DE LA RECHERCHE À L'INDUSTRIE

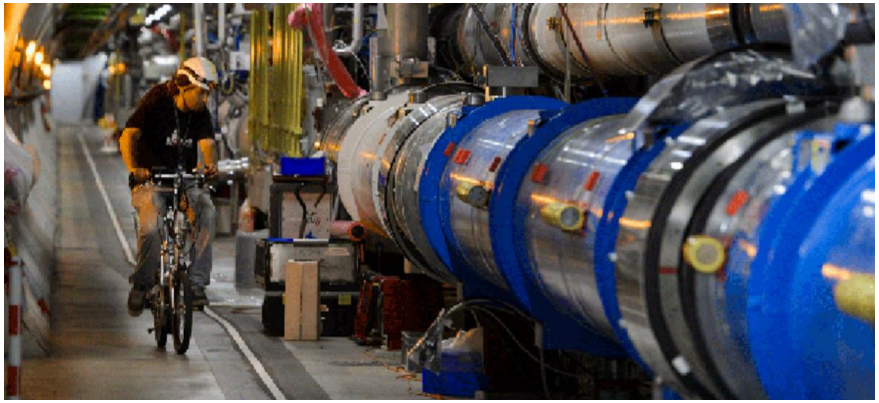
INTRODUCTION AUX PLANS D'ÉCHANTILLONNAGE

www.cea.fr
www-instn.cea.fr

Introduction

ANALYSER SON ENVIRONNEMENT : 2 APPROCHES

1/ Fixation des paramètres – mesure de la réponse



2/ Collecte des paramètres – mesure de la réponse

FIXER SES PARAMÈTRES : PLANS D'EXPÉRIENCE

1/ Réglage et contrôle des paramètres – mesure de la réponse

Comportement : « actif »
notion de programme, de procédé, …

Lieu : le laboratoire, l'usine

Méthode : synthèse, production

Outil : projet, **plan d'expérience** – expérience - mesure

2/ Collecte et enregistrement des paramètres – mesure de la réponse

Comportement « passif »
notion de surveillance, de classification/clustering, ,,,,

Lieu : l'environnement au sens large (populations importantes)

Méthode : observation et/ou analyse destructive,

Outil : **plan d'échantillonnage –** prélèvements - mesure

PRINCIPAUX DOMAINES D'UTILISATION

Avec un minimum de données à analyser :

- Suivre l'évolution d'une population

  Détecter un événement, une structure ou une propriété particulière (rare, inattendue, intéressante ou indésirable)

  Surveillance, contrôle de matière première, pollution, crue, défaillances, …

- Caractériser une population (état des lieux)

  Sa taille, son étendue, sa densité

- Distribution ou Répartition des individus dans un milieu suivant leurs caractéristiques

  répartition des étoiles, étendue d'une maladie

POURQUOI ÉCHANTILLONNER ?

Il est impossible de tout mesurer …. et de tous surveiller
Le plan d'échantillonnage doit permettre à partir d'**un nombre minimal de mesures** de :

1. Déterminer des relations qui peuvent exister entre les différents paramètres et de mettre en évidence les paramètres les plus importants

2. D'inventorier les individus suivant leur propriétés et avec leur effectif. Les ensembles ainsi constitués peuvent être des compartiments qui seront utilisés dans une modélisation

3. Mettre en évidence des structures et de constituer des cartes (voir méthodes de Krigeage)

Ce nombre minimal de mesure est souvent obtenu par des **opérations de regroupement ou de tri conditionnel** sur les unités de la population étudiée

NOTIONS PRINCIPALES : TECHNIQUES DE RÉDUCTION

La population (absence/présence) avec son effectif

- Un ensemble de nœud et d'arrêtes pour un graphe
- Un ensemble de parcelles pour une carte
- Un ensemble de durées pour une durée
- Un ensemble de personnes pour une population humaine

Pour faciliter sa compréhension cette **population est regroupée en sous-ensembles appelées strates** suivant leurs propriétés. Plus le nombre de strates est élevé, plus leur effectif est faible

- Regroupement en importance de connexion pour les nœuds d'un graphe
- Regroupement en écosystèmes pour une carte
- Regroupement en saisons pour des longues durées
- Regroupement en habitats pour une population humaine

Mais chacun de ces regroupements peut, à son tour faire l'objet de nouveaux **regroupement successifs donnant plusieurs niveaux ou degré** au plan. Plus le niveau est élevé, plus il y a de subdivision, plus l'effectif ce ces subdivisions est faible.

- Division des connexions suivant leur portée
- Division de la forêt en hêtraies, chênaies puis suivant l'âge des arbres,
- Division en mois puis en jours
- Division en villes puis quartiers puis pavillons

PRINCIPAUX PLANS D'ÉCHANTILLONNAGE

Plan non structuré dans sa totalité : c'est le plan aléatoire

On pourra cibler sur des unités ou des éléments d'un graphe (arrêtes et/ou nœud)

Plan grossièrement structuré par classes (ce qui permet de réduire la totalité à un ensemble de sous groupes)

- Classes de même taille : plan systématique
- Classe d'unités aux propriétés semblables : plan stratifié (une classe = une strate)
- Classe d'unités aux propriétés rangées ou ordonnées : plan multiniveaux (une classe = un niveau). Un niveau pouvant être lui-même constitué de plusieurs strates,

→Pratiquement un certains nombre de classes sont sélectionnées aléatoirement, puis une autre sélection est effectuée sur les unités des classes qui ont été sélectionnées Chaque étape de sélection aléatoire est un degré

Les plans progressifs qui sont une sélection itérative jusqu'à réalisation d'un critère

NOTIONS PRINCIPALES : TECHNIQUES DE SÉLECTION

Les deux techniques de base sont le tirage aléatoire ou le recensement

La taille du tirage aléatoire peut être proportionnel à l'effectif des classes (strates ou niveaux). Une optimisation de la taille de l'échantillon est obtenue par **allocation**,

Attention, certains plans dits d'échantillonnage sont en fait des sélections particulières :

- Le plan mélange ou composite est un plan où tous les échantillons prélevés sont mélangés. Suivant l'efficacité du mélange, cela revient à constituer un plan aléatoire empirique.

- Le plan en grappe est en fait un plan subdivisé (en strates ou niveaux) pour lequel toutes les unités des subdivisions aléatoirement sélectionnées sont analysées : autrement dit, un recensement y est effectué.

- Le plan en « boule de neige » qui est l'analogue du plan en grappe mais pour un réseau ou graphe (étude des interactions).

CHRONOLOGIE « TYPE » DE LA RÉALISATION

1. Définir son **cadre** : objectifs, population, milieu, propriétés

2. Identifier les **contraintes** : normes, budgets, effectif, temps, accès, cout), … et en déduire la taille de l'échantillon

    3. Rechercher des **informations** déjà existantes (archives, expert, pré-échantillonnage)

    4. **Pré**-échantillonnage et regroupement/clustering

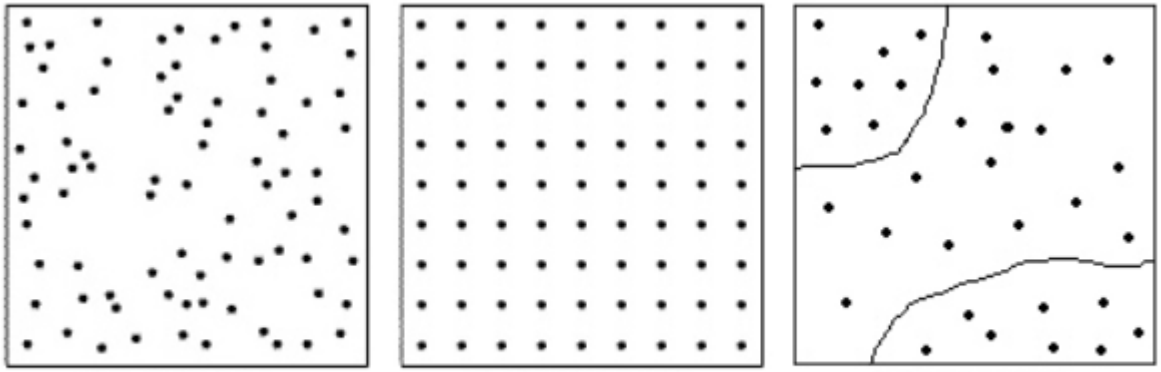
5. Choix du plan d'échantillonnage

1. Méthode de **collecte**, taille, allocation

2. Évaluation de l'efficacité du plan d'échantillonnage D_{eff}

PROPOSITION DE SYNTHÈSE

Sélection / Réduction	Aléatoire	Recensement	Ponctuel	Jugement
Aucune	Plan aléatoire			« au pif »
Par division	Plan systématique	Plan en grappe		
Par équivalence	Plan stratifié			
Par ordre	Plan multiniveaux		RSS	
Par tri conditionnel	Plan progressifs	Boulle de neige		
Par jugement	Plan jugement			



Population, statistique et modélisation
Classification et regroupement

ECHANTILLONNAGE SANS REMISE

Après sélection, les unités peuvent être :

- réintroduites : **échantillonnage avec remise**
- gardées : **échantillonnage sans remise**

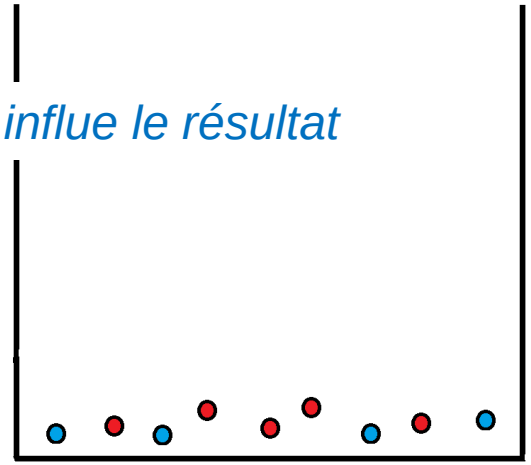
Quand n ~ N l'action d'échantillonnage a un effet sur le résultat

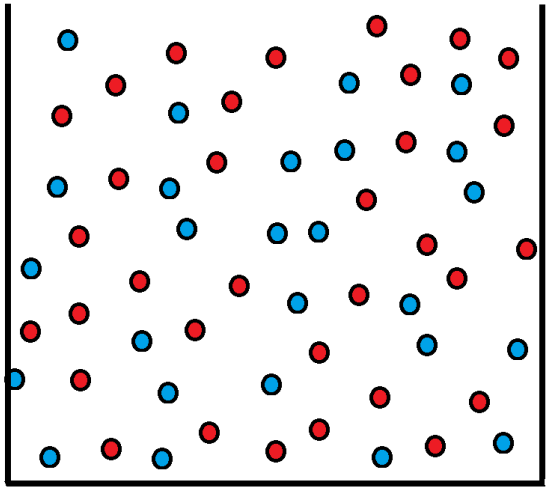
Il en résulte un « effet de bord » qu'il faut corriger dans la probabilité de sélection

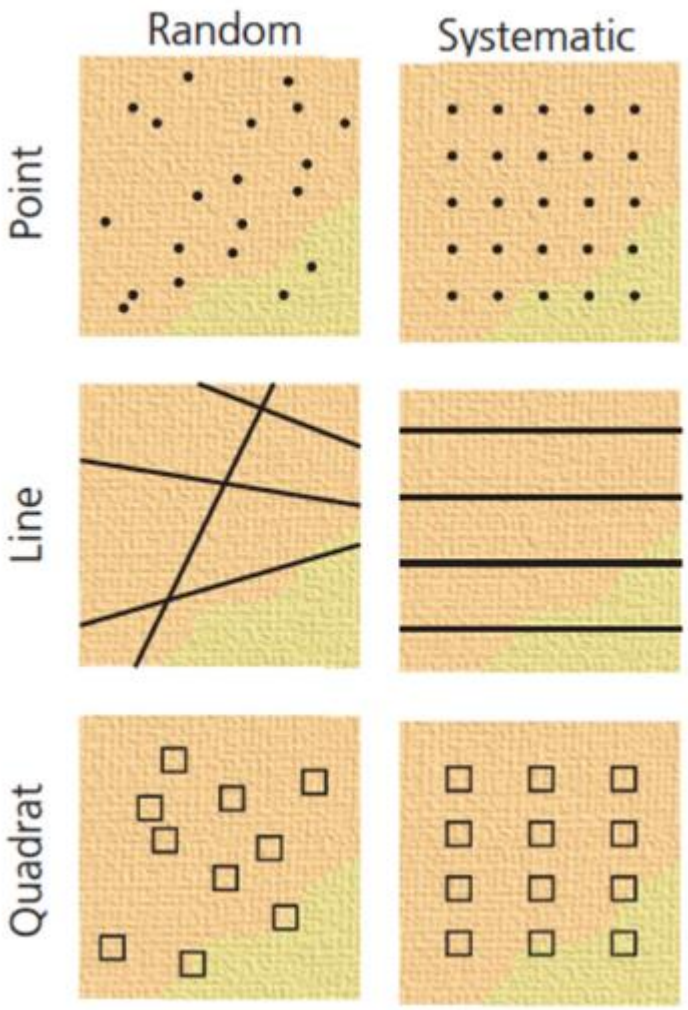
$$\frac{1}{\pi_i} = \frac{1}{C_n^N} = \frac{n!(N-n)!}{N!}$$

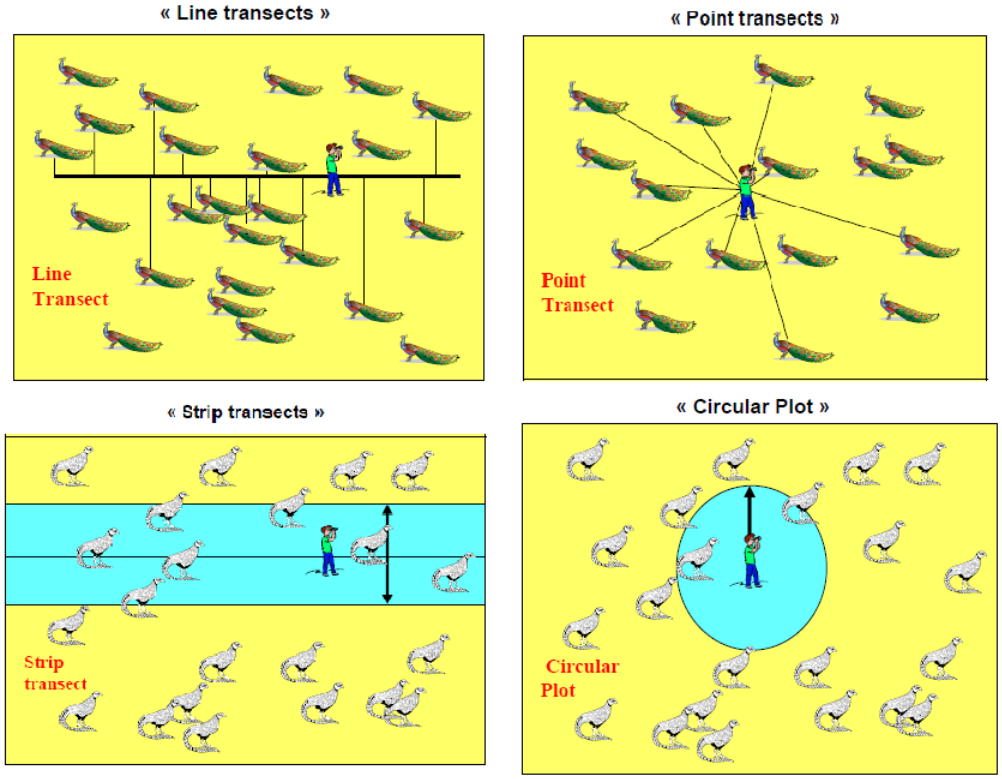
La remise n'influe pas le résultat

La remise influe le résultat






Random
Systematic
Point
Line
Quadrat


« Line transects »
Line
Transect
« Point transects »
Point
Transect
« Strip transects »
Strip
transect
« Circular Plot »
Circular
Plot

