Populations d'étude, Echantillonnage, Estimations

Ter Tiero Elias DAH MD,PhD
AHU Santé Publique – Option Epidémiologie
Université de Ouahigouya
Désiré Lucien Dahourou MD, PhD
Chargé de Recherche – Santé Publique/Epidémiologie
IRSS/CNRST

Plan

1. Populations à enquêter ou populations d'étude
2. Echantillonnages
   i. Echantillonnages non aléatoires
   ii. Echantillonnages aléatoires
3. Estimations des paramètres dans un échantillon
4. Calcul de la taille d'un échantillon

I. Populations d’étude

Quelle population faut-il étudier ? (1)

Quelques définitions…(Larousse)

Population

- Ensemble d'éléments (êtres humains, êtres vivants ou objets) soumis à une étude statistique
    - Ensemble des adolescentes et jeunes filles (âgées de 15-24 ans) inscrites dans les établissements d'enseignements publics ou privés du Yatenga en 2021
    - Ensemble des dossiers médicaux des patients admis aux UM du CHUR de Ouahigouya en 2020.

Unité statistique

- Tout élément faisant partie de la population visée
- Elément pour lequel des données sont recueillies
    - Une adolescente de 15 ans inscrite dans un établissement public ou privé du Yatenga en 2021
    - Un dossier médical d'un patient admis aux UM du CHUR de Ouahigouya en 2020

Quelle population faut-il étudier ? (2)

Quelques définitions

Population cible

- Population qui est définie dans la formulation de l'objectif général. La précision d'informations sur les personnes, le temps et le lieu permet de la définir.
- Population qui a motivé au départ la mise en place de l'enquête (population conceptuelle)
- Population à laquelle on appliquera les résultats de l'étude.

<u>Exemple:</u> si nous voulons mettre en place une étude dont l'objectif général est:

Estimer la prévalence d'utilisation des contraceptifs modernes chez les adolescentes et jeunes filles âgées de 15-24 ans inscrites dans les établissements d'enseignements publics ou privés du Yatenga en 2021.

Population cible ?

adolescentes et jeunes filles âgées de 15-24 ans inscrites dans les établissements d'enseignements publics ou privés du Yatenga en 2021.

Quelle population faut-il étudier ? (3)

Quelques définitions

Population source

- Sous-ensemble de la population cible
- la population accessible lors de l'enquête selon la stratégie définie.
- C'est donc au sein de cette population-source que les participants de l'enquête (c'est-à-dire l'échantillon) seront sélectionnés.

<u>Dans l'exemple:</u>

Population source ?

adolescentes et jeunes filles âgées de 15-24 ans inscrites dans les établissements d'enseignements publics ou privés du Yatenga en 2021

  - *Présentes à l'école le jour de l'enquête*
  - *Qui accepteront de participer*
  - *Dont les parents accepteront la participation (assentiment) pour celles qui ont moins de 18 ans*

NB: Dans certains cas, populations cible & source peuvent être les mêmes

Quelle population faut-il étudier ? (4)
Quelques définitions

Couverture (de la population cible par la population source)

- Sous-entend 2 critères
  - Quantitatif
    - correspond à la proportion que représente l'effectif de la population source par rapport à l'effectif de la population cible
  - Qualitatif
    - correspond au caractère de ressemblance ou de dissemblance des unités appartenant à la population-cible et non à la population-source par rapport aux unités appartenant aux deux populations. Autrement dit, les unités de la population cible exclues de la population source sont-elles des unités de la population cible incluses de la population source?

<u>Dans l'exemple:</u>

Critère quantitatif: Calculer la % d'absents, de refus de participer, et de non autorisés à participer.

Critère qualitatif: Est-ce que l'utilisation de méthodes contraceptives diffèrent entre la population source et l'ensemble des filles qui ne seront pas inclus? (Difficile à savoir. On peut néanmoins faire des hypothèses)

Que faire pour améliorer la couverture? Bonne communication

Quelle population faut-il étudier ? (5)
Quelques définitions

Recenser ou échantillonner ?

Dès lors que la population-source est définie, 2 issues sont possibles pour obtenir les unités qui seront incluses dans l'enquête :

- **Recensement**
  Inclusion de la totalité des unités statistiques de la population source dans l'enquête
- **Echantillonnage ou sondage**
  Sélection d'un sous-ensemble de la population source pour participer à l'enquête

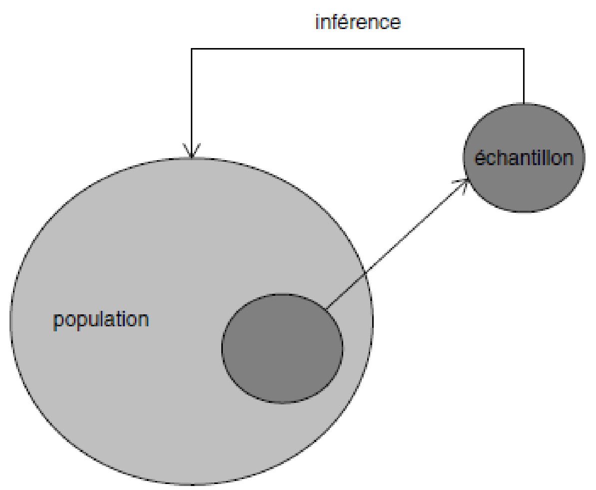
Comment faire le choix entre un recensement et un échantillonnage ? **2 principaux critères** :

- **Taille de la population source**
  - Plus elle est importante, plus il est difficile de faire un recensement (logistique & coûts)
- **Durée de l'enquête**
  - Plus la période de réalisation de l'enquête est courte, plus il est difficile de faire un recensement

