
4	6	10.0 %	100.0 %
Total	60	100 %	

F_i répond à : « Quelle proportion a une valeur $\leq i$? »

Regrouper les valeurs en k classes d'amplitude égale.

Règle de Sturges :

$$k = 1 + \log_2(n)$$

Règle racine carrée :

$$k = \lceil \sqrt{n} \rceil$$

Conseil pratique

Choisir des bornes rondes ayant un sens agronomique.

Ex. : classes de 0.5 t/ha d'amplitude pour le rendement.

```
# n = 80, Sturges -> k = 8  
bornes <- seq(1.0, 5.5, by = 0.5)  
classes <- cut(rdt, breaks = bornes, right = FALSE)
```

Tableau de fréquences : rendements maïs ($n = 80$)

Classe [a, b[	Effectif n_i	Fréq. f_i	Fréq. cum. F_i
[1.0 ; 1.5[	2	2.5 %	2.5 %
[1.5 ; 2.0[	5	6.3 %	8.8 %
[2.0 ; 2.5[	8	10.0 %	18.8 %
[2.5 ; 3.0[	14	17.5 %	36.3 %
[3.0 ; 3.5[	19	23.8 %	60.0 %
[3.5 ; 4.0[	17	21.3 %	81.3 %
[4.0 ; 4.5[	10	12.5 %	93.8 %
[4.5 ; 5.5[	5	6.3 %	100.0 %
Total	80	100 %	

Histogramme

- Barres **contiguës** (variable continue)
- Hauteur $\propto$ effectif ou densité
- Polygone des fréquences : relier les centres

Courbe ogive

- Fréquences cumulées F_i aux bornes supérieures
- Lecture graphique des percentiles
- Médiane : intersection avec $F = 50\%$

Attention : Histogramme $\neq$ diagramme en bâtons

Dans un histogramme, les barres se **touchent** (classes contiguës). Un diagramme en bâtons laisse des espaces (variable discrète ou qualitative).

1. Pourquoi l'échantillonnage ?
2. Méthodes d'échantillonnage
3. Taille d'échantillon
4. Structure et encodage des données
5. Tableaux de fréquences
6. Résumé et exercices

Objectifs atteints :

1. Nécessité et conditions de l'échantillonnage
2. EAS, stratifié, systématique, mise en œuvre dans R
3. Calcul de taille d'échantillon ($n \approx (1.96 s/e)^2$)
4. Format *tidy*, bonnes pratiques de saisie
5. Tableau de fréquences : variables discrètes et continues
6. Histogramme et courbe ogive

Prochain module

M3 : Représentation graphique :
barres, histogrammes, boîtes à moustaches, nuages de points.

Exercice 2.1 : Choisir une méthode d'échantillonnage

Pour chaque situation, choisir EAS, stratifié ou systématique :

1. Estimer la teneur en matière organique d'une parcelle expérimentale homogène de 2 ha (20 prélèvements).
2. Enquête sur les rendements en igname dans 3 zones agro-écologiques : 400 + 250 + 150 exploitations.
3. Contrôle qualité d'un lot de 5 000 sachets de semences certifiées (50 sachets à prélever).

Exercice 2.2 : Tableau de fréquences et ogive

Hauteurs de plant de sorgho (cm), $n = 30$:

115	122	108	131	119	125	112	138	127	118
105	133	121	116	129	123	110	136	124	117
130	114	128	120	107	134	126	113	122	119

1. Nombre de classes (Sturges) ; proposer des bornes rondes.
2. Tableau de fréquences (n_i , f_i , F_i).
3. Proportion de plants > 125 cm ?
4. Histogramme et ogive dans R.

Merci de votre attention

Statistiques Descriptives · Module 2

Institut National Polytechnique Félix Houphouët-Boigny · INPHB