RECENSEMENT, PLAN DE JUGEMENT ET AUTRES …

Ensemble des préjugés possibles (**causes** possibles), Idée que l'on se fait

L'Absolu
Les paramètres vrais

Prédiction des valeurs paramètres pour l'ensemble des **effets** attendus

Adaptation
Correction de ses préjugés
Maximum de vraisemblance

Ensemble des résultats (effets) obtenus

Biais

Calcul des paramètres

Estimateurs

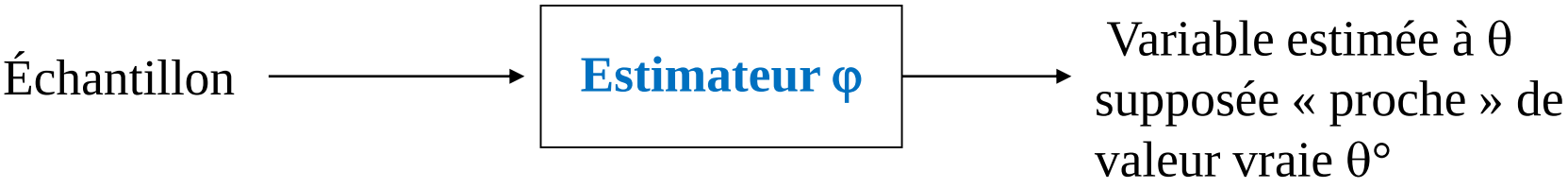
Ensemble des estimations possibles pour les paramètres à déterminer

Echantillonnage & mesure

Population à décrire ou caractériser

PARAMÈTRES À ESTIMER

Estimation la valeur vraie inaccessible



Exemple de paramètres à estimer :

- ***La moyenne***
- ***La population totale T (ou τ)***
- *La variance ou l'écart-type $\sigma^2 = Var(X)=V(X)$*
- ***Les proportion p***
- *Le coefficient de variation CV*

DEUX PARAMÈTRES UTILES POUR DÉCRIRE

Tendance des résultats : la moyenne

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Fluctuation ou dispersion des résultats : l'écart type

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

Cet écart type caractérise la dispersion intrinsèque de la variable

Par la suite cet écart type sera noté σ_{exp}

Et pour avoir un paramètre sans dimension : le Coefficient de Variation CV

$$CV = \frac{\sigma_{exp}}{\bar{x}}$$

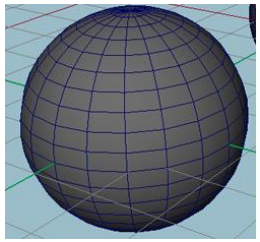
ET POUR ALLER PLUS LOIN : COVARIANCE

Expression général de la variance :

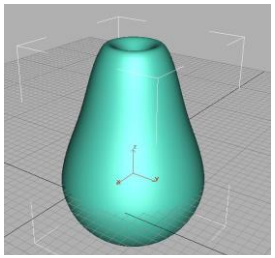
$$Var(Xi + Xj) = Var(Xi) + Var(Xj) + 2.Cov(Xi, Xj)$$

Autre notation $\sigma_{XiYj} = 2.Cov(Xi, Xj)$

La covariance rappel le cosinus, elle estime **interdépendance des variables**



$\sigma_{Xi,Xj} = Cov(X_i, X_j) = 0$



$\sigma_{Xi,Xj} = Cov(X_i, X_j) \neq 0$

On utilisera un outil plus pratique que la covariance ou la corrélation : le variogramme

COVARIANCE ET CORRÉLATION

Covariance : met bien en évidence les interdépendances
Mais elle dépend de l'échelle choisie

⇒ Pour s'affranchir de la dimension, on travaille en relatif (**normalisation**).

Covariance de deux variables x et y :

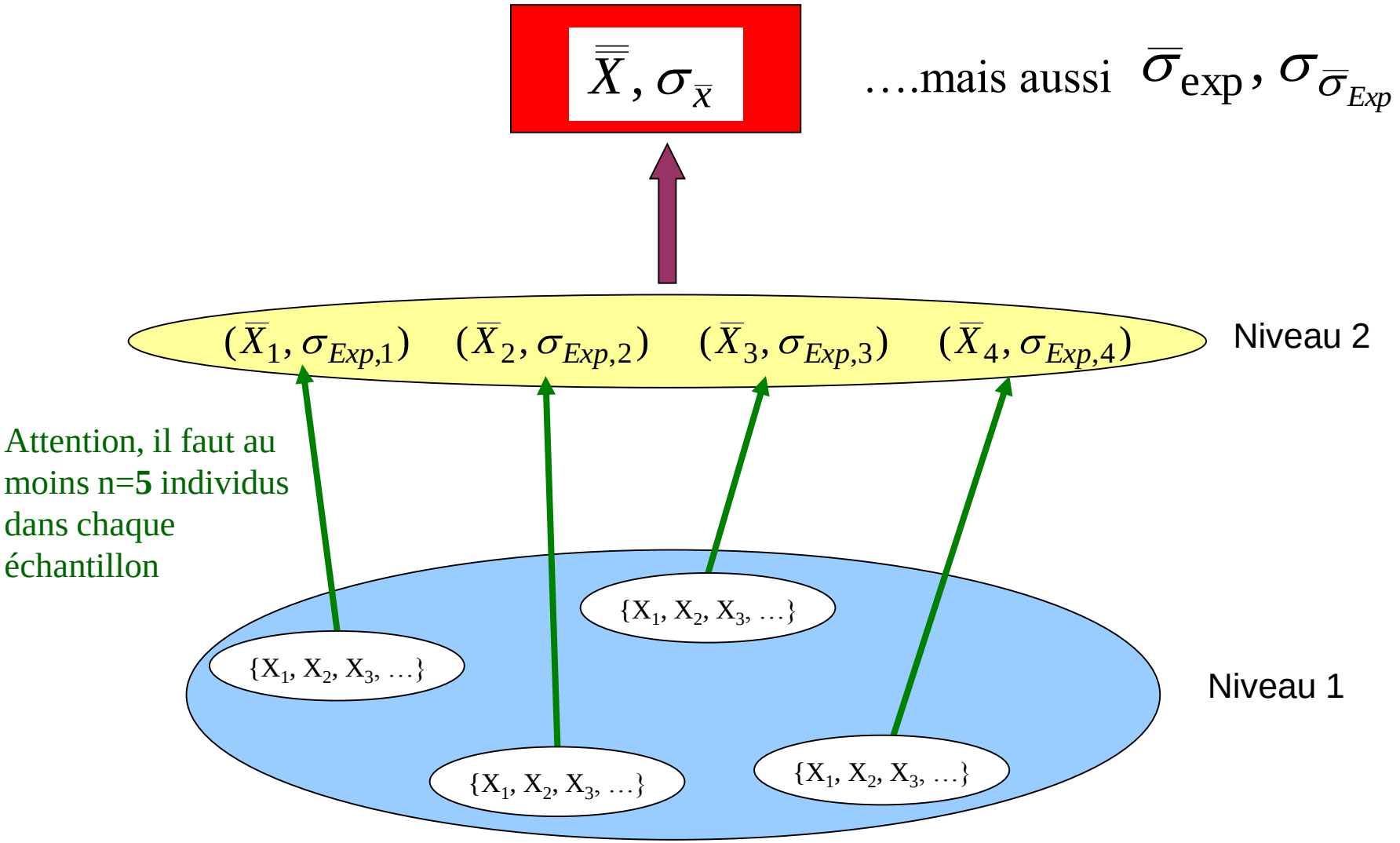
$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Normalisation

$$r_p = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \longrightarrow r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Coefficient de corrélation

ESTIMATION PLUSIEURS « ÉCHANTILLONNAGES »



VARIANCE DES MOYENNES

$$\sigma^2_{\bar{x}}\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n^2}.\sigma^2(x_1 + x_2 + \ldots + x_n) = \frac{n}{n^2}.\sigma^2_{exp} = \frac{\sigma^2_{exp}}{n}$$

Tous les échantillonnage sont indépendants les uns des autres (sinon il faut aussi prendre la covariance)
Les variables aléatoire X_i suivent la même loi

σ_{exp} est le même pour tous les échantillonnage

Finalement :

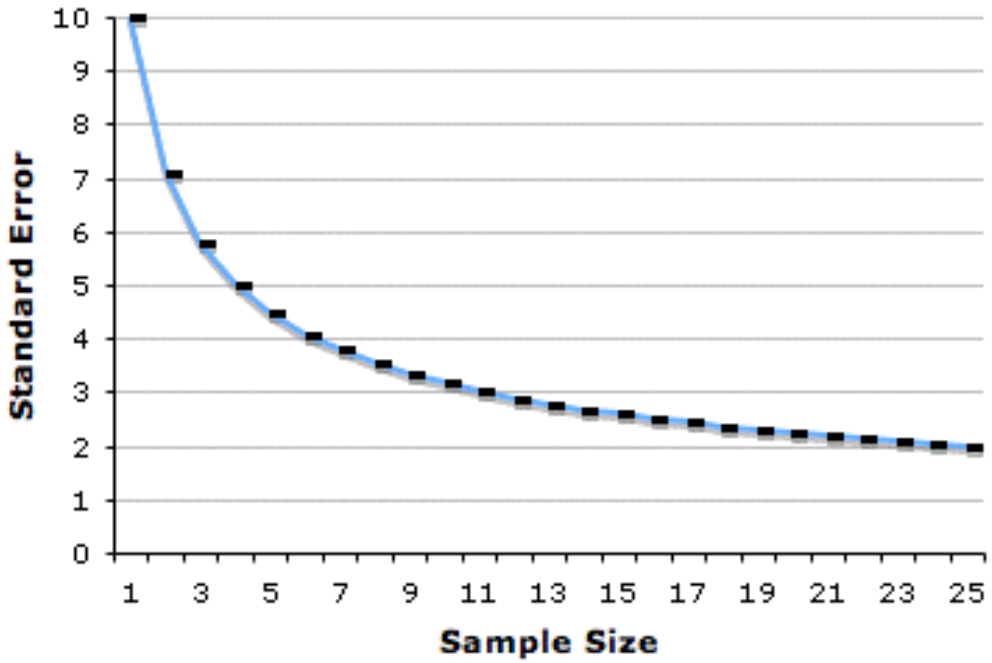
$$\sigma^2_{\bar{x}} = \frac{\sigma^2_{exp}}{n}$$

Cela suppose que σ_{exp} est constant pour tout l'échantillon (variabilité constante)

CONSÉQUENCES : RÉSULTAT INTUITIF

Cet écart type sur la moyenne prend en compte

1. L'écart type expérimental (qui caractérise les fluctuations intrinsèques au phénomène)
2. Le nombre d'échantillons réalisés



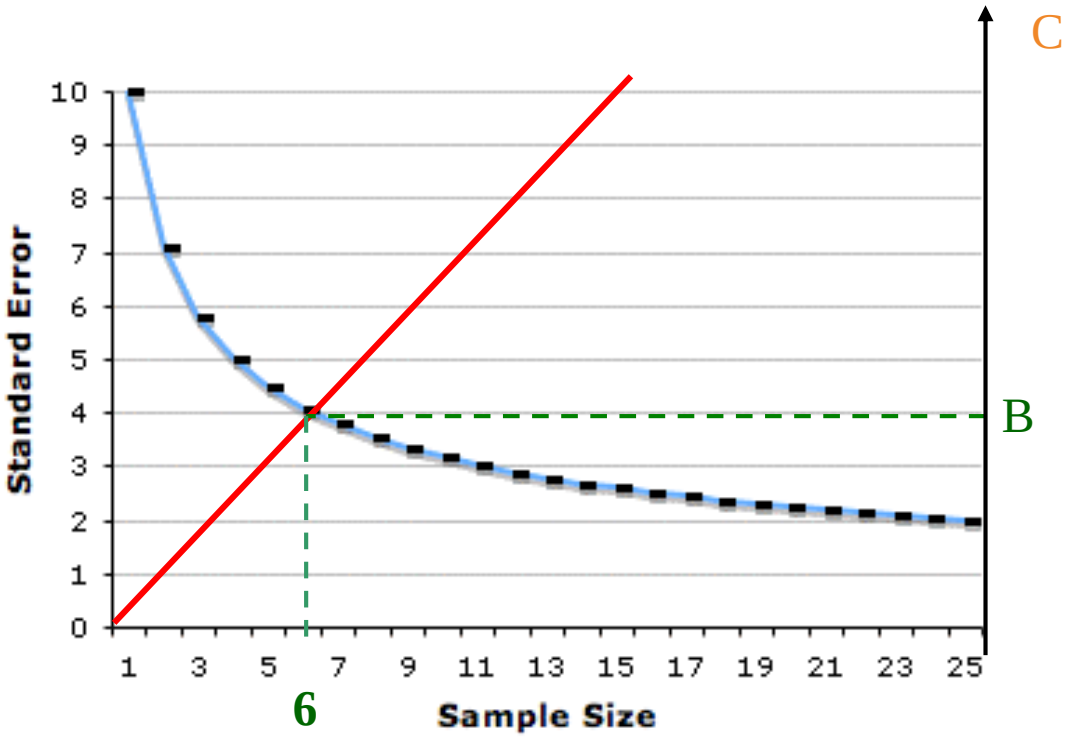
$$\sigma_{moyenne} = \frac{\sigma^{\exp}}{\sqrt{n}}$$

APPLICATION

Hypothèses :

- Le coût C des observation est linéaire C = a.n
- On dispose d'un budget total de B

Dans l'exemple suivant, à budget fixe, il faut prévoir 6 échantillons (la précision est alors fixée)



Hypothèse : les variables aléatoires sont indépendantes
Après n données, l'écart type de la moyenne est estimé par : $\sigma_{\bar{x}} = \frac{\sigma_{exp}}{\sqrt{n}}$

Finalement pour une exigence $\Delta°$ donnée, le nombre d'échantillon est :

$$n = \left(\frac{t_{1-\frac{\alpha}{2},n} . \sigma_{exp}}{\Delta° . \bar{x}} \right)^2$$

CV

Règle de Thumb : $n = \frac{16}{(\Delta°)^2}$

Cas de l'utilisation d'un plan d'échantillonnage d'efficacité D_{eff} :

$n_{eff} = \frac{n_{alea}}{D_{eff}^2}$ où n_{eff} est la taille d 'échantillonnage effective

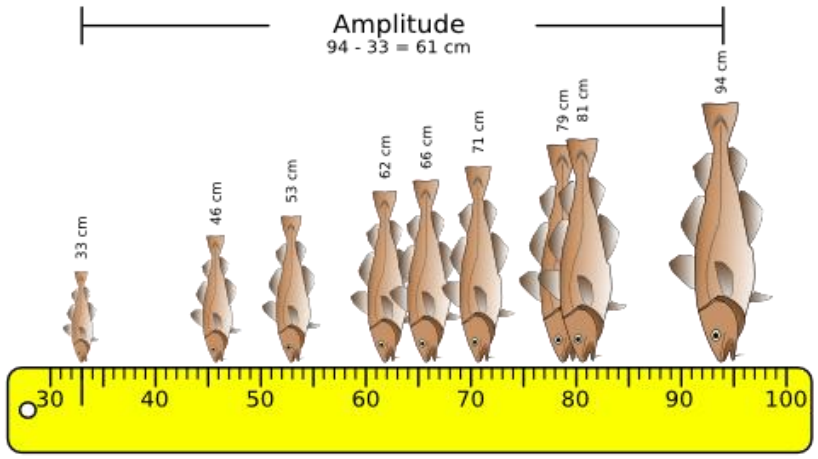
ESTIMATION À PRIORI DE LA VARIANCE

Pour tous les plans d'échantillonnage, il est nécessaire d'évaluer σ_{exp} ou le coefficient de variation CV pour estimer une taille d'échantillon.