Quelle population faut-il étudier ? (6)
Quelques définitions

Echantillon

Un sous-ensemble d'unités statistiques de la popu

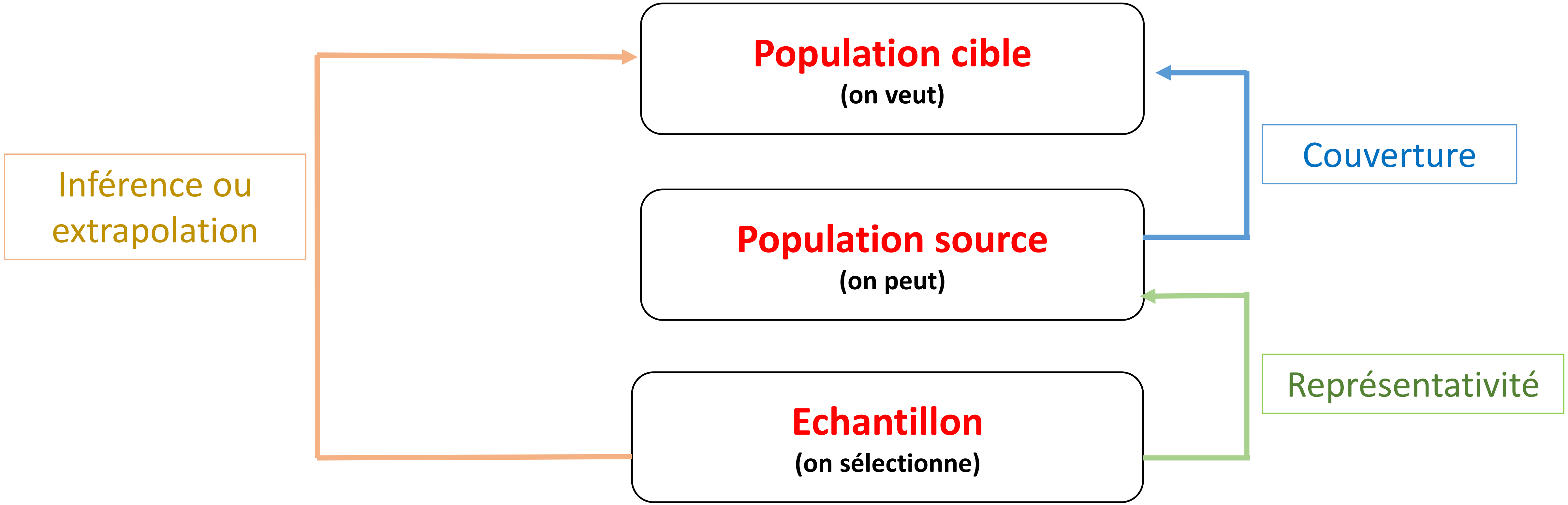


Echantillonnage ou sondage

Processus de sélection d'un échantillon au sein de la population source

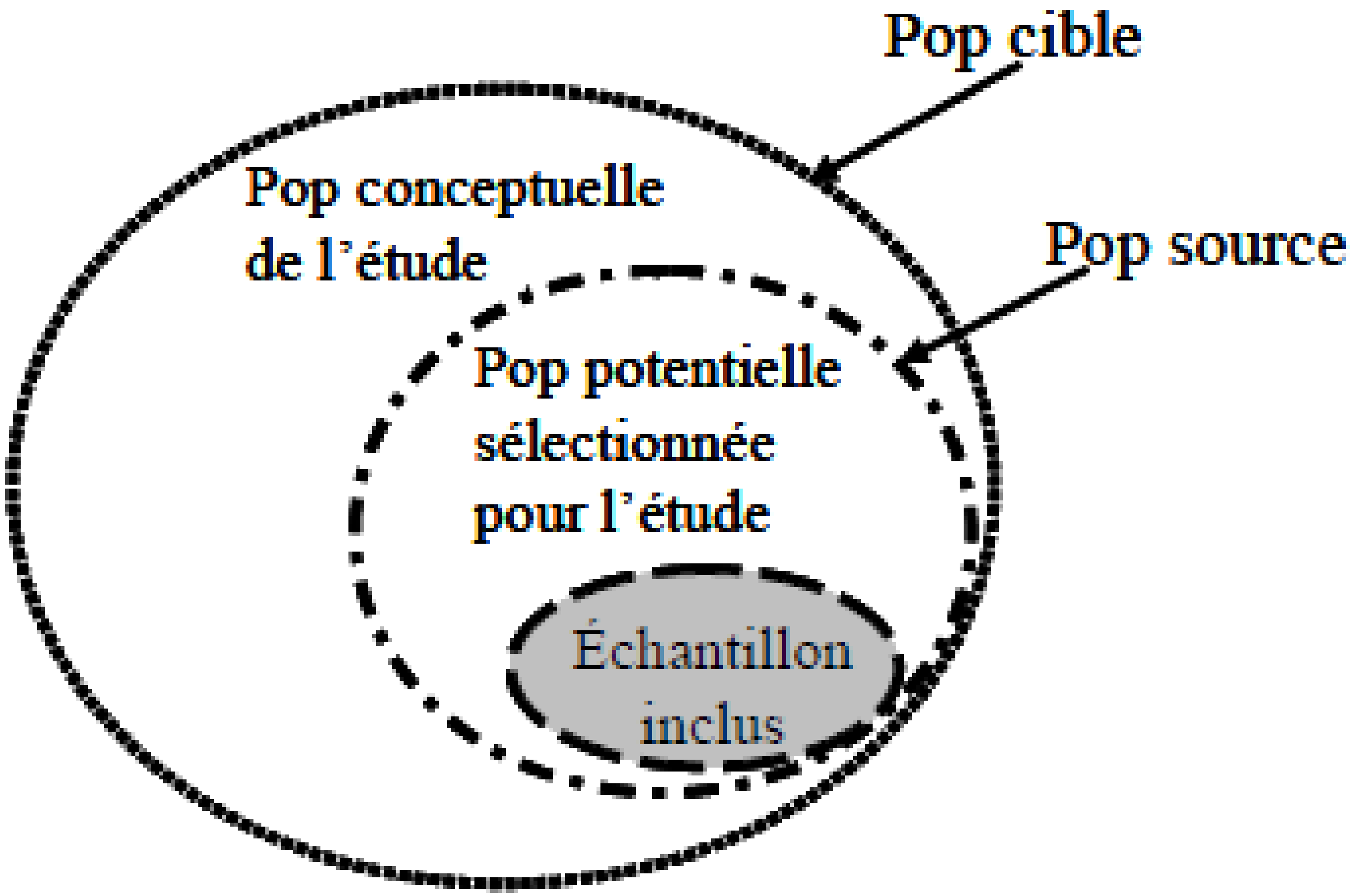
Quelle population faut-il étudier ? (7)

Résumé



Quelle population faut-il étudier ? (8)

Résumé



II. Echantillonnage ou sondage

Echantillonnage ou sondage (1)

Echantillon

- Ensemble représentatif d’une « population-mère » possédant les mêmes caractéristiques » (Larousse)
- Un sous-ensemble permettant de prévoir les informations inconnues de la population source

Echantillonnage

- Procédure de sélection d’un échantillon au sein de la population source.

2 grands groupes selon le type de procédure

- **Echantillonnage non aléatoire ou non probabiliste ou empirique**
  - l’échantillon se construit par inclusion choisie par l’enquêté ou l’enquêteur
  - Il n’y a donc pas de hasard.
  - Certaines personnes n’ont aucune chance d’être sélectionnées
- **Echantillonnage aléatoire ou probabiliste**
  - Chaque unité statistique de la population est incluse dans l’échantillon à la suite d’un tirage au sort
  - Toute unité statistique (personne) a une chance d’être incluse dans l’échantillon

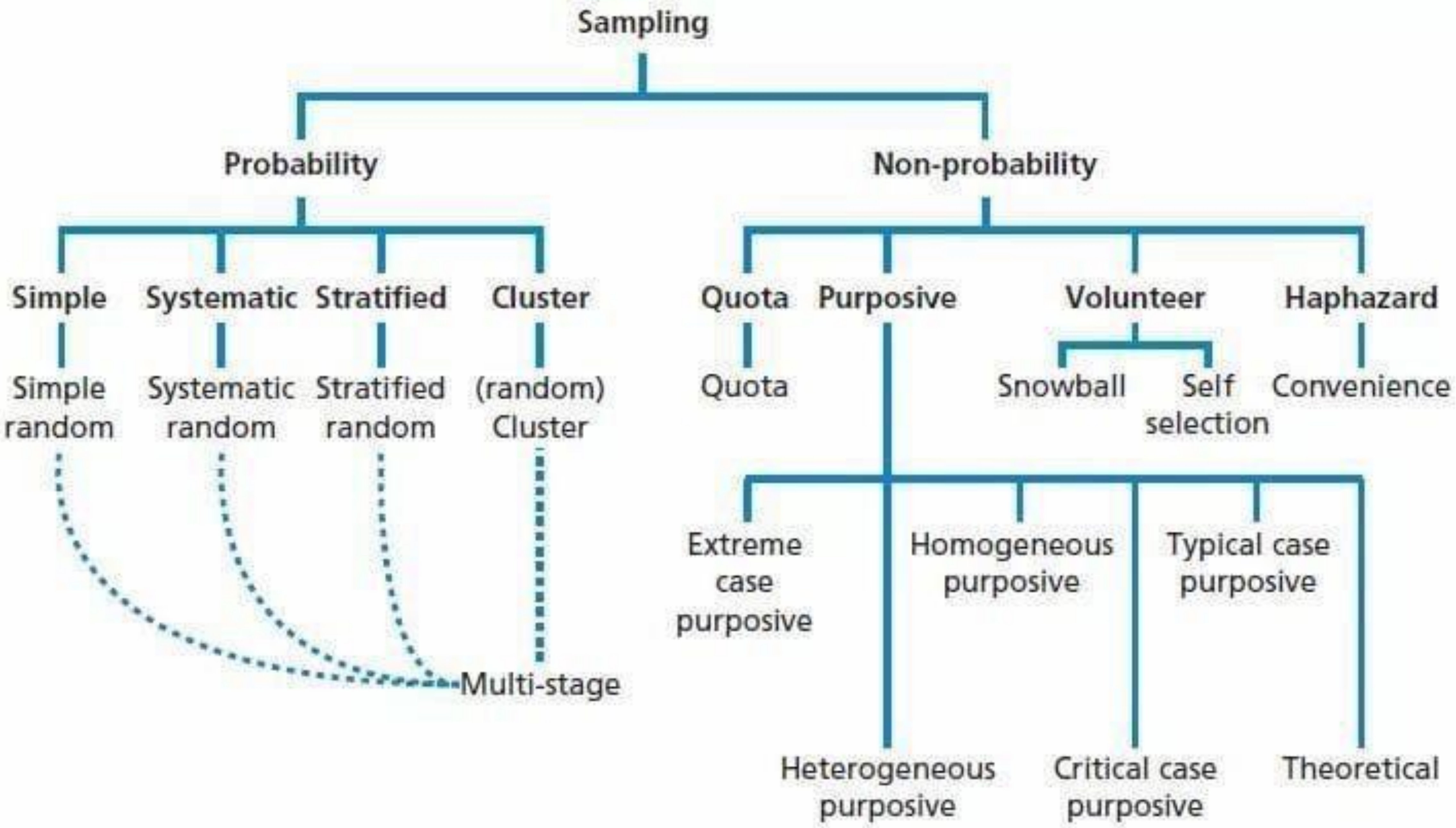
Echantillonnage ou sondage (2)