→ Faire une bibliographie pour trouver des situations proches
→Faire une étude pilote grossière ou pré-échantillonnage
→approche bayesienne plus adaptée ?

$$\sigma_{exp} \approx \frac{E(X_{max}) - E(X_{min})}{6} = \frac{\text{Etendue des valeurs de X}}{6}$$

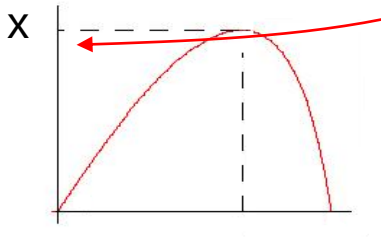
$$\sigma_{exp} \approx \frac{R_x}{6}$$



MODÉLISATION DES ALÉAS, DISTRIBUTION

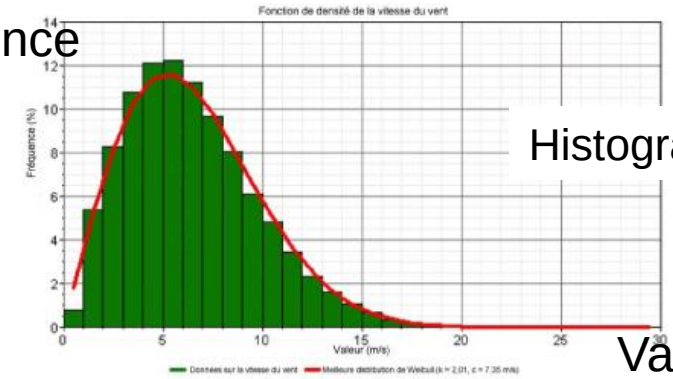
La loi en mécanique classique :

$$m\frac{d\vec{v}}{dt} = \sum_i \vec{F}_i = m\vec{g} \quad \Rightarrow \quad x = \frac{1}{2}g.t^2 + v_0.t + x_0$$



La loi en probabilité Probabilité **P** = **f**(**X**$_{\text{variable aléatoire}}$)

Fréquence

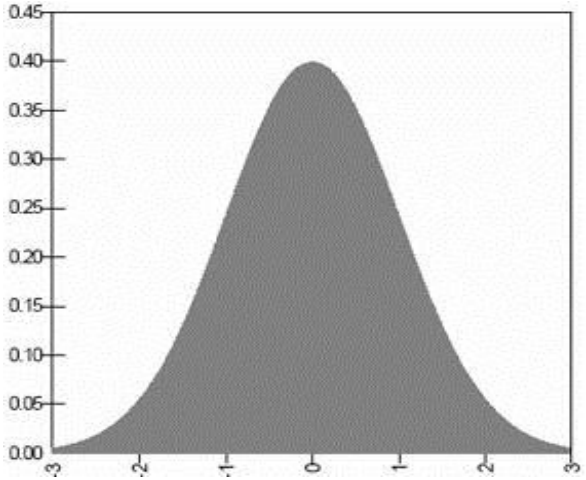


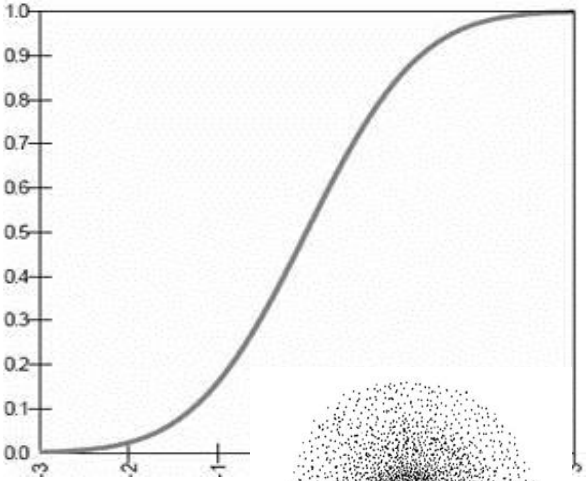
Histogramme

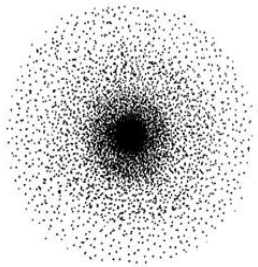
Valeur observée

EXEMPLE DE DISTRIBUTION

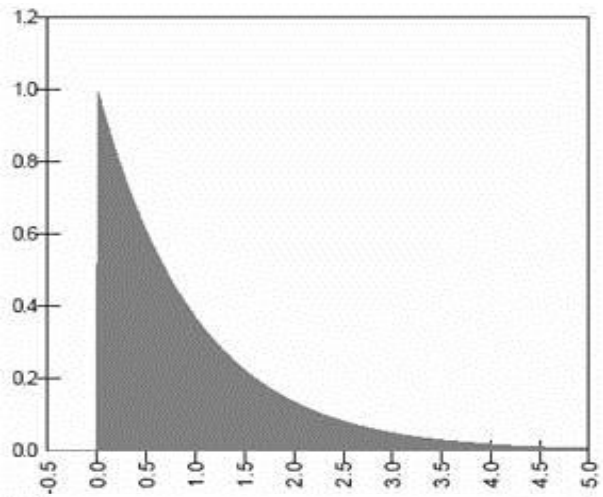
Loi normale

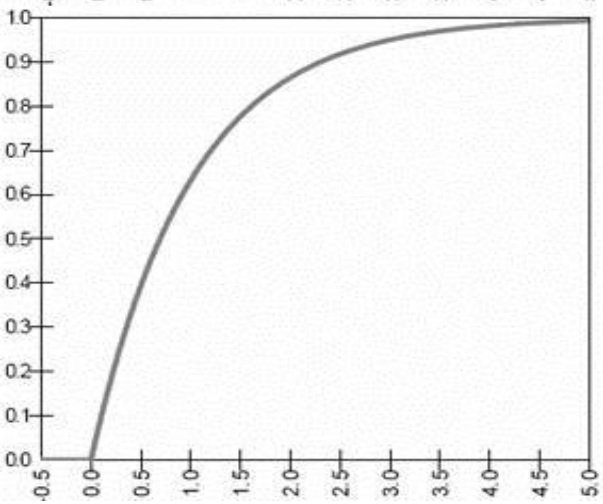




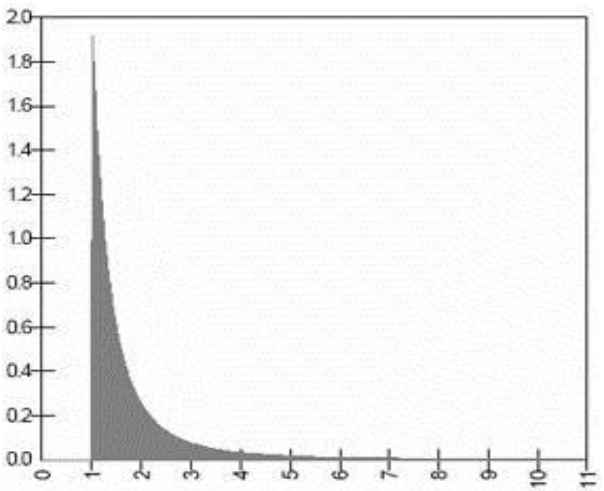


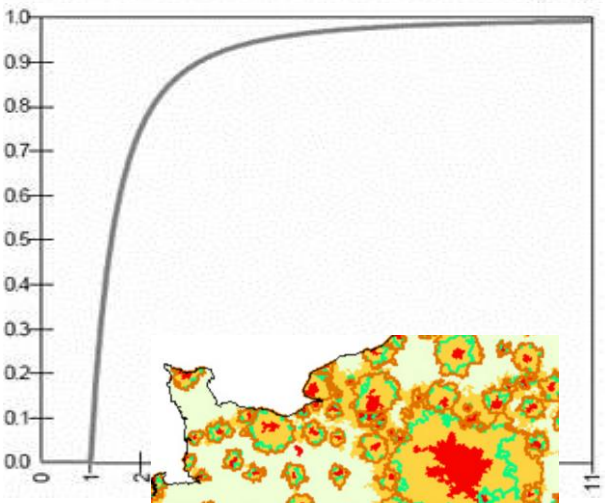
Loi exponentielle





Loi puissance





INTRODUCTION A LA CLASSIFICATION

Il s'agit d'apporter de l'information en structurant par regroupement d'unités d'une population

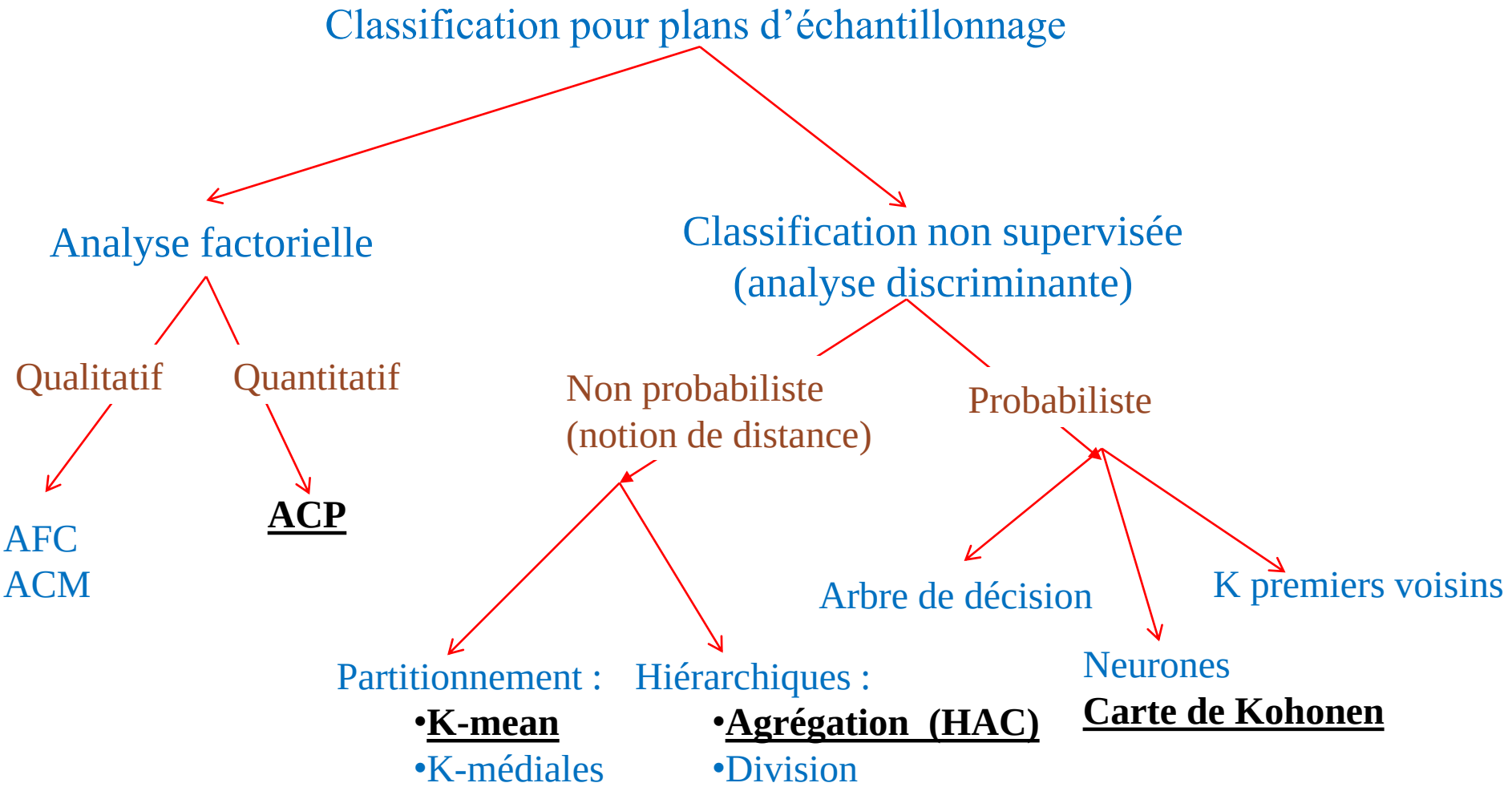
→ Avec un échantillon on pourra ensuite extrapoler au reste de la population

→ En comparant et en regroupant suivant leurs propriétés un échantillon (pré-échantillonnage), on pourra optimiser l'échantillonnage suivant (en taille et en lieu de prélèvement)
Remarque : un pré échantillonnage s'apparente donc à un plan à 2 niveaux

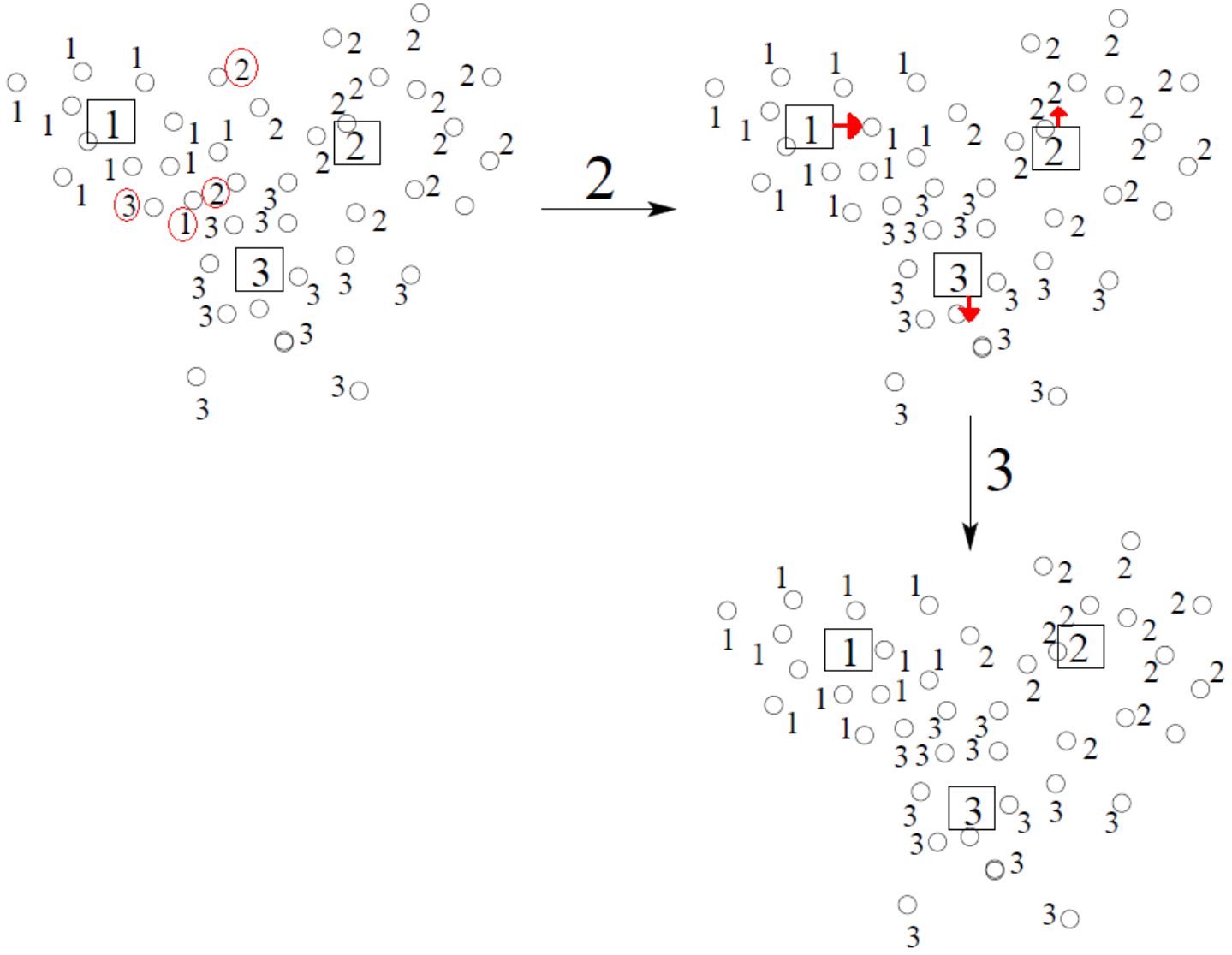
MÉTHODOLOGIE POUR LE CLUSTERING

1. Quantifier les variables associées aux unités
2. Calculer des distances ou des valeurs
3. Identifier un seuil pour tester ou optimiser
4. Fixer les conditions d'arrêt
5. Choisir et établir l'algorithme (programme séquentiel de calcul)
6. Préciser les conditions initiales et lancer l'algorithme
7. Définir des indicateurs de performances et de qualité de la classification

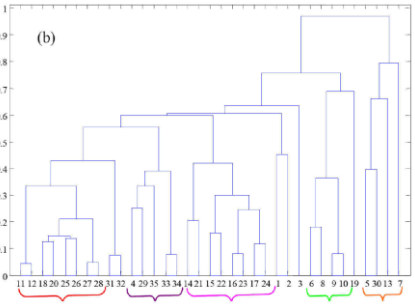
REGROUPEMENT - CLASSEMENT (*CLUSTERING*)



EVOLUTION GRAPHIQUE



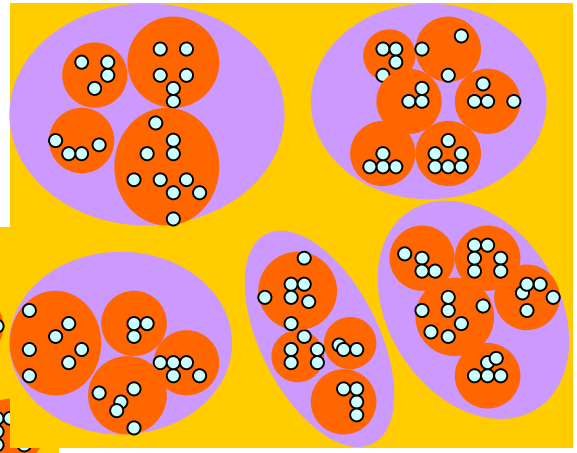
CLASSIFICATION HIÉRARCHIQUE

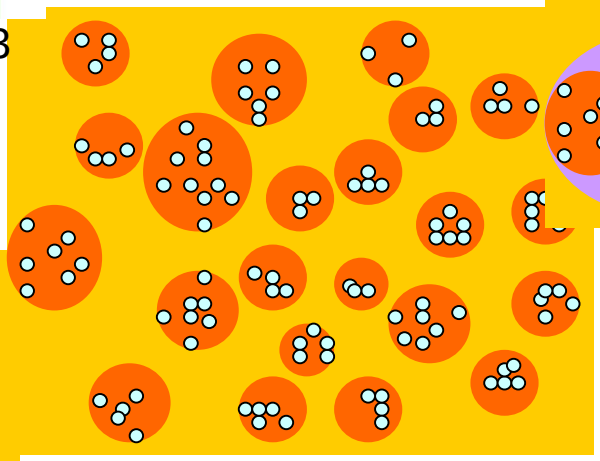


⇒

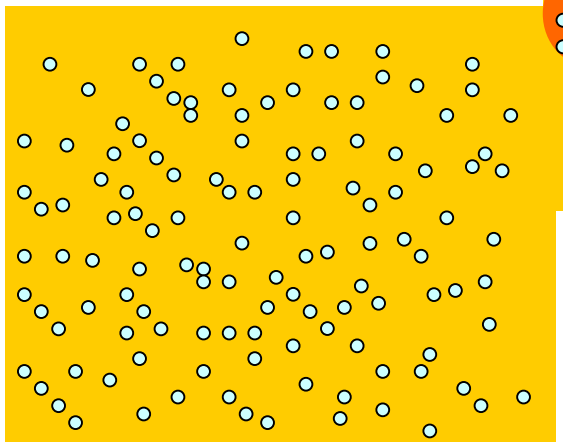
N = 23

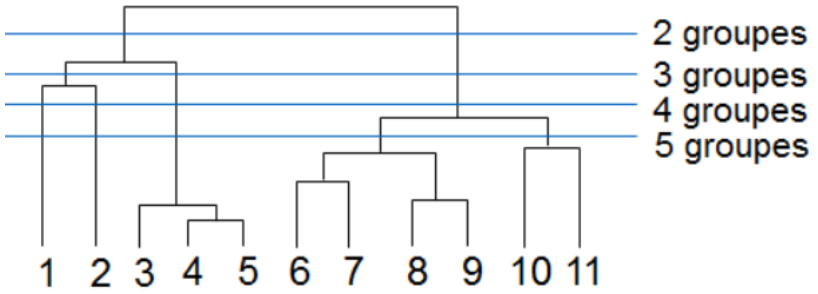
N=5





N ≈ 123



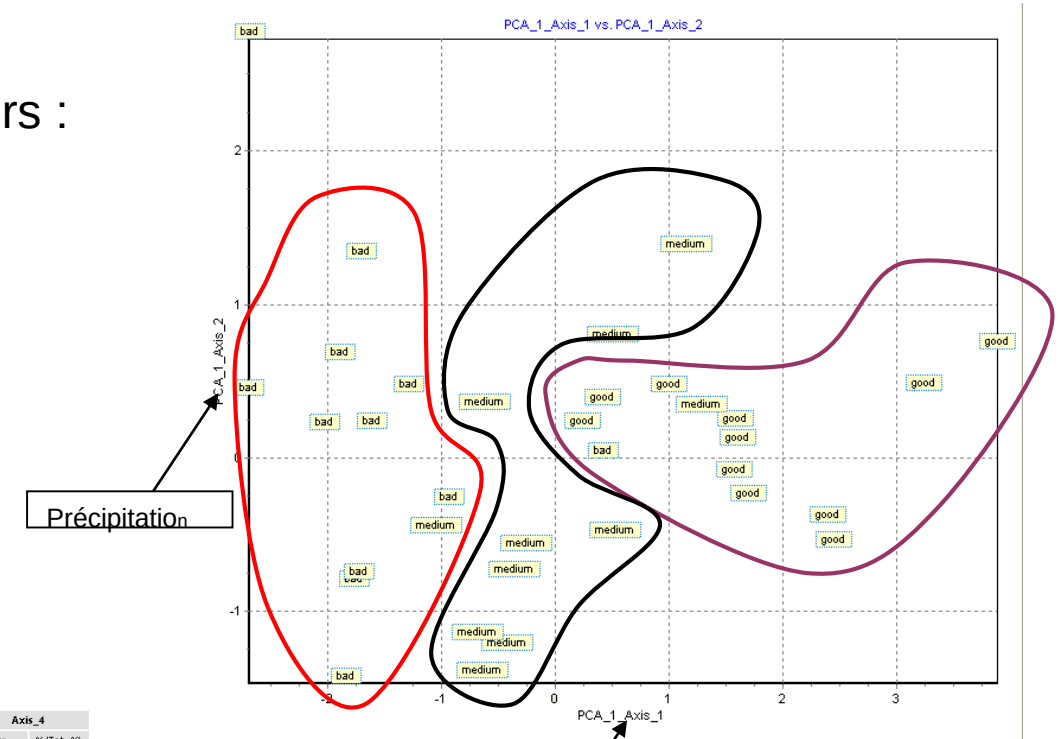


ANALYSE EN COMPOSANTES PRINCIPALES ACP

Utile pour l'identification des stratificateurs (v. Plans stratifiés)

Exemple de classificateurs :

- Précipitation
- Température



Précipitation

Température
Soleil
Durée d'ensoleillement

Eigen values

Matrix trace = 4,00

Axis	Eigen value	% explained	Histogram	% cumulated
1	2,791467	69,79%		69,79%
2	0,714470	17,86%		87,65%
3	0,365928	9,15%		96,80%
4	0,128136	3,20%		100,00%
Tot.	4,000000	-	-	-

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Temperature (°C)	0,9197	85 % (85 %)	0,2580	7 % (91 %)	-0,1255	2 % (93 %)	-0,2681	7 % (100 %)
Sun (h)	0,8577	74 % (74 %)	0,0072	0 % (74 %)	0,5115	26 % (100 %)	0,0524	0 % (100 %)
Heat (days)	0,8964	80 % (80 %)	0,2683	7 % (88 %)	-0,2666	7 % (95 %)	0,2312	5 % (100 %)
Rain (mm)	-0,6376	41 % (41 %)	0,7589	58 % (98 %)	0,1321	2 % (100 %)	0,0089	0 % (100 %)
Var. Expt.	2,7915	70 % (70 %)	0,7145	18 % (88 %)	0,3659	9 % (97 %)	0,1281	3 % (100 %)

Eigen vectors -- Factor Scores

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3	Axis_4
Temperature (°C)	3157,882353	139,092598	0,550458	0,305215	-0,207543	-0,748843

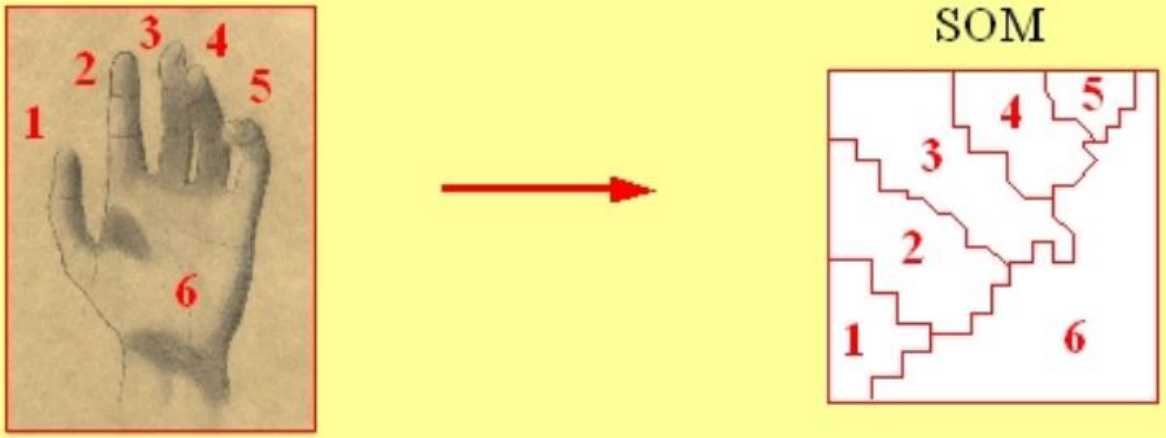
CARTES KOHONEM (SOM)

Principe

⇒ Classification de données sous formes de cartes dites auto-organisées ou encore topologiques (catégorisation des exemples).