Echantillonnage non aléatoire

- Echantillonnage de convenance
- Echantillonnage de volontaires
- Méthode des unités types
- Méthode des quotas

Echantillonnage aléatoire

- Echantillonnage aléatoire simple
- Echantillonnage systématique
- Echantillonnage (aléatoire) stratifié
- Echantillonnage en grappes
- Echantillonnage aléatoire à plusieurs degrés
- Echantillonnage aléatoire en plusieurs étapes (phases)

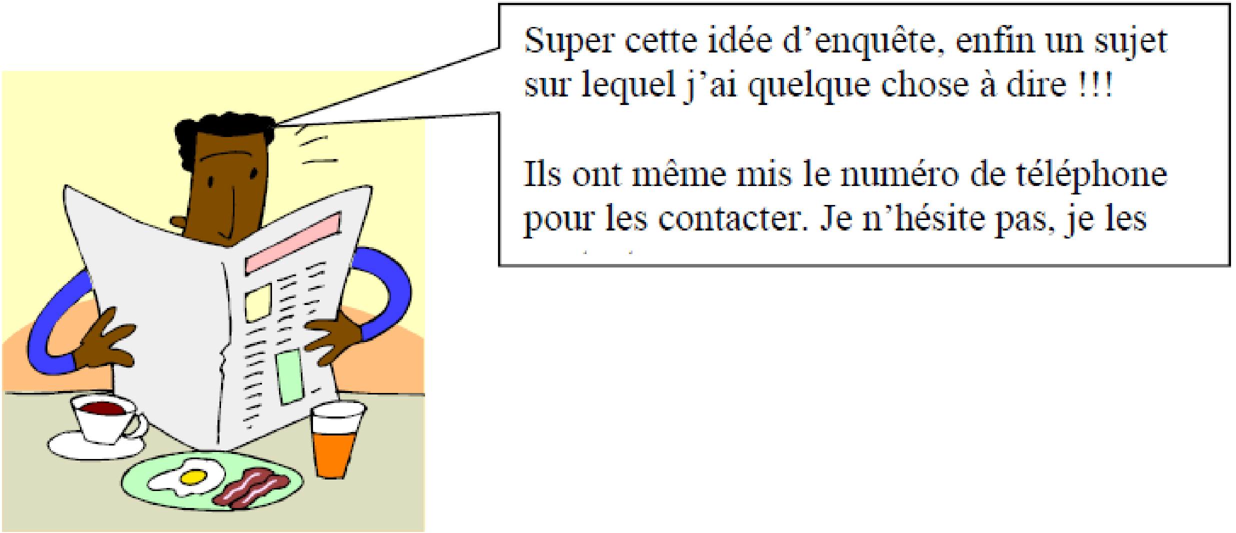

Sampling
Probability
Non-probability
Simple
Systematic
Stratified
Cluster
Simple random
Systematic random
Stratified random
(random) Cluster
Multi-stage
Quota
Purposive
Volunteer
Haphazard
Quota
Snowball
Self selection
Convenience
Extreme case purposive
Homogeneous purposive
Typical case purposive
Heterogeneous purposive
Critical case purposive
Theoretical

II.1. Echantillonnage non aléatoire

Echantillonnage non aléatoire (1)

Echantillonnage par volontariat ou de volontaires

- Les unités se sélectionnent d’elles-mêmes
- Méthode de sélection éthique car elle laisse la possibilité aux gens de participer
- Mais il est difficile de juger de la représentativité de l’échantillon
  (mesurer les différences entre les volontaires & la population source)
- Boule de neige: exemple d’échantillonnage utilisé pour les enquêtes auprès de populations cachées et/ou difficiles à joindre (toxicomanes, homosexuels...)

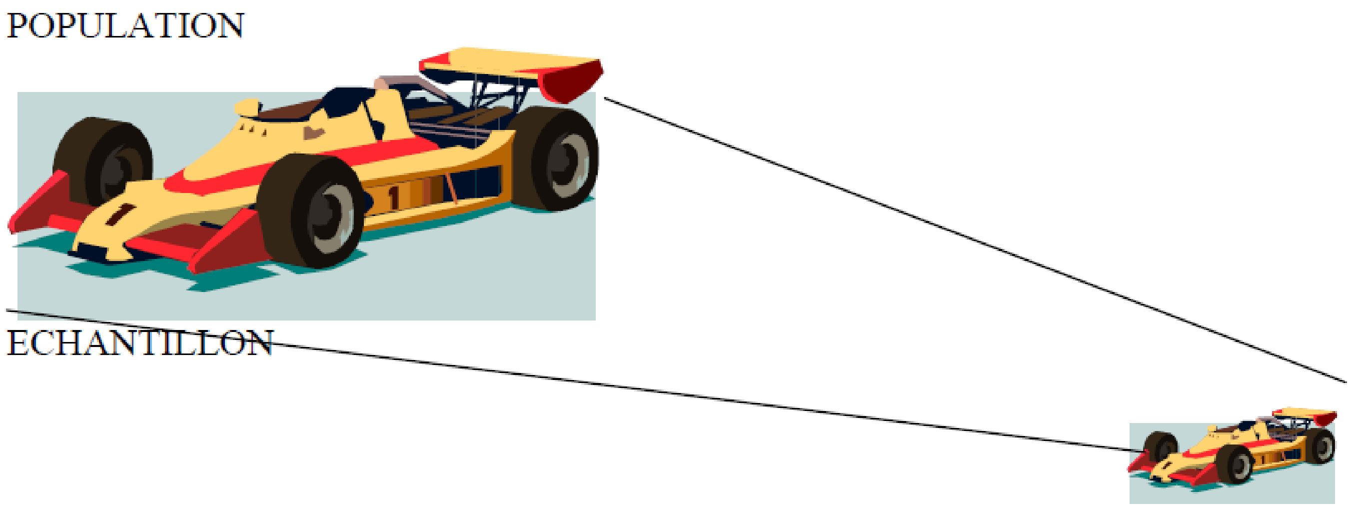


Echantillonnage non aléatoire (2)

Méthode des quotas

- Construction d’un échantillon qui soit une maquette, un modèle réduit de la population étudiée selon certaines caractéristiques, et en conservant les mêmes proportions
- Elle est beaucoup utilisée car plus rapide et moins couteuse que les méthodes alétaoires
- Mais elle est moins fiable.

<u>Ex:</u> Supposons que notre population soit composée de trois profils : il y a 30% des d’étudiants, 20% de travailleurs de la fonction publique et 50% de commerçants. Le constructeur de l’échantillon choisit de manière arbitraire 10 personnes en conservant la même répartition. Il prend donc 3 d’étudiants, car cela correspond à 30% de 10, puis 2 travailleurs de la fonction publique et enfin 5 commerçants. Le choix des unités de réponses (individu) peut donc être laissé au bon vouloir de l’enquêteur dès lors qu’il respecte dans son échantillon la répartition connue de la population.