Intérêt

⇒ Préservation de la topologie
(contrairement à d'autres méthodes de classification) :



ALLOCATION

Adapter le nombre d'unité à prélever en fonction des informations apportées par un plan d'échantillonnage

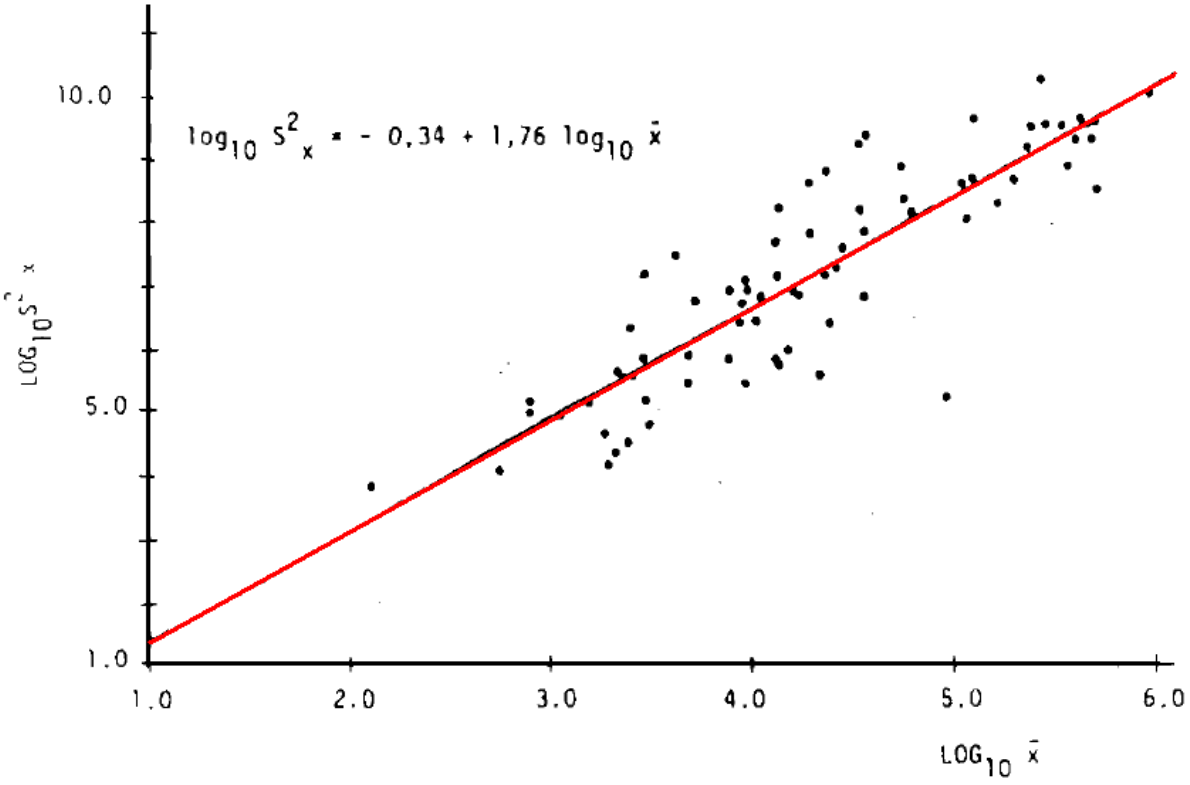
Allocation **équilibrée** : échantillon de **même taille** pour toutes les classes

Allocation **proportionnelle** : échantillon de taille proportionnelle à la **taille de la classe**

Allocation de **Newman** : échantillon dont la taille est adapté à l'**hétérogénéité de la classe**

Plan d'échantillonnage pour évaluer la pollution fécale dans les étangs de Thau

75 prélèvements d'eau, 5 mesures par prélèvements



Loi de Taylor : $\log(s^2_{exp}) = 1,76.\log(x_{moy}) - 0,34$

X_{moy} : dénombrement des bactéries hétérotrophes revivifiables

Les plans triviaux

- Plan aléatoire (*random sampling*)
- Recensement (*census*)
- Plan de jugement (*judgemental sampling*)

PLAN ALÉATOIRE

Correspond plus à la façon dont sont sélectionnées les unités

Les autres façons de sélectionner étant :

- Sélection de l'unité suivante suivant la valeur de l'unité présente (v plan progressif)
- Sélection de l'ensemble des unités en relation avec l'unité présente (v plan boule de neige)

ECHANTILLONNAGE SANS REMISE

Après sélection, les unités peuvent être :

- réintroduites : **échantillonnage avec remise**
- gardées : **échantillonnage sans remise**

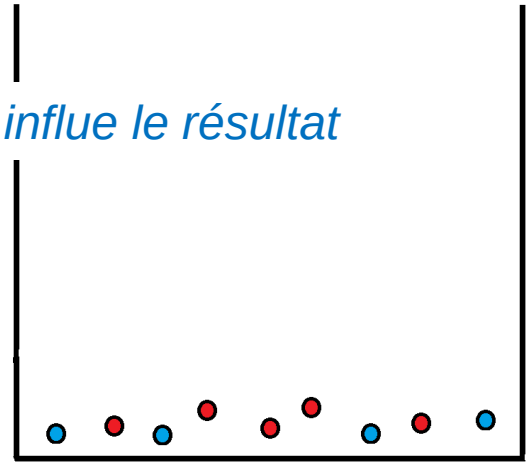
Quand n ~ N l'action d'échantillonnage a un effet sur le résultat

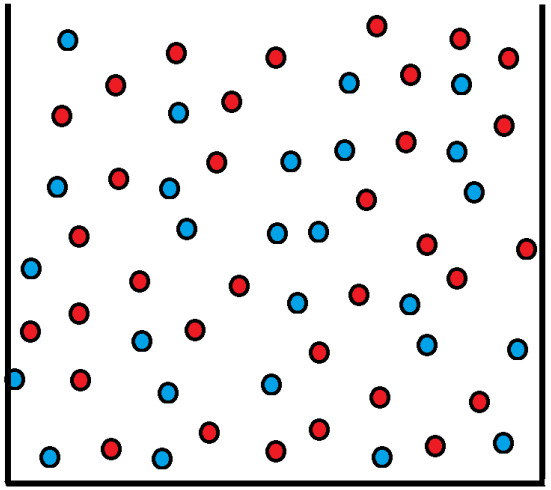
Il en résulte un « effet de bord » qu'il faut corriger dans la probabilité de sélection

$$\frac{1}{\pi_i} = \frac{1}{C_n^N} = \frac{n!(N-n)!}{N!}$$

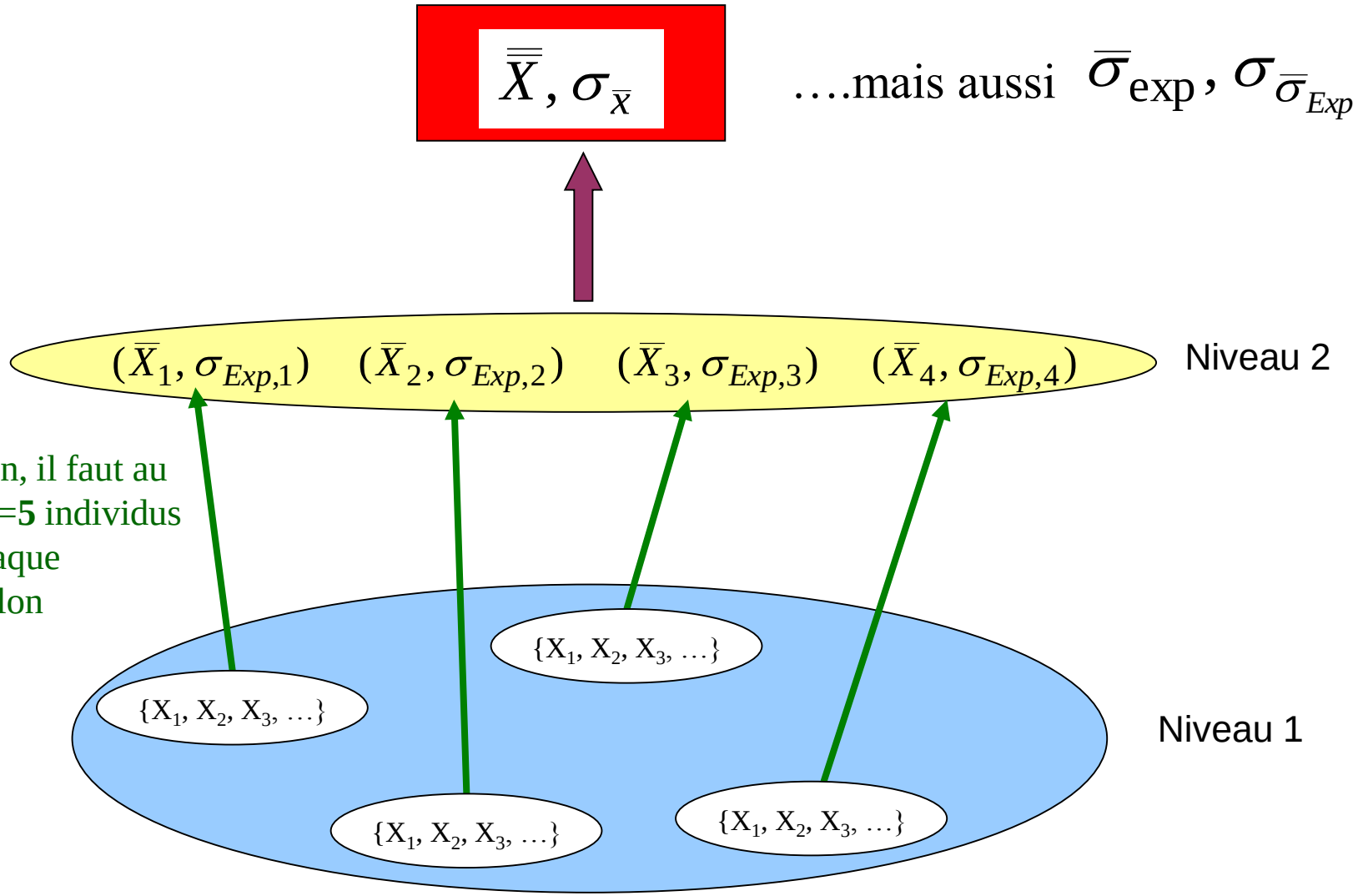
La remise n'influe pas le résultat

La remise influe le résultat





NOUVEL APPORT D'INFORMATION : PLUSIEURS ÉCHANTILLONAGES



VARIANCE DES MOYENNES (INDISCERNABILITÉ DES INDIVIDUS = PLAN ALÉATOIRE)

$$\sigma^2_{\bar{x}}\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \frac{1}{n^2}.\sigma^2(x_1 + x_2 + \ldots + x_n) = \frac{n}{n^2}.\sigma^2_{exp} = \frac{\sigma^2_{exp}}{n}$$

Tous les échantillonnage sont indépendants les uns des autres (sinon il faut aussi prendre la covariance)
Les variables aléatoire X_i suivent la même loi

σ_{exp} est le même pour tous les échantillonnage

Finalement :

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_{exp}}{n}$$

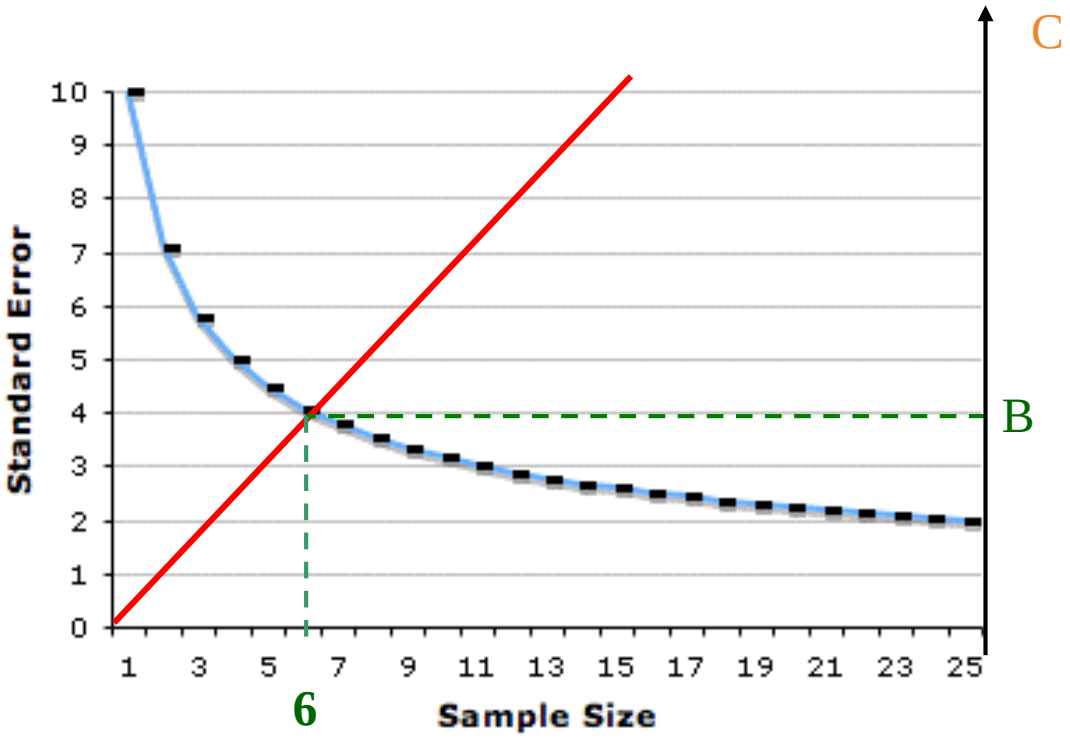
Cela suppose que σ_{exp} est constant pour tout l'échantillon (variabilité constante)

APPLICATION : ESTIMATION DE LA TAILLE DE L'ÉCHANTILLON

Hypothèses :

- Le coût C des observation est linéaire C = a.n
- On dispose d'un budget total de B

Dans l'exemple suivant, à budget fixe, il faut prévoir 6 échantillons (la précision est alors fixée)



EXEMPLE DE GÉNÉRATEURS D'ALÉA

Générateurs artificiels :

- Les tables
  - Suivant le nombre de chiffres servant à repérer les unités
- Les algorithmes
  - Basés sur des lois statistiques

Générateurs physiques :

- La radioactivité
- Les bruits thermiques
- Les bruits électromagnétiques
- Mécanique quantique

La fonction **=ALEA()** :

Elle renvoie un nombre aléatoirement compris entre 0 et 1

Pour des nombres compris entre a et b : **=ALEA()*(b-a)+a**

Attention, ce qui s'apparente à de l'aléa ne l'est pas forcément :

Cas des systèmes chaotiques où derrières l'apparence d'un comportement aléatoire se cache un système déterministe avec un système d'équation non linéaires

Les plans par sélection régulière

- Plan systématique (*systematic sampling*)
- Carte de contrôle (*control charts*)

APPLICATION

•Plan très **intuitif** et donc très populaire (impression d'un échantillonnage méthodique et « déterministe »)

•Permet d'avoir une **couverture** uniforme sur un domaine d'étude

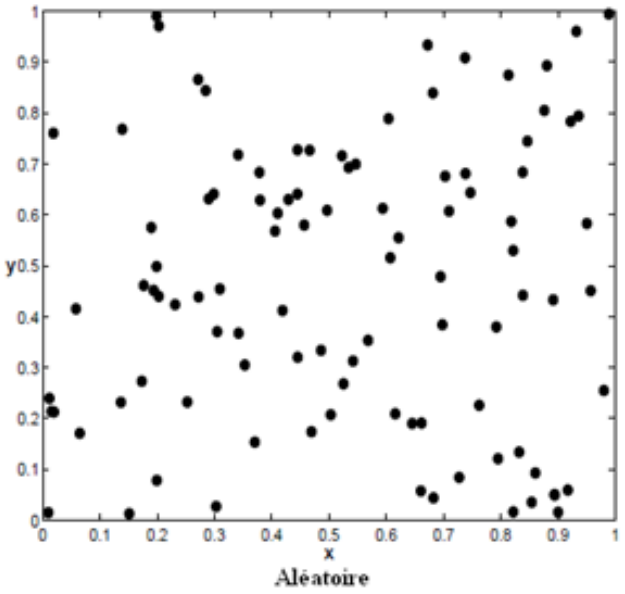
•Facile d'utilisation et souple (adaptation de la taille du maillage à l'hétérogénéité)

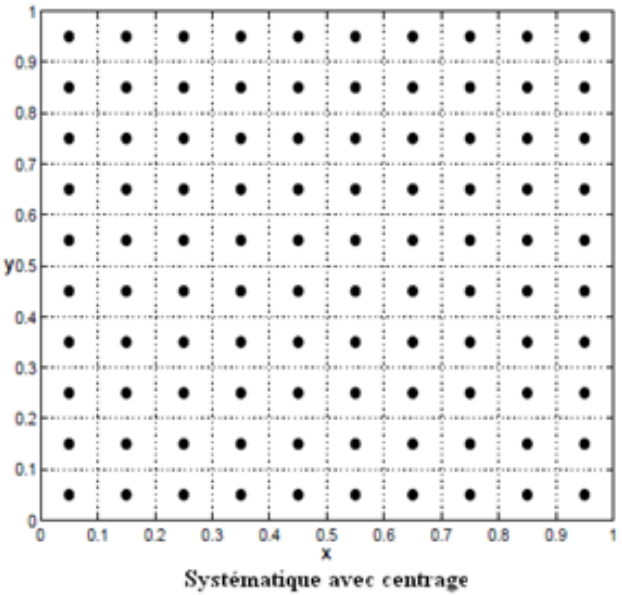
Plan adapté pour :

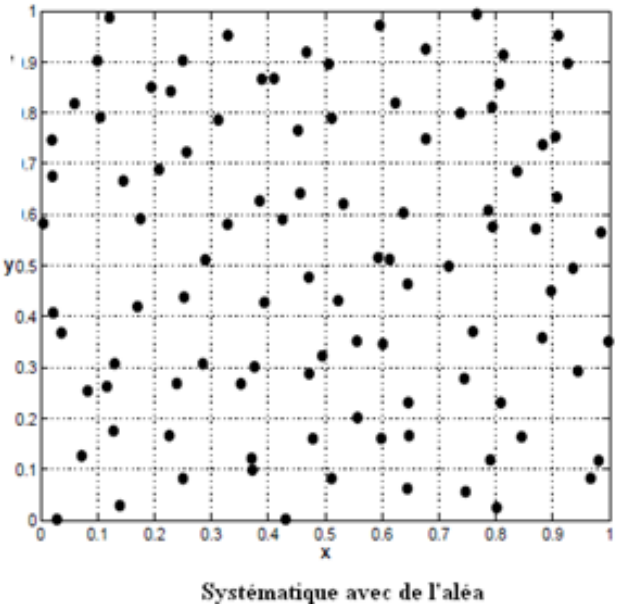
- •La recherche de corrélations
- •Estimer leurs importances
- •Les phases exploratoires (pré-échantillonnage)

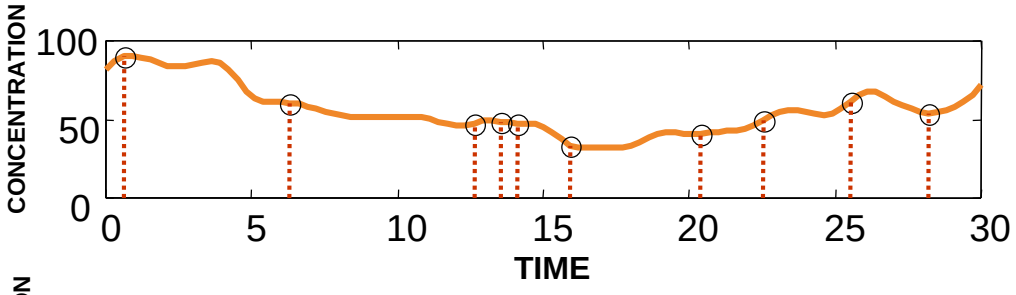
Plan souvent comparé avec les plans aléatoires et stratifiés

REPRÉSENTATION SHÉMATIQUE

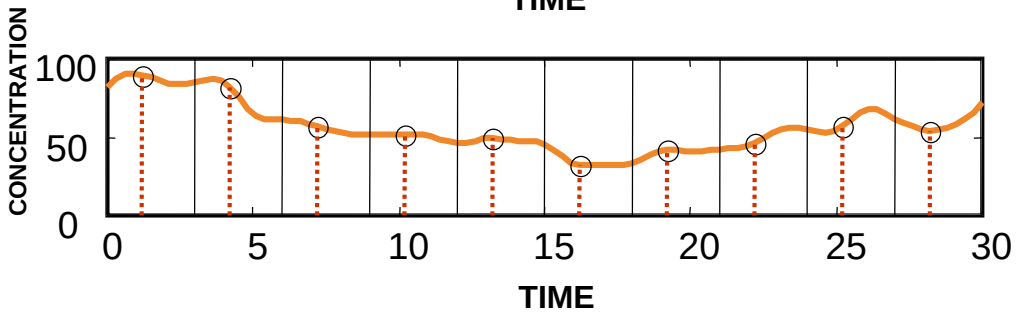








Random selection



Systematic selection

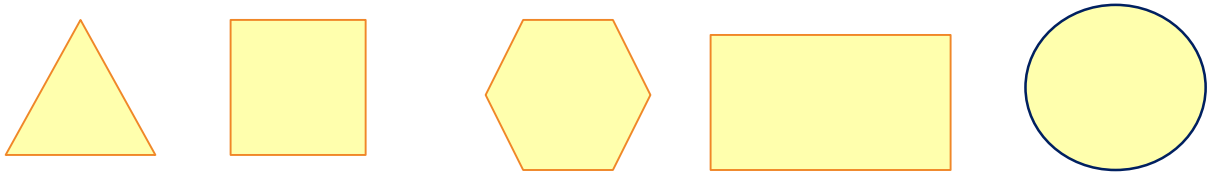
MÉTHODOLOGIE

1. Vérifier l'existence d'un système de coordonnées
2. Les unités sont repérées par ces coordonnées
3. Choisir une périodicité k
4. Découpage du domaine d'étude en k mailles régulières
5. Sélection aléatoire de la première unité
6. Sélection d'une unité dans chacune des mailles (distantes d'un multiple de k)

LA MAILLE

Elle varie suivant :

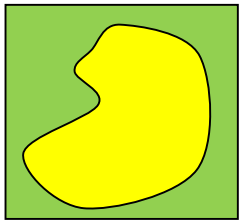
- Leur forme



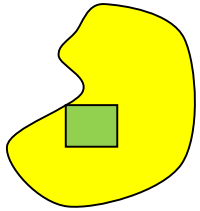
Le triangle équilatéral semble être le plus performant
Les mailles carrées sont les plus utilisées

- Leur orientation
- Leur taille (liée à la périodicité de K)

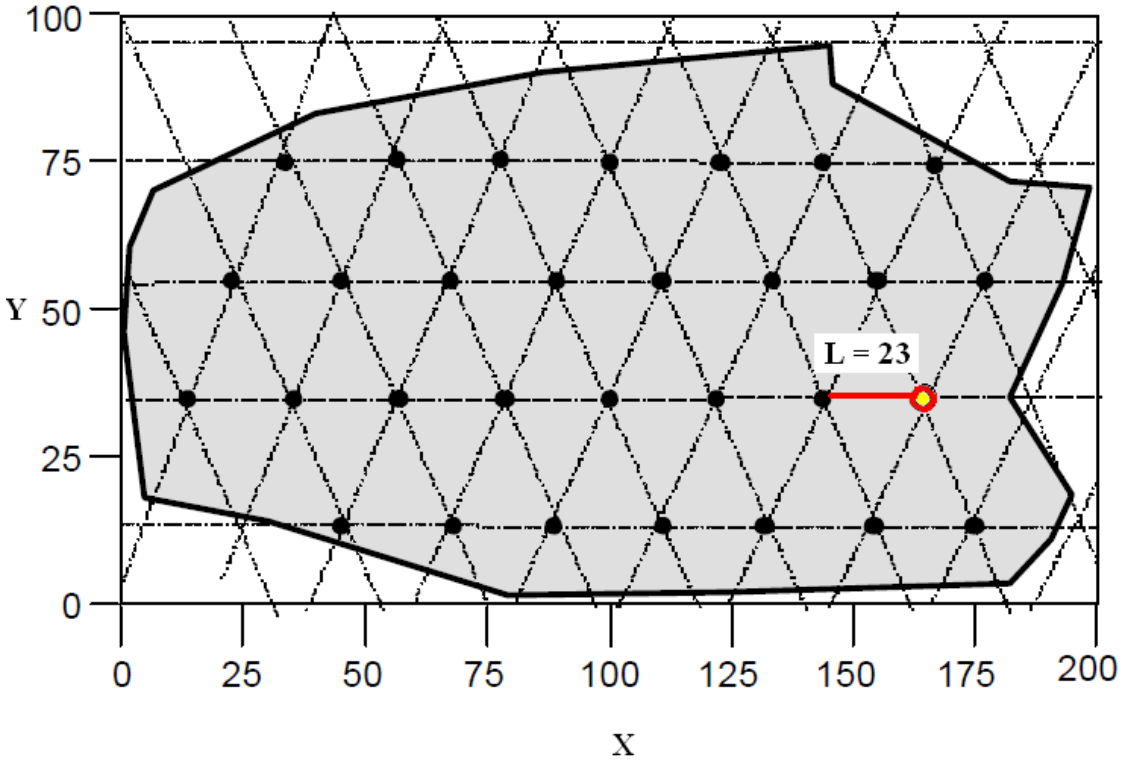
Trop grand



Trop petit



MAILLAGE CORRESPONDANT



ÉCHANTILLONNAGE SYSTÉMATIQUE OU STRATIFIÉ ?

Il est impossible d'évaluer la variance de la moyenne avec un seul échantillonnage

Mais en réalisant n échantillonnage (n_{max} ~ N/k), cette variance peut être calculée, en particulier pour un échantillon n (qui correspond à une classe) la variance sur la moyenne s'écrit :

$$\hat{\sigma}^2_{\bar{x}n,\, n\times Syst} = \left(1 - \frac{1}{N}\right)\frac{S^2_{tot}}{n}[1+(n-1)]$$

Pour un échantillonnage aléatoire dans une des strates d'un échantillonnage stratifié, on aurait eu :

$$\hat{\sigma}^2_{\bar{x}h,aléa} = \left(1 - \frac{n}{N}\right)\frac{S^2_{tot}}{n}$$

Il en résulte que l'échantillonnage systématique est plus précis que l'échantillonnage stratifié lorsque :

$$\hat{\sigma}^2_{\bar{x}n,\, n\times Syst} < \hat{\sigma}^2_{\bar{x},aléa} \quad soit : \rho < \frac{-1}{n.k-1}$$

CONSÉQUENCES

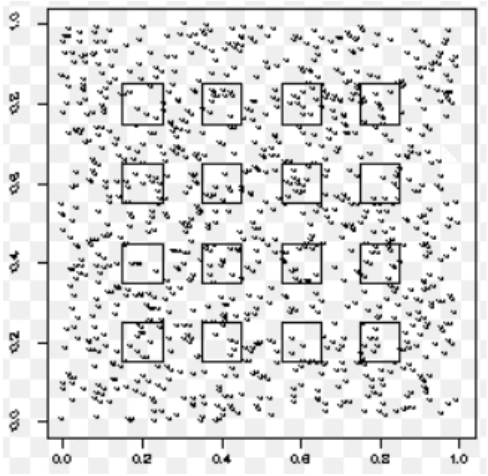
Si $\rho < 0$ l'échantillonnage systématique est avantageux
Si $\rho > 0$ l'échantillonnage systématique est DESavantageux

Si on considère que les deux plans apportent une même précision, le coefficient ρ est de la forme :

$$\rho = \frac{1}{1 - n.k} = \frac{1}{1 - N} \text{ qui est sensiblement egalà } 0$$

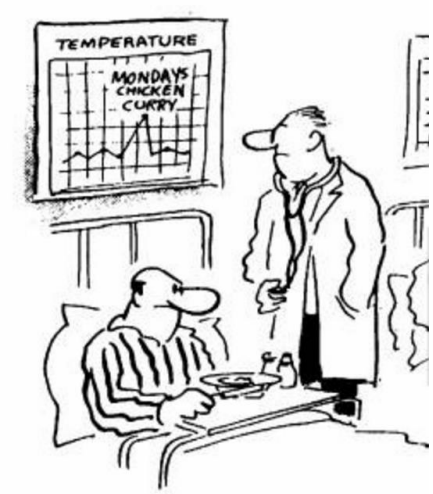
L'efficacité du plan étant : $D_{eff} = \frac{N-1}{n.(k-1)} [\ 1 + (n-1)\rho\ ]$

QUELQUES REMARQUES



Cas d'un plan à deux niveaux avec

- 1er niveau : plan systématique
- 2ème niveau : plan aléatoire



La variable auxiliaire est le temps, c'est la cas de certaines cartes de contrôle

→ Voir plus loin dans la présentation

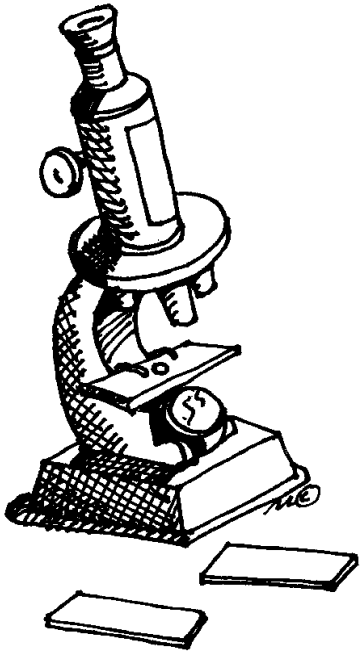
Les plans par regroupement

- Plan stratifié (*stratified sampling*)
- Echantillonnage en grappes (cluster sampling – hot spot)

ANALOGIE AVEC LE MICROSCOPE

Il s'agit d'adapter la sélection progressivement comme on le ferait en choisissant le grossissement d'un microscope ou d'un télescope