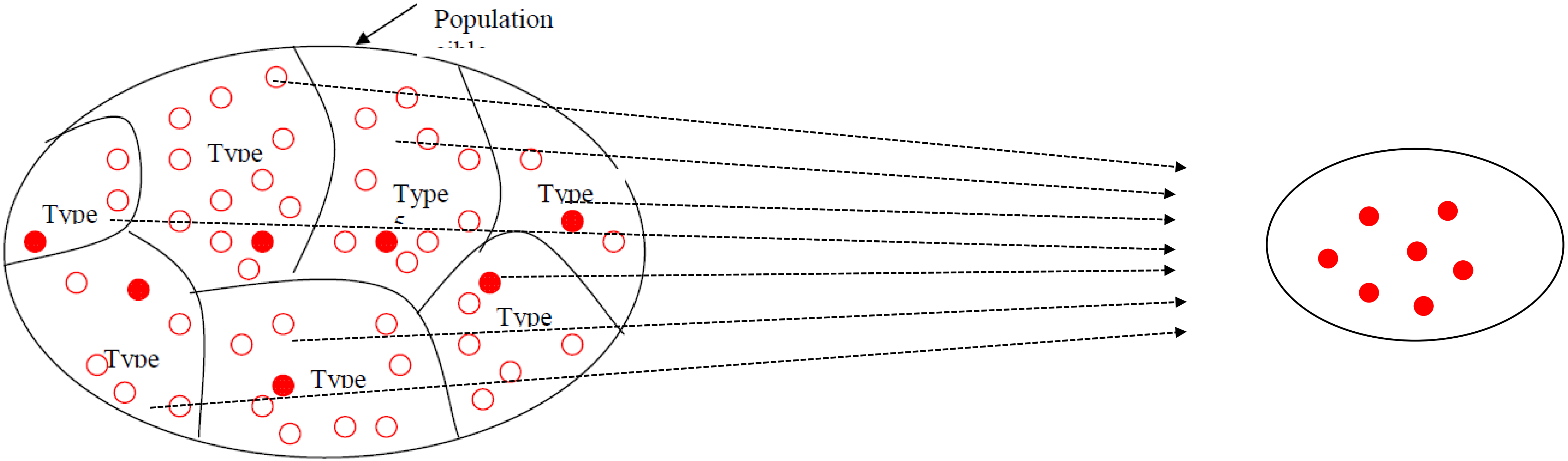


Méthode des unités types

Philosophie : *les différentes variables attachées à un individu n'étant pas indépendantes entre elles, un individu se situant dans la moyenne pour un certain nombre de caractères importants, le sera pour les autres.*

- Etape 1. Diviser la population en groupes homogène (typologie) à partir d'informations connues
- Etape 2. Sélectionner un sujet dans chaque type ou groupe prédéfini = unité type

Méthode souvent utilisée pour les enquêtes avec peu de sujets, et avec l'entretion comme type de



II.2. Echantillonnage aléatoire

Echantillonnage aléatoire (1)

Echantillonnage aléatoire simple (EAS)

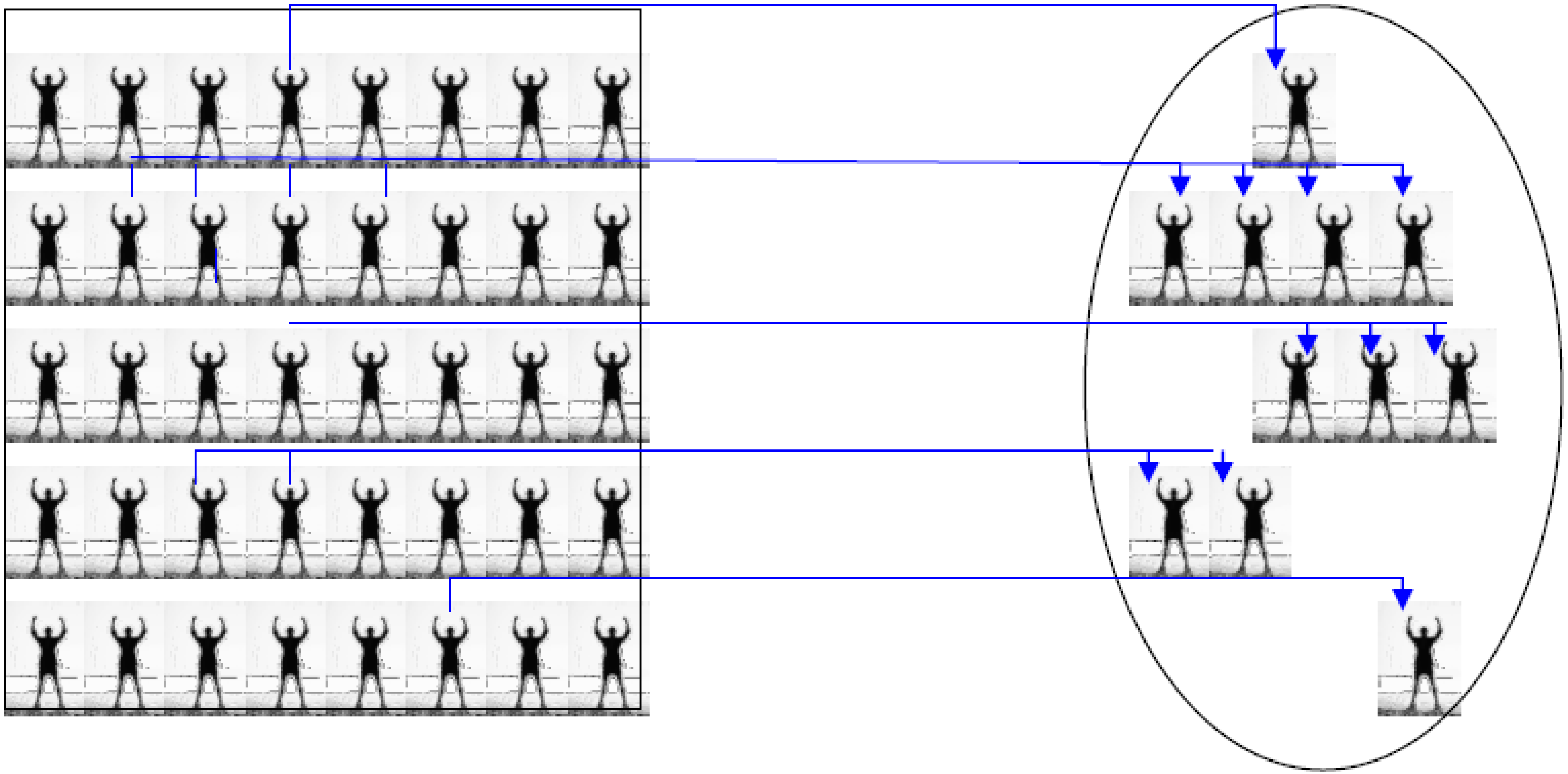
- Consiste à choisir des individus de telle sorte que chaque membre de la population a une chance égale de figurer dans l'échantillon.
- Ce choix peut se faire avec remise ou sans remise :
  - **Avec remise,** un individu peut être choisi plusieurs fois
  - **Sans remise,** un individu déjà choisi ne peut l'être de nouveau. C'est le cas habituel.
- **Avantage de cette méthode** : On peut espérer un échantillon «représentatif » puisque la méthode donne à chaque individu de la population une chance égale. Cependant ceci n'est pas systématiquement vérifié.
- **Difficultés :** la méthode n'est applicable que lorsqu'il existe une liste exhaustive de toute la population appelée **base de sondage ou base d'échantillonnage**

<u>Ex:</u> Soit une enquête ayant pour objectif de décrire l'état des connaissances vis-à-vis du VIH Sida des étudiants en Médecine de Ouahigouya. Le nombre d'étudiants N est de 10 000, donc très élevé et difficile à enquêter, On souhaite sélectionner un échantillon de taille n = 1000.

L'EAS consistera à tirer au sort 1000 étudiants parmi 10 000. On peut utiliser à cet effet plusieurs outils dont une table de nombres aléatoires, ou la fonction alea dans excel.

Sondage aléatoire simple

Population → **Echantillon**



Echantillonnage aléatoire (2)

Echantillonnage systématique

- Nécessite aussi une **base de sondage** (=liste exhaustive de la population où chaque individu est noté de 1 à N)
- Notons n la taille de l'échantillon. Le nombre entier voisin de N/n sera noté **r** et appelé **raison de sondage ou pas de sondage**
- Pour constituer l'échantillon il faut :
  - Choisir au hasard un entier naturel **d** entre 1 et r (cet entier sera le point de départ),
    - L'individu dont le numéro correspond à d est le premier individu
  - Pour sélectionner les autres, il suffit **d'ajouter à d le pas de sondage** : les individus choisis seront alors ceux dont les numéros correspondent à
    - d + r
    - d + 2r
    - d + 3r

    etc.