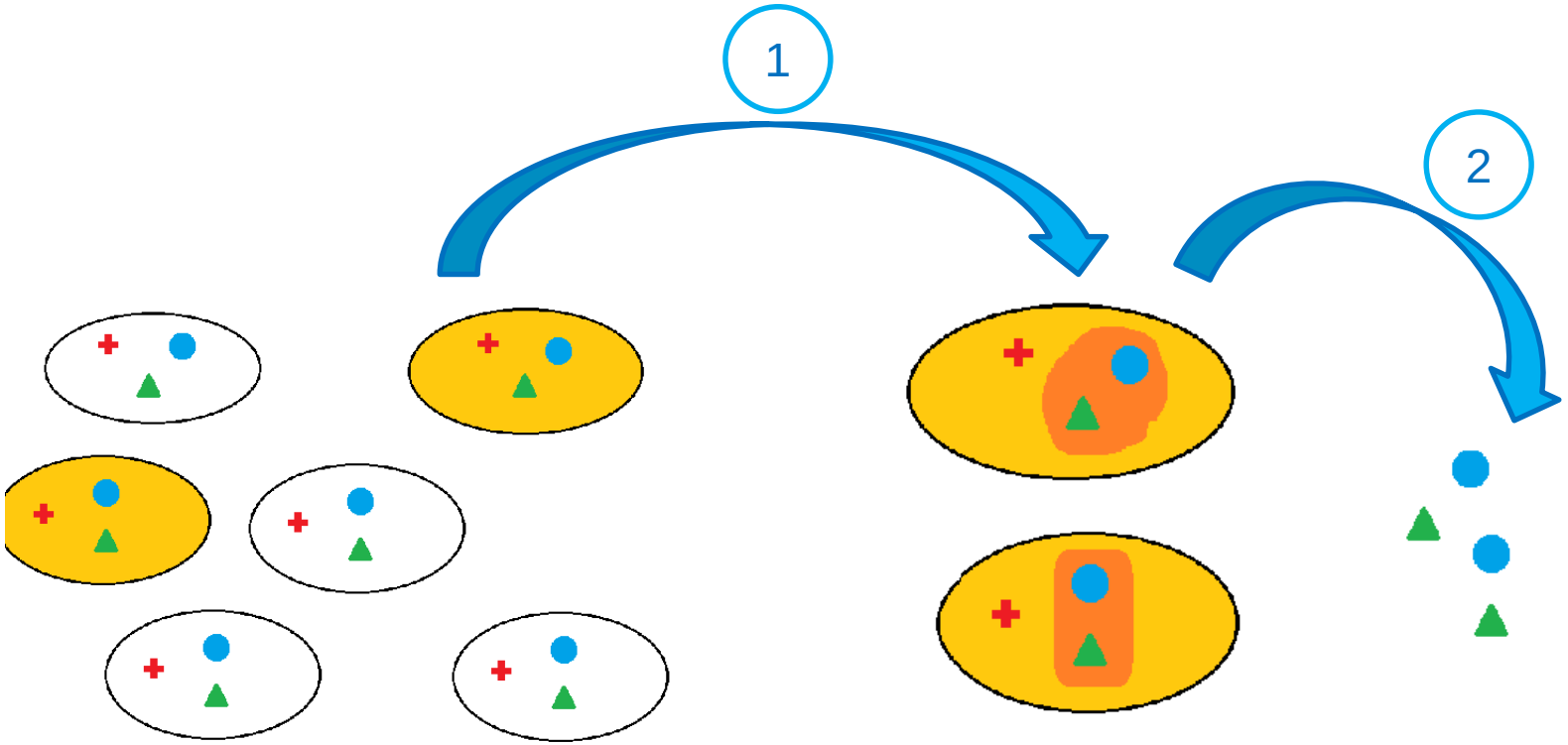
Changer de niveaux c'est un peu changer de monde



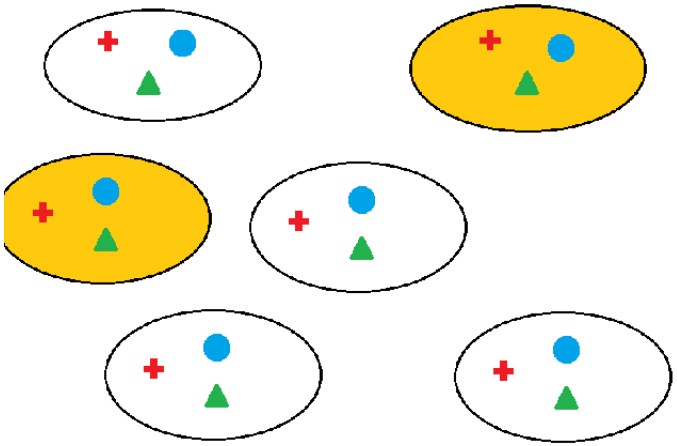
Dans un plan à 2 niveaux ou 2 degrés on a :

- **Un niveau macroscopique**
- **Un niveau microscopique**

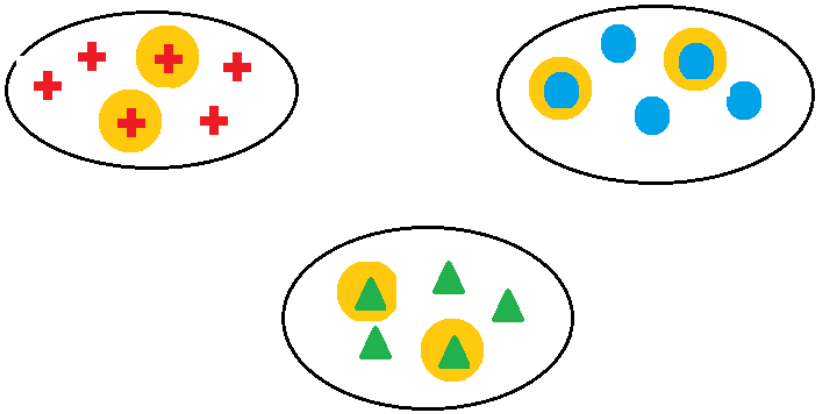
PLAN 2 NIVEAUX OU 2 DEGRÉS



CAS PARTICULIERS DE PLAN A 2 NIVEAUX
STRATES ET GRAPPES



- Sélection de 2 unités au niveau **macro** (grappes)
- Grappes hétérogènes (pas de regroupements)
- Exemples : villes, écosystèmes, …



- Sélection de 2 unités au niveau **micro** (strates)
- Strates homogènes (unités semblables rassemblées en sous-groupes)
- Exemples : groupe d'âge, groupe de fruits, groupe de …

COMPARAISON PLAN STRATIFIÉ ET EN GRAPPE

Concernant le mode de sélection

	Plan stratifié	Plan en grappes
Sélection "Macro"	Exhaustif	Aléatoire
Sélection "Micro"	Aléatoire	Exhaustif

Concernant la variance

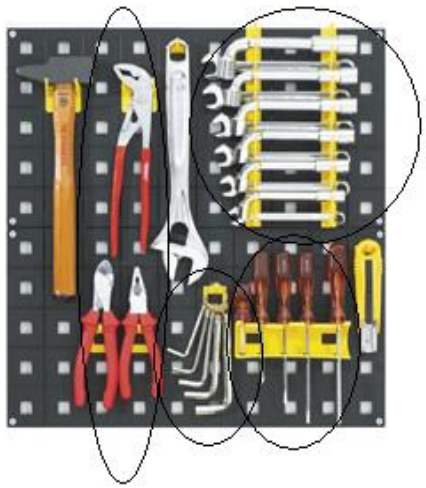
	Plan stratifié	Plan en grappes
Variance dans les classes	Faible	Importante
Variance entre classes	Importante	Faible

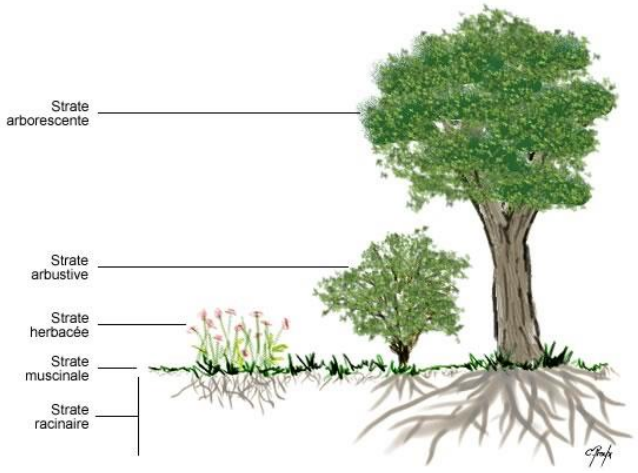
PRINCIPE

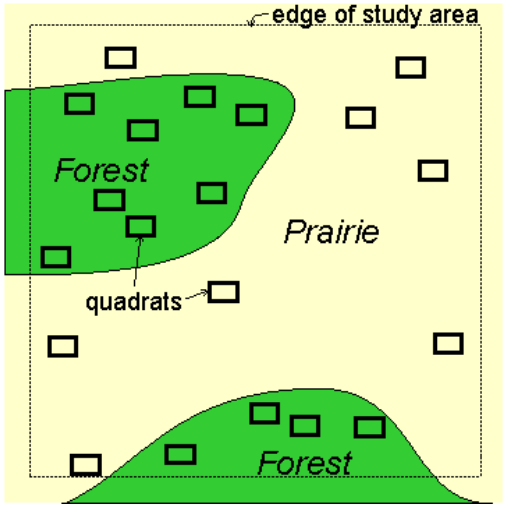
Idée : regrouper les unités suivant leurs ressemblance en **sous-ensembles homogènes** appelés strates

Un échantillonnage dans ces strates s'apparente à la réalisation d'un **plan aléatoire**

Rappel : plan aléatoire = plan dont on ne peut plus extraire d'information







Règle de Thumb pour l'estimation de la taille d'une population :

- ***Choix de 6 strates***
- ***Pour chaque strate un échantillon de 5 à 10 unités***

APPLICATION

Problématique : distribution des fourmis dans les bois

Choix des stratificateurs : (variables qui influencent la présence de fourmis), intuitivement (plan de jugement) :

1. L'altitude
   3 zones : 380-800, 801-1200, >1200m
2. L'exposition au soleil
   2 zones : favorable et défavorable
3. La pente (stabilité de la fourmilière)
   2 zones 1 – 20° et 25 – 45°)
4. Couvert végétal
   2 zones : lisière et pleine forêt

1. Identification des 3×2×2×2 = 24 strates sur une carte
2. Tirage au sort de 10 quadrat de 1 ha dans les 24 strates
   → soit 240 unités d'échantillonnage
3. Plan en transect :
   - unités parallèles espacées de 15 m,
   - recensement des fourmilières et identification des espèces

ESTIMATION DES PARAMÈTRES

Pour chaque strate h, on défini un poids $W_h = \frac{N_h}{N}$

La moyenne totale est alors : $\overline{\overline{x}} = \sum_{h=1}^{m} w_h . \overline{x}_h$

Et la variance sur cette moyenne : $\sigma^2_{\overline{\overline{x}}} = \sum_{h=1}^{m} w_h^2 . \sigma_h^2$

EXEMPLE D'APPLICATION

Population : 210 communes - Ces communes sont stratifiées en 4 strates :

	Strate 1	Strate 2	Strate 3	Strate 4
Nh	105	63	21	21
	↓	↓	↓	↓
Échantillonnage	7	14	23	24
	10	11	36	110
	8	14	2	17
	2	7	9	47
	7	17	24	32
	6	19	Ech3	Ech4
	2	9	$n_3=5$	$n_4=5$
	6	Ech2		
	Ech1	$n_2=7$		
	$n_1=8$			

CALCUL DE LA MOYENNE

Strate	1	2	3	4
Effectif	105	63	21	21
Taille de l'échantillon	8	7	5	5
Moyenne	2,775	4,2817	13,4052	37,4767
Ecarts-types de l'échantillon	2,7775	4,2817	13,4052	37,4767
$\sqrt{1-f_h/n_h}$	0,3398	0,3564	0,3904	0,3904
Ecarts-types des estimateurs	0,9438	1,5258	5,2329	14,6294
Poids de la strate	0,5	0,3	0,1	0,1

Calcul de la moyenne totale :

$$\bar{\bar{x}} = \sum_{h=1}^{m} w_h . \bar{x}_h = (0,5.6) + (0,3.13) + (0,1.18,8) + (0,1.46) = 13,38$$

CALCUL DE LA VARIANCE

On calcule

$$\hat{\sigma}^2_{\bar{y}1_1} = (1 - \frac{8}{105}) \frac{2{,}7775^2}{8} = 0{,}6886$$

$$\hat{\sigma}^2_{\bar{y}2} = 2{,}3280$$

$$\hat{\sigma}^2_{\bar{y}3} = 27{,}3829$$

$$\hat{\sigma}^2_{\bar{y}4_1} = 214{,}019$$

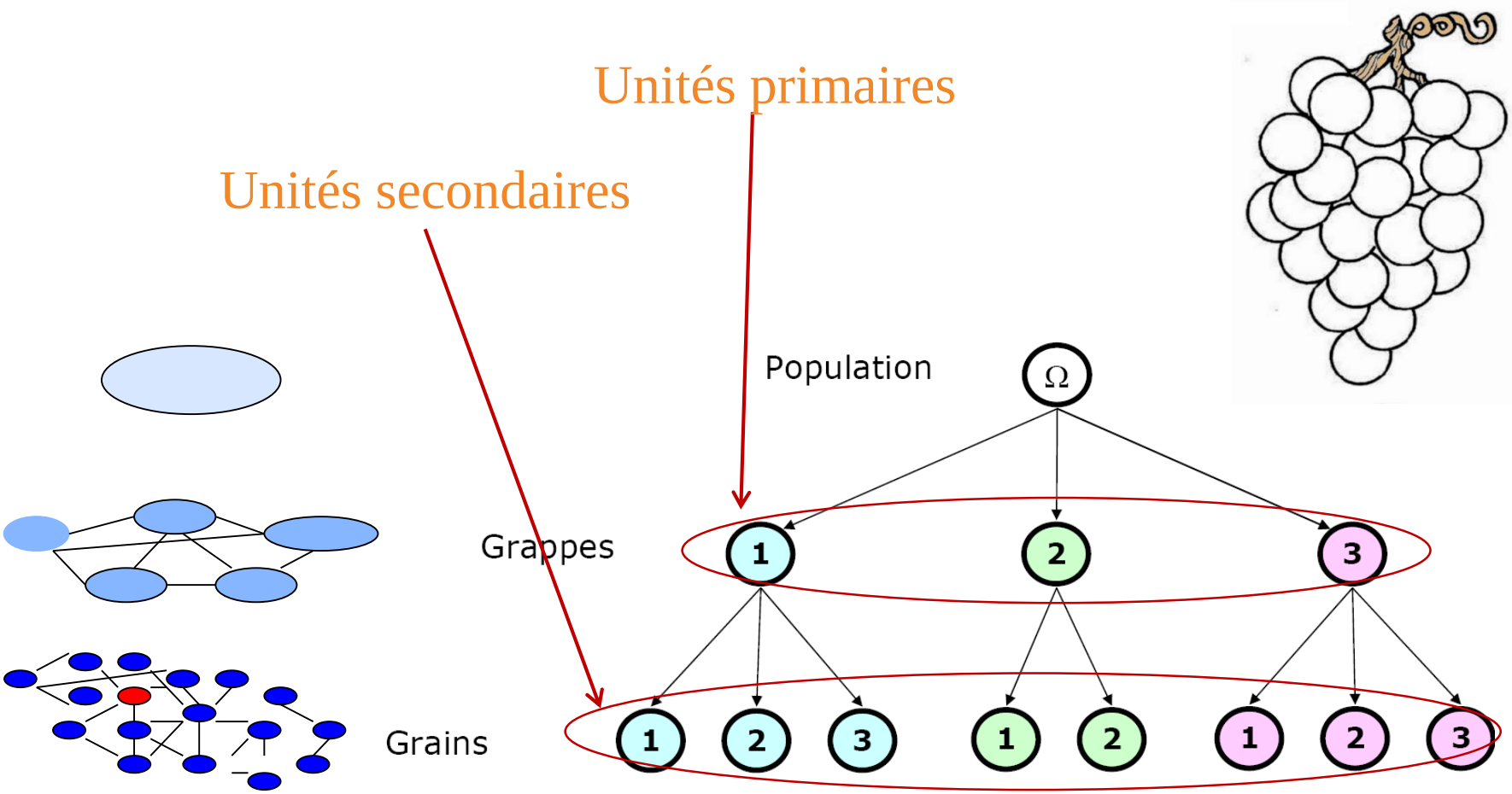
Finalement :

$$\sigma^2_{\bar{\bar{x}}} = \sum_{h=1}^{m} w_h^2 . \sigma_h^2 = (0{,}5^2 . 0{,}6886) + (0{,}3^2 . 2.3280) + (0{,}1^2 . 27{,}3829) + (0{,}1^2 . 214{,}019) = 2{,}7957$$