Echantillonnage aléatoire (3)

Echantillonnage stratifié

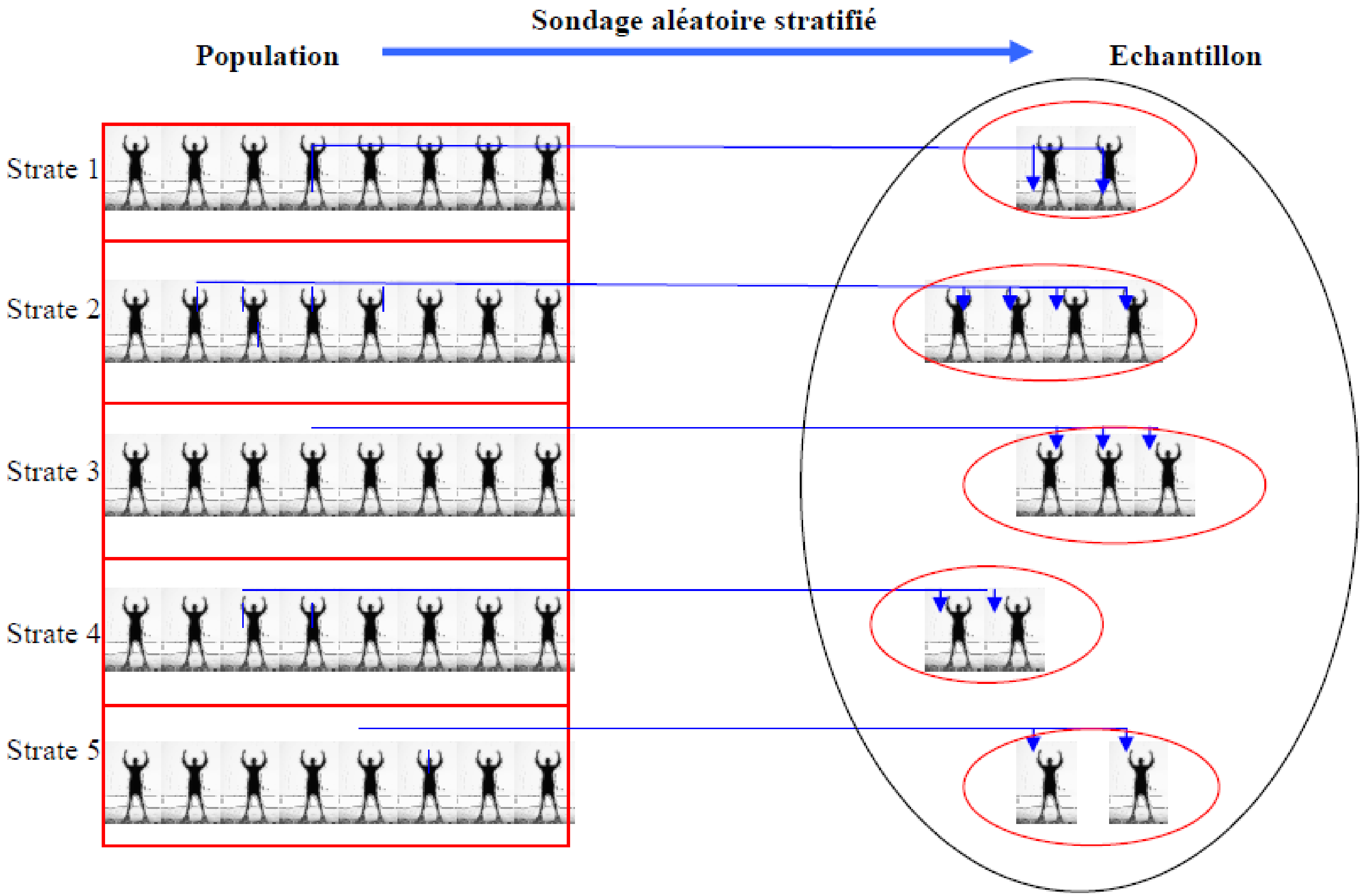
- Chaque **strate** est considérée comme une sous-population d'individus comparables vis-à-vis de la variable de stratification. On dit qu'une strate est une section transversale de la population, c'est-à-dire un **sous-groupe homogène**.
- Démarche de sélection :
  1. On subdivise la population en strates qui sont mutuellement exclusives
  2. Proportionnellement à son importance dans la population, on calcule combien il faut d'individus au sein de l'échantillon pour représenter chaque strate.
  3. Dans chacune des strates, on choisit au hasard le nombre nécessaire d'individus
- Les variables de starification doivent être :
  - Simple à utiliser (Ex: Age, sexe, niveau scolaire etc.)
  - Facile à observer
  - Étroitement reliées au thème de l'enquête

Echantillonnage aléatoire (4)

Echantillonnage stratifié (2)

- **Avantages :** Il est peu probable de choisir un échantillon absurde puisqu'on s'assure de la présence proportionnelle de tous les divers sous-groupes composant la population.
- **Désavantages :** La méthode suppose l'existence d'une base de sondage. Il faut aussi connaître comment cette population se répartit selon certaines strates.

Exemple : choisir par échantillonnage stratifié 10 étudiants dans un groupe de 60, en tenant compte du fait que 50% d'entre eux sont en PCEM1, 30% en PCEM2 et 20% en DCEM1.



Echantillonnage aléatoire (5)

Echantillonnage en grappes

- Dans les méthodes précédentes, l'unité statistique était choisie individuellement.
- La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou grappes. On appelle **grappe** un sous-ensemble de la population source.
- On sélectionne au hasard un certain nombre de grappes (unités primaires) pour représenter la population.
- **On sélectionne tous les individus des grappes choisies**

Attention:

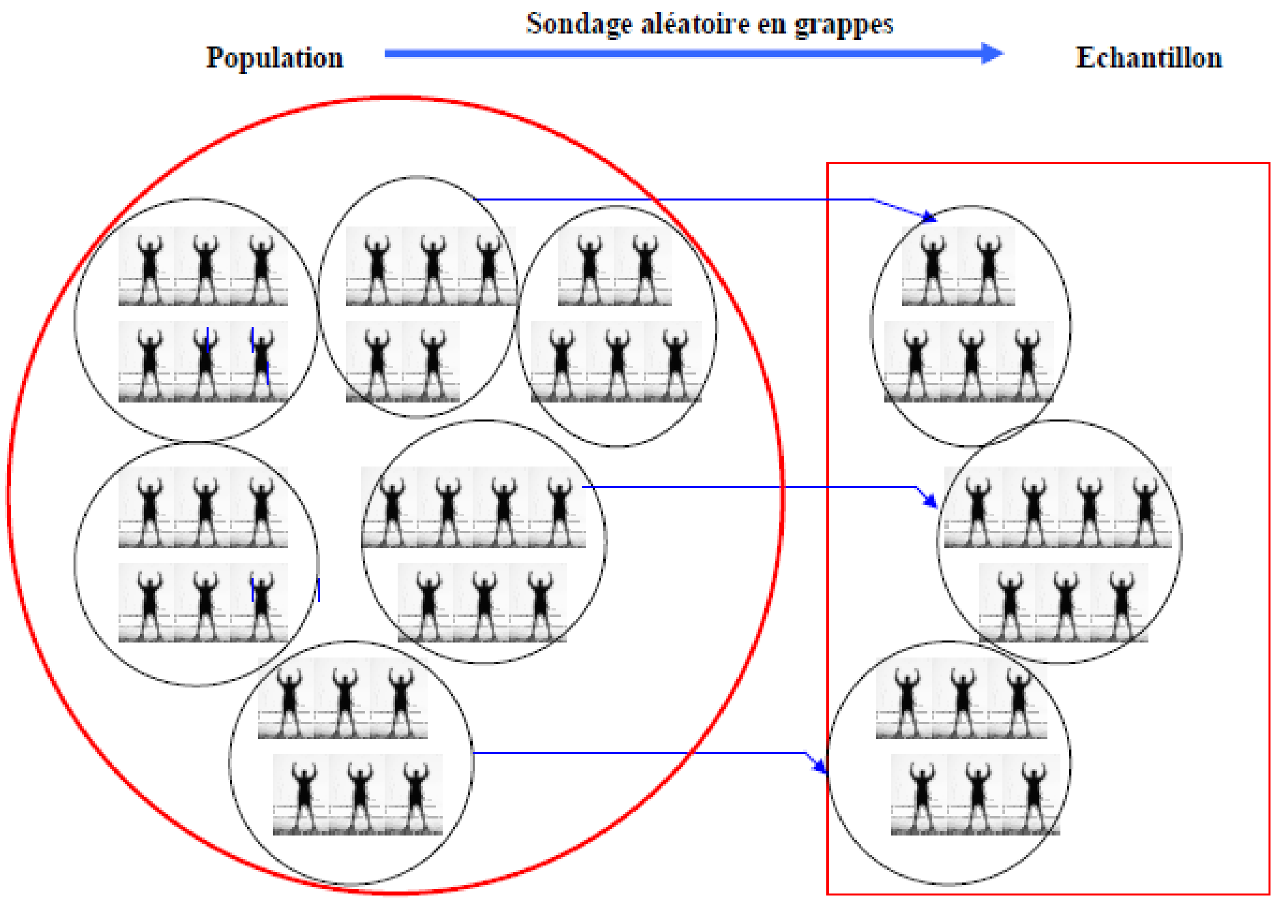
La composition de ces grappes est antérieure au plan de sondage et **n'a pas été réalisée *a priori* par rapport à une caractéristique** corrélée au critère d'intérêt contrairement à ce qui prévaut souvent lors de la construction des strates.

Echantillonnage aléatoire (6)

Echantillonnage en grappe (2)

- **Avantages** : la méthode ne nécessite pas une base de sondage puisque seules les individus inclus dans les grappes comptent. Elle permet de limiter l'échantillon à des groupes compacts ce qui permet de réduire les coûts de déplacement, de suivi et de supervision.
- **Désavantage** : la méthode peut entraîner des résultats imprécis (moins précis que les méthodes précédentes) puisque les unités voisines ont tendance se rassembler. Elle ne permet pas de contrôler la taille finale de l'échantillon. « Qui se ressemblent s'assemblent)

Exemple : Choisir par grappes 600 individus à l'aide d'un certain nombre de ménages.



Echantillonnage aléatoire (7)

Echantillonnage à plusieurs degrés

- Ressemble à l'échantillonnage en grappes, sauf que dans ce cas on prélève un échantillon à l'intérieur de chaque grappe
- On a au moins 2 degrés
  - On identifie au premier degré les grandes grappes (unités primaires). Ces grappes renferment plus d'unités qu'il n'en faut dans l'échantillon
  - Au second degré, à l'intérieur de chaque grappes, on sélectionne les unités (unités secondaires) qui vont faire partie de l'échantillon
- On peut utiliser plus de 2 degrés :
  - Niveau 1 : Ville
  - Niveau 1 : Établissement de santé
  - Niveau 3 : Médecins

Echantillonnage aléatoire (8)

Echantillonnage à plusieurs degrés(2)

Avantage :

Échantillon plus concentré ce qui réduit les coûts, pas besoin de disposer de la liste de toutes les unités. La méthode permet de contrôler la taille de l'échantillon notamment par stratification.

Désavantage :

précision des résultats

Dans certaines situations, l'accès à la liste des unités statistiques composant la population d'étude n'est pas possible ; En revanche, on peut disposer d'un répertoire de groupes d'individus.

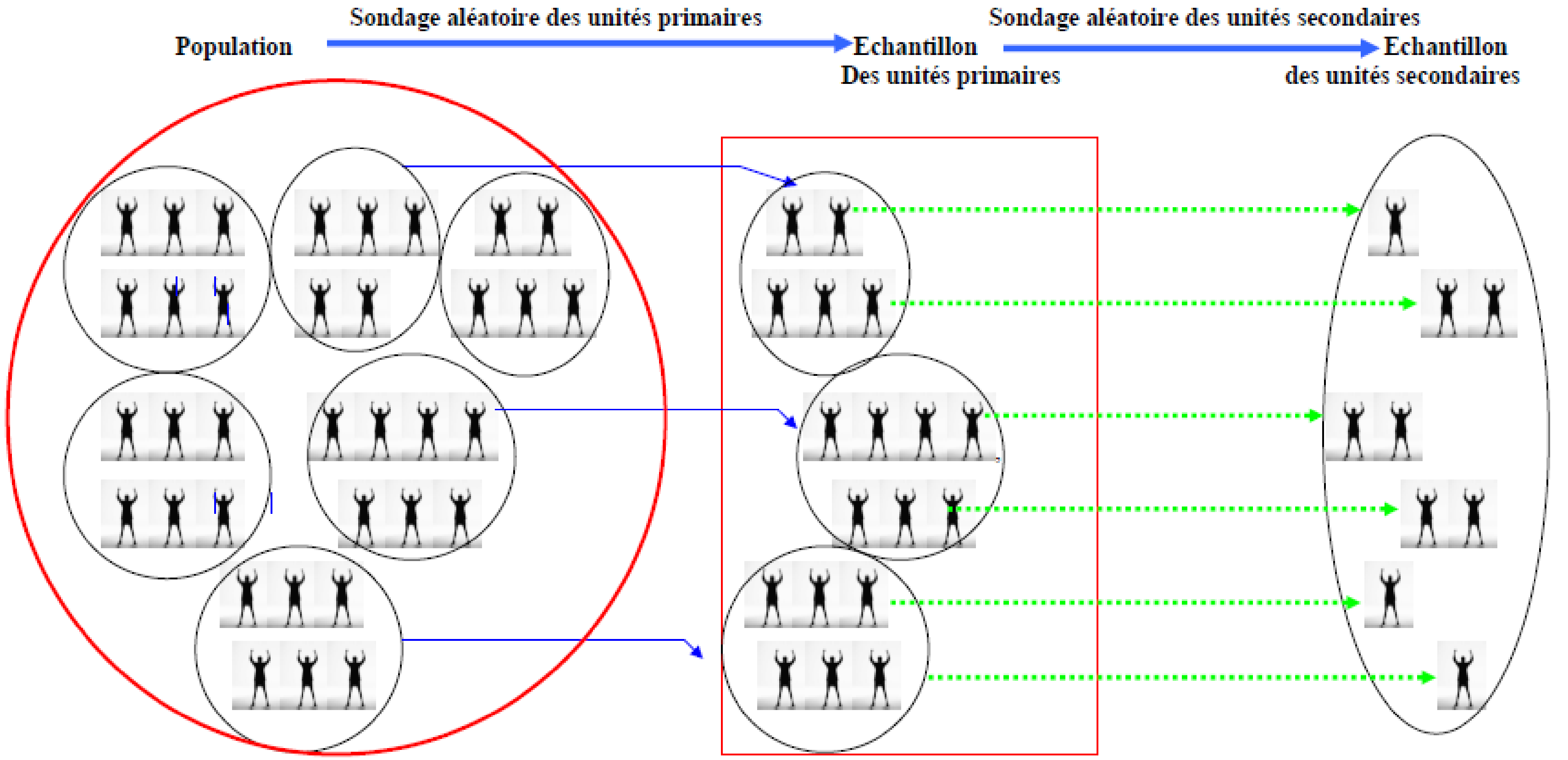


Figure 1 : constitution d'un échantillon à partir d'une population lors d'un sondage aléatoire à deux degrés

III. Estimations

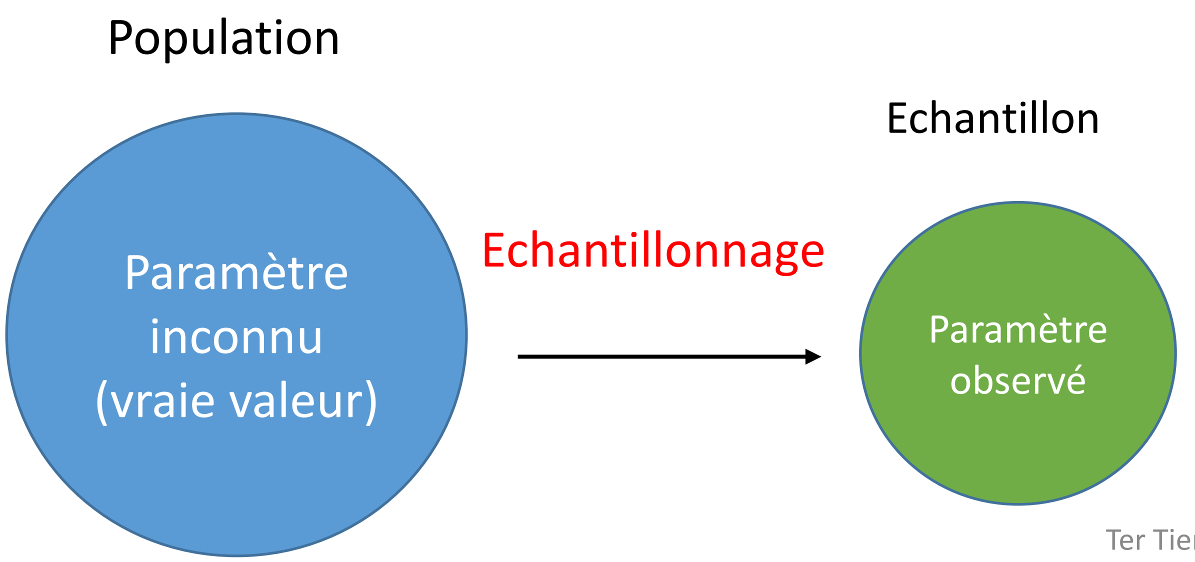
Définitions (1)