L'intervalle de confiance est :

$$\bar{\bar{y}} - 2.\hat{\sigma}_{\bar{\bar{y}}} \leq \mu° \leq \bar{\bar{y}} + 2.\hat{\sigma}_{\bar{\bar{y}}} \Rightarrow \mathbf{10{,}2859 \leq \mu° \leq 16{,}9741}$$

GRAPPES


Unités primaires
Unités secondaires
Population
Ω
Grappes
Grains
1
2
3
1
2
3
1
2
3
1
2
3

EXEMPLES DE GRAPPES

Reproduction de la chouette Effraie en Côte d'Or.

- Visite systématique des clochers pour recensement des nichées.
- Une fraction des clochers abritant une nichée a été tirée au hasard.
- Tous les jeunes (grains) des nichées sélectionnées (grappes) ont été pesés.

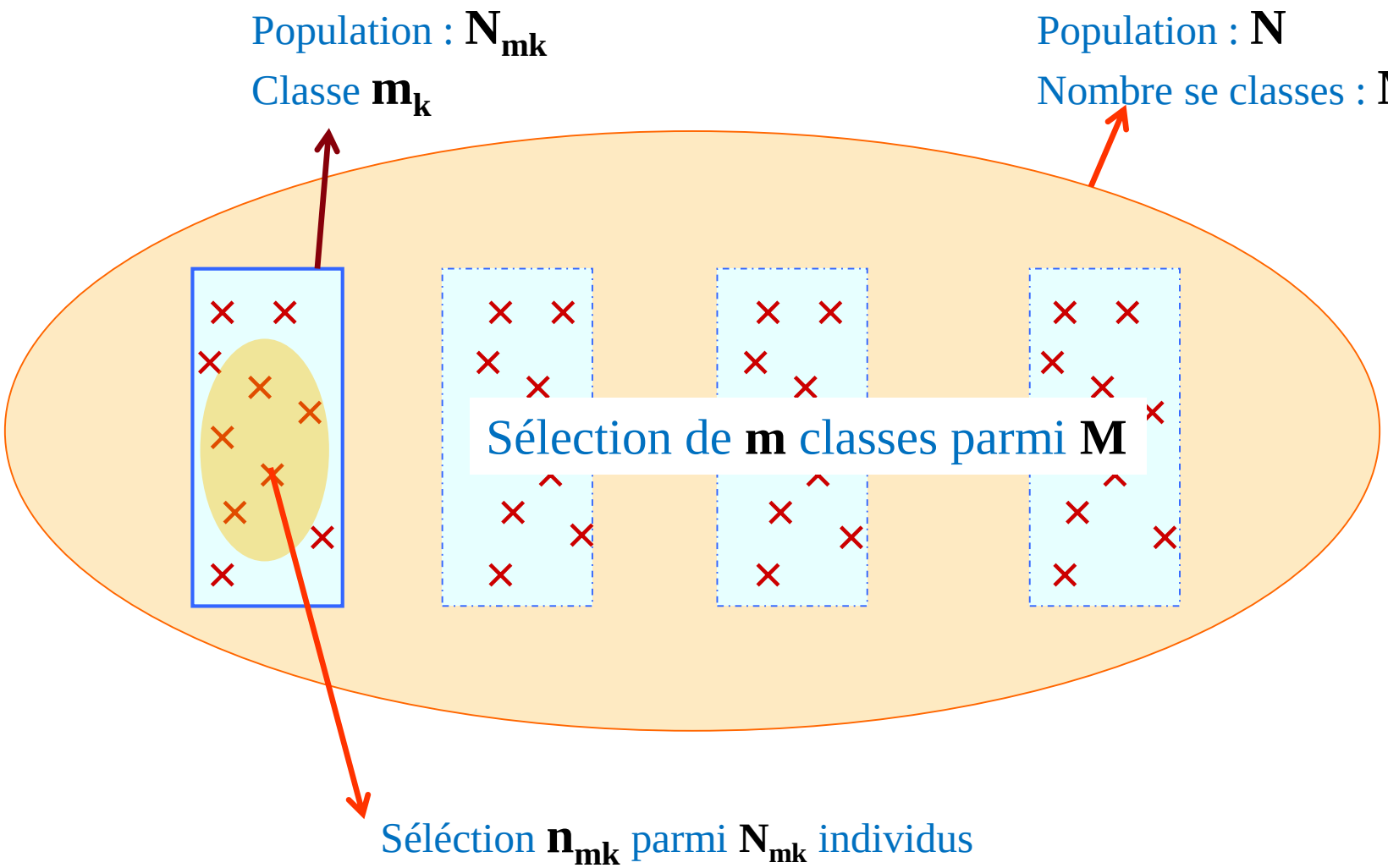
Etude une caractéristique (ex: attaque par un parasite) de la population de bovins de moins d'un an dans une région.

- Sélection au hasard de cheptels dans cette région
- Identification de la valeur de la caractéristique étudiée sur tous les bovins de moins d'un an

Les plans par regroupement

- Plan multiniveaux (ou à degrés) (multistage sampling)
- Echantrionnage (*Rank Set Sampling*)

POUR RÉSUMER



ABRÉVIATIONS ÉTUDIÉES

Nombre total d'individus : N
Nombre total de classes : M
Nombre de classe sélectionnées : m
Identification d'une des classes sélectionnée, indice : mk
Population de la classe mk : N_{mk}
Taille de l'échantillon dans la classe mk : n_{mk}
Proportion d'individus sélectionnées dans la classe mk : f_{mk}
Estimation de la variance expérimentale d'une classe mk : $s^2_{dans} = s^2_{mk}$
Estimation de la moyenne d'une classe mk : $\bar{x}_{mk}$
Estimation de la variance de la moyenne d'une classe mk : $s^2_{\bar{x},mk}$
Estimation de la variance expérimentale à partir des classes sélectionnées : s^2_M
Estimation de la moyenne des classes sélectionnées : $\bar{\bar{x}}_m$
Estimation de la variance de la moyenne des classes sélectionnées : $s^2_{\bar{M}}$
Estimation de la moyenne des classes sélectionnées pondérée par le nombre : $\bar{x}_m$
Estimation de la variance à partir des moyennes pondérées : $s^2_{entre} = s^2_m$
Estimation de la moyenne totale : $\bar{\bar{x}}$
Estimation de la variance sur la moyenne totale : $s^2_{\bar{\bar{x}}}$

DESCRIPTION INTERNE D'UNE CLASSE (MICROSCOPIQUE)

Moyenne d'une classe m_k: $\bar{x}_{mk} = \frac{1}{n_{mk}} \cdot \sum_{i=1}^{n_{mk}} x_{i,mk}$

Écart type **expérimental** de la classe m_k: $s^2_{mk} = \frac{SSW}{n_{mk} - 1}$

Écart type sur la **moyenne** de la classe m_k : $s^2_{\bar{x},mk} = (1 - f_{mk}) \frac{s^2_{mk}}{n_{mk}}$

Moyenne des écarts types obtenus pour chaque classe (permet d'évaluer l'importance de l'hétérogénéité qu'il y a dans chacun d'eaux) dans le cas où les classes ont toute la même taille n

$$s^2_{dans} = \frac{SSW}{m.(\bar{n} - 1)}$$

Avec $SSW = \sum_{i=1}^{n} \left(x_{i,mk} - \bar{x}_{mk} \right)^2$

DESCRIPTION EXTERNE, ENTRE LES CLASSES (MACROSCOPIQUE)

Moyenne **non pondérée**

Avec : $SSB = \sum_{mk=1}^{m} \left(\bar{x}_{mk} - \bar{\bar{x}}_{m}\right)^2$

Moyenne : $\bar{\bar{x}}_m = \frac{1}{m} \sum_{k=1}^{m} \bar{x}_{mk}$

Écart type **expérimental** : $s^2_{entre} = s^2_M = \frac{SSB}{m-1}$

Écart type sur la **moyenne** : $s^2_{\bar{M}} = \frac{(1 - F_m)}{m.(m-1)} SSB$

*Paramètres sur la population "class" (moyenne **pondérée** par le nombre)*

Moyenne : $\bar{x}_m = \frac{1}{m} \sum_{k=1}^{m} N_{mk} . \bar{x}_{mk}$

Écart type expérimental : $s^2_{entre} = s^2_m = \frac{1}{m-1} . \sum_{mk=1}^{m} \left[\left(N_{mk} . \bar{x}_{mk} \right) - \bar{x}_m \right]^2$

SYNTHÈSE DES DEUX ANALYSES

ECRAT TYPE expérimental

VARIANCE $_{\text{expérimentale}}$ **TOTALE =**
MOYENNE DES VARIANCES (expérimentales des classes) **+ VARIANCE** $_{\text{expérimentales}}$ **DES MOYENNES** (des classes)

- La moyenne des variances c'est : $s^2_{dans} = \frac{SSW}{m.(\bar{n}_m - 1)}$
- La variances des moyennes c'est : $s^2_{entre} = s^2_M = \frac{SSB}{m-1}$

$$s^2_{tot} = \frac{SSB}{m-1} + \frac{SSW}{m.(\bar{n}_m - 1)}$$

ECRAT TYPE sur la moyenne totale

$$s^2_{\bar{\bar{x}}} = \left[\left(\frac{M}{N}\right)^2 .(1-F_m).\frac{s^2_{entre}}{m}\right] + \left[\frac{1}{N^2}\frac{M}{m}\sum_{mk=1}^{m} N^2_{mk}.(1-f_{mk}).\frac{s^2_{m_k}}{n_{mk}}\right]$$

SÉLECTION

Simplification : classes et unités sélectionnées de même taille, respectivement m et n

Nombre total d'unités sélectionnées

- Pour la taille des classes $N_{mk} = \overline{N} = \frac{N}{M}$

- Pour la taille de la sélection dans les classes $n_{mk} = \overline{n} = \frac{n}{m}$

$$s^2_{\overline{\overline{x}}} = \left[\frac{1}{\overline{N}^2}.(1-F_m).\frac{s^2_{entre}}{m}\right] + \left[\frac{1}{M.\overline{n}}.(1-\overline{f}).s^2_{dans}\right]$$

Par ailleurs, pour donner une forme plus simple à cette équation, on pose :

$$S^2_{entre} = \frac{s^2_{entre}}{\overline{N}^2} \quad \text{et} \quad S^2_{dans} = \frac{s^2_{dans}}{M}$$

$$s^2_{\overline{\overline{x}}} = \left[(1-F_m).\frac{S^2_{entre}}{m}\right] + \left[(1-\overline{f}).\frac{S^2_{dans}}{n}\right]$$

COMPARAISON AVEC UN PLAN PUREMENT ALÉATOIRE

Taille de l'échantillon : $n = \bar{n}.m$

Conséquences : $s^2_{\bar{\bar{x}},alea} = \left(1 - \frac{m.\bar{n}}{N}\right).\frac{s^2_{tot}}{m.\bar{n}} = \left(1 - \frac{m.\bar{n}}{M.\bar{N}}\right).\frac{s^2_{tot}}{m.\bar{n}}$

Dans le cas où la population est importante : $s^2_{\bar{\bar{x}},alea} = \frac{s^2_{tot}}{m.\bar{n}}$

En considérant l'équation : $\rho = \frac{S^2_{entre}}{S^2_{total}} = \frac{S^2_{entre}}{S^2_{entre} + S^2_{dans}}$

Ce qui permet d'évaluer l'efficacité du plan $D_{eff} = \frac{s^2_{\bar{\bar{x}}}}{s^2_{aléa}} = 1 + (n-1).\rho$
(sous réserve des hypothèses approximations faites)

GÉNÉRALISATION À 3 NIVEAUX

Estimation de la moyenne

$$x = \frac{1}{n_i.n_j.n_k} \sum_i \sum_j \sum_k x_{ijk}$$

Estimation de la variance

$$\sigma_{\bar{x}} = \frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_1.n_2} + \frac{\sigma_3}{n_1.n_2.n_3}$$

Optimisation des couts

$$C_t = C_0 + n_1 C_1 + n_1.n_2 C_2 + n_1 n_2 n_3 C_3$$

MÉTHODOLOGIE

C'est un échantillonnage en 2 niveau, il faut donc choisir un premier plan
Technique intéressante quand le prix de la mesure en laboratoire est élevée

1. Sélection aléatoire de m^2 unités du domaine d'étude
2. Regroupement aléatoirement de ces m^2 unités en m ensembles de m unités
3. Faire du rangement dans chacun des m groupes : à chacune des m unités est attribué une place i
4. Sélectionner des unités rangées suivant la règle :
   dans le groupe i, on ne sélectionne que l'unité située à la place i
   (les autres sont laissées)

A la fin de cette opération, on dispose de m unités rangés suivant un place i

Pour avoir une statistique, il est nécessaire d'avoir plusieurs unités pour une place donnée