Estimation

- Définir ou calculer les paramètres d'une population (inconnus) à partir des paramètres observés sur un échantillon.
- Le paramètre peut être grandeur quantitative (moyenne, écart-type) ou qualitative (proportion).
- L'estimation est **ponctuelle** et **par intervalles** (à cause des fluctuations d'échantillonnage)

Estimateur

- Formule mathématique qui permet de calculer une estimation (paramètre estimé)



Notations

	Population P (valeur théorique	Echantillon E (v. observée)
Variable quantitative		
Moyenne	M ou μ	m ou $\hat{\mu}$
Variance	S^2 ou σ^2	s^2 ou $\hat{\sigma}^2$
Ecart-type	S ou σ	S ou $\hat{\sigma}$
Variable qualitative		
Fréquence	P ou π	p ou f

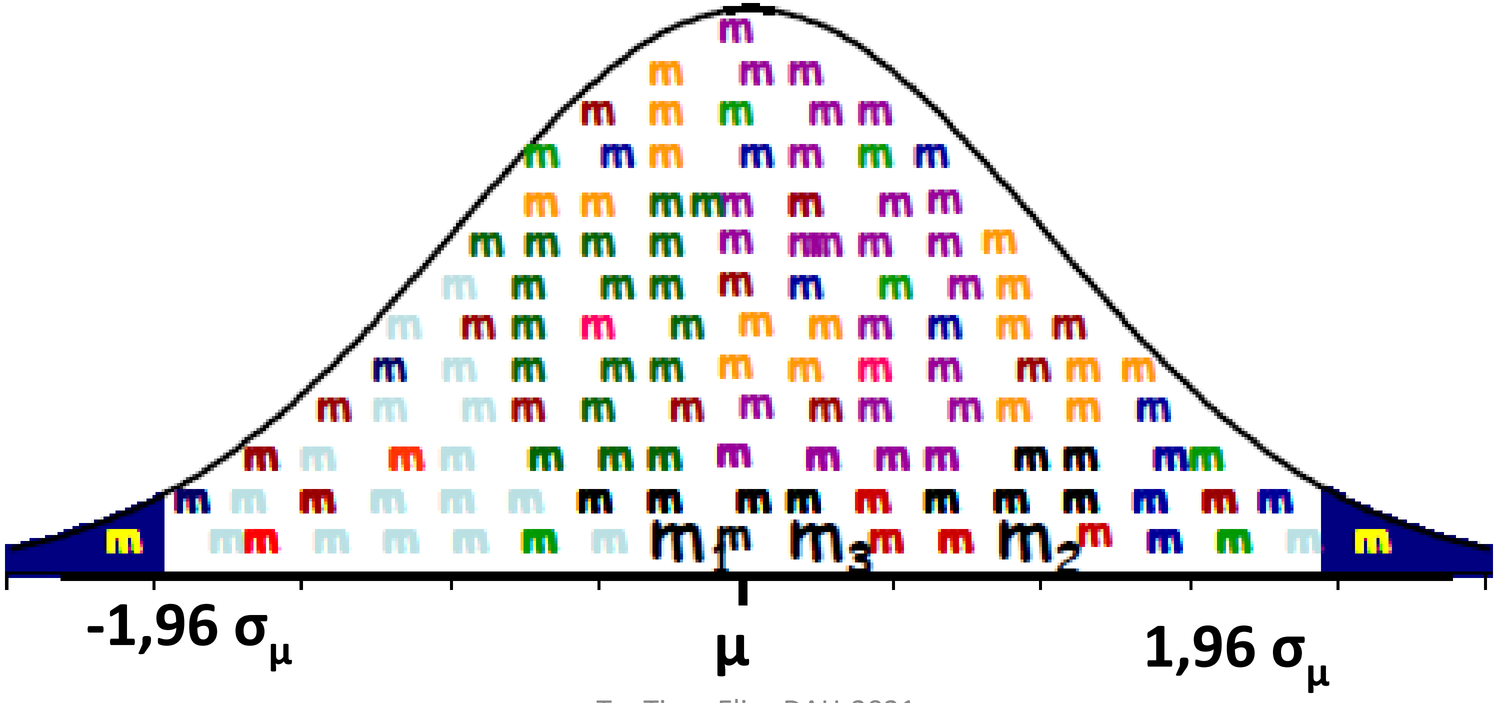
Estimation de la moyenne & de la variance (1)

Comment estimer μ en connaissant m ?

Une moyenne observée dans un échantillon

- Est une variable aléatoire qui suit une loi normale
- Centrée sur la moyenne μ et d'écart-type σ_μ

Si α = 5%, alors 95% des valeurs de m sont comprises entre -1,96 et +1,96 écart type de μ



Estimation de la moyenne & de la variance (2)

- **Estimation ponctuelle**

$$m = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{et} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$\boldsymbol{\sigma} = \sqrt{\sigma^2} \simeq \frac{s}{\sqrt{n}}$$

- **Estimation par intervalle**

Si $\alpha = 5\%$, dans 95% des cas $\mu - 1{,}96\ \sigma_\mu < m < \mu + 1{,}96\ \sigma_\mu$

On peut aussi écrire $\mu < m + 1{,}96\ \sigma_\mu$ et $\mu > m - 1{,}96\ \sigma_\mu$

Intervalle de confiance à 95% de la moyenne (IC 95%): $m - 1{,}96 \frac{s}{\sqrt{n}} < \mu < m + 1{,}96 \frac{s}{\sqrt{n}}$

n: taille de l'échantillon

m : moyenne de l'échantillon

s : écart-type de l'échantillon

Estimation de la moyenne & de la variance (3)

- **Estimation par intervalle**

  Intervalle de confiance d'une moyenne au risque 1-α%

$$\mathbf{m - U_{\alpha} \frac{s}{\sqrt{n}} < \mu < m + U_{\alpha} \frac{s}{\sqrt{n}}}$$

$$\mathbf{IC_{1\text{-}\alpha\%}\ \mu = m \pm U_{\alpha} \frac{s}{\sqrt{n}}}$$

Précision Δ

α = erreur fixée

U_{α} lue dans la table de la loi normale

- Conditions d'application de la formule de calcul
  - Taille de l'échantillon grande, c'est à dire n > 30
  - Taille de l'échantillon n négligeable par rapport a la taille de la population N (n/N < 10%)
  - Ou X est N (μ,σ) dans P

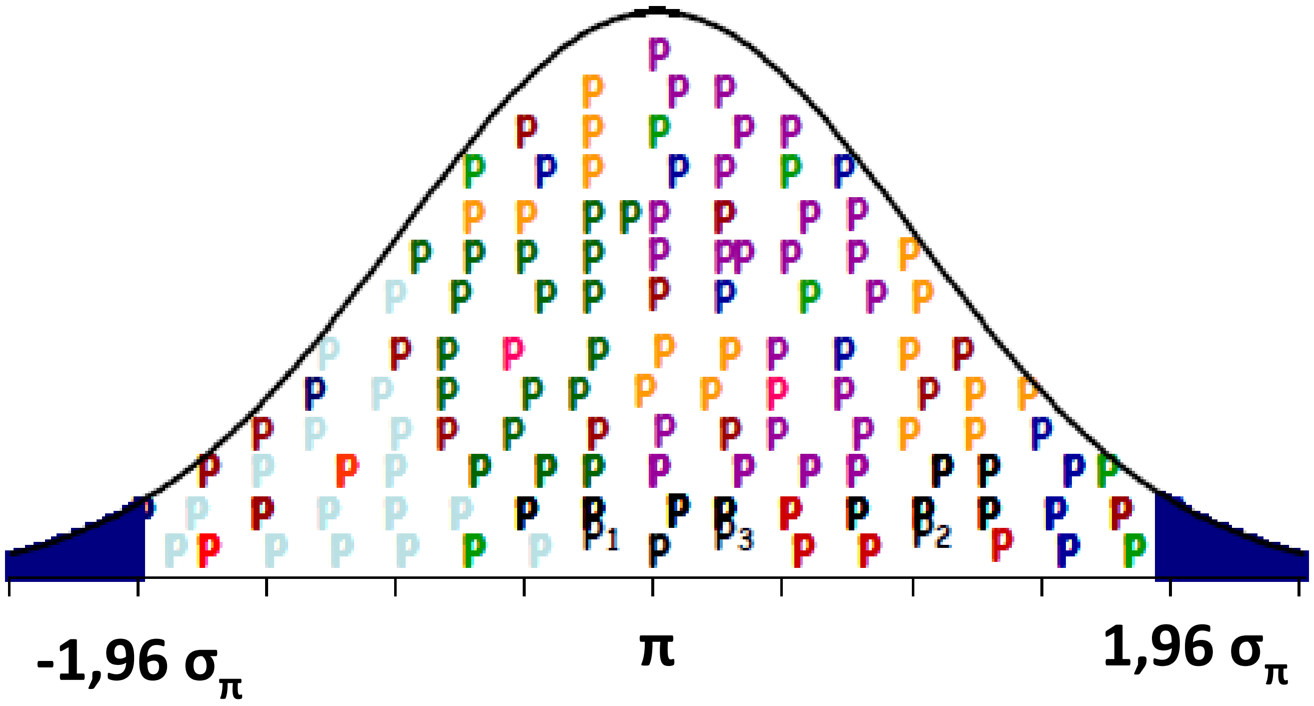
Estimation d'une proportion ou un pourcentage (1)

Comment estimer π en connaissant p ?

Une proportion observée sur un échantillon de taille n

- Est une variable aléatoire qui suit une loi normale
- Centrée sur la proportion π et d'écart-type σ_π

Si α = 5%, alors 95% des valeurs de p sont comprises entre -1,96 et +1,96 écart type de π



Estimation d'une proportion ou d'un pourcentage (2)

- **Estimation ponctuelle**

Soit une population dans laquelle une proportion π d'individus possède un caractère qualitatif (par exemple **proportion π de sujets malades**), un estimateur de π est donné par la variable proportion d'échantillonnage :

$$p = P = \sum X_i / n$$

ce qui correspond à la proportion d'individus possédant le caractère dans un échantillon de taille n.

Espérance mathématique de (p) = E(p) = E(P) = p

Variance (p) = σ^2= p (1-p) / n

Ecart-type (p) = σ_p = $\sqrt{\frac{p\,(1-p)}{n}}$

Estimation d'une proportion ou d'un pourcentage (3)

- **Estimation par intervalle**

  Si $\alpha = 5\%$, dans 95% des cas $\pi - 1{,}96\ \sigma_p < p < \pi + 1{,}96\ \sigma_p$

  On peut aussi écrire $\pi < p + 1{,}96\ \sigma_\mu$ et $\pi > p - 1{,}96\ \sigma_\mu$

IC 95% d'une proportion : $p - 1{,}96\sqrt{\dfrac{\mathbf{p\ (1-p)}}{\boldsymbol{n}}} < \pi < p + 1{,}96\sqrt{\dfrac{\mathbf{p\ (1-p)}}{\boldsymbol{n}}}$

n: taille de l'échantillon

p : fréquence (proportion) dans l'échantillon

Estimation d'une proportion ou d'un pourcentage (3)

- **Estimation par intervalle**

  Intervalle de confiance d'une proportion au risque 1-α%

$$\mathbf{p - U_\alpha\sqrt{\frac{p\,(1-p)}{n}} < \pi < p + U_\alpha\sqrt{\frac{p\,(1-p)}{n}}}$$

ou

$$\mathbf{IC_{1\text{-}\alpha\%}\ \pi = [b_{inf}\ p\ ;\ b_{sup}\ p] = p \pm U_\alpha\sqrt{\frac{p\,(1-p)}{n}}}$$

Précision Δ

α erreur fixée

U_α lue dans la table de la loi normale

- **Conditions d'application de la formule de calcul de l'IC d'un pourcentage**
  - Taille de l'échantillon n grande, vérifiée à priori par : **$np \geq 5$ et $n(1-p) \geq 5$**
    - Vérifier aussi les conditions à posteriori après le calcul des bornes inf & sup de l'I
      - **$nb_{inf} \geq 5$, $(n\ 1- b_{inf}) \geq 5$, $nb_{sup} \geq 5$, $(n\ 1- b_{sup}) \geq 5$**
  - Taille de l'échantillon n négligeable par rapport a la taille de la population N (n/N < 10%)

III. Calcul de la taille d’un échantillon

Calcul de la taille n d'un échantillon (1)

Dépend

- Objectif principal de l'étude
  - Question de recherche, hypothèse
    - Paramètre que l'on veut estimer (épidémiologie descriptive)
    - Association ou différence que l'on veut montrer (épidémiologies analytique & évaluative)
- Type d'étude
- Type d'échantillonnage
- Précision (Erreur α, dispersion)
- Loi de distribution de la variable
- % de refus (de participer et/ou de répondre)
- Puissance statistique (1-β)
- % de perdus de vue

Essentiellement pour les objectifs d'épidémiologie
- Analytique
- évaluative

Nous verrons uniquement le calcul d'échantillon pour les études (épidémiologie) descriptives

Calcul de la taille d'un échantillon (2)

Rappel

- L'intervalle de confiance d'un paramètre estimé (moyenne ou proportion) a la forme suivante: **[paramètre ± Δ].** Δ est appelée la **précision,** et correspond:
  - $\Delta = U_\alpha \frac{s}{\sqrt{n}}$ si estimation d'une moyenne
  - $\Delta = U_\alpha \sqrt{\frac{p\,(1-p)}{n}}$ si estimation d'une proportion
- De cette formule, on peut déterminer **la taille n ou (nombre de sujets nécessaire**) à inclure dans l'échantillon pour obtenir une précision souhaitée Δ dans l'estimation du paramètre dans la population.

$$\boxed{n = \frac{(U_\alpha)^2\, s^2}{\Delta^2}}$$

pour une moyenne pour une moyenne

$$\boxed{n = \frac{(U_\alpha)^2\, p(1-p)}{\Delta^2}}$$

pour une moyenne pour une proportion

Les valeurs de p et de s^2 sont à rechercher dans la littérature.

Si cela n'est pas disponible, considérer p = 0,5 ou 50 %, et s^2 = variance attendue que l'on souhaite

Résumé

- **Estimation, précision et taille d'échantillon**

Dans l'estimation d'un paramètre (moyenne ou proportion), la dispersion (écart-type ou intervalle de confiance (IC)) vaut 2 fois la précision. Ex: IC = 2 X Δ.

Par ailleurs, on sait que plus la dispersion est petite, plus l'estimation est précise.

De même, on sait aussi que plus α est petit, plus Uα est grand

De ce fait, la précision d'une estimation est liée aux facteurs suivante :

- la taille n de l'échantillon
- l'erreur α

L'épidémiologiste devra donc jouer sur ses 2 facteurs pour assurer une bonne estimation du paramètre observé : une erreur α acceptable et une taille d'échantillon grande

$$\mathbf{\Delta = U_{\alpha}\ \frac{s}{\sqrt{n}}}\ ;\ \text{Moyenne}$$

$$\mathbf{\Delta = U_{\alpha}\sqrt{\frac{p\,(1-p)}{n}}}\ ;\ \text{Proportion}$$

α	U_{α}
20%	1,28
10%	1,65
5%	1,96
2%	2,33
1%	2,58

Pour aller loin…

2 liens utiles pour le calcul de taille d’un échantillon.

https://biostatgv.sentiweb.fr/

https://www.openepi.com/Menu/OE_Menu.htm

Merci pour votre attention

Documents ressources

1. Jean Bouyer. Méthodes statistiques. Médecine-Biologie. 2017.
2. Thierry Ancelle. Statistique-Epidémiologie, 4ème édition. 2017.
3. François Dabis, Jean Claude Desenclos. Epidémiologie de terrain. Méthodes et applications 2017
4. Simone Benazeth, Michel Boniface, Catherine Demarquilly, Virginie Lasserre, Mohamed Lemdani, Ioannis Nicolis. Biomathémathiques. Analyses, algèbre, probabilités, statistiques. Pharmacie, Médecine 1ère & 2ème années, Masson. 2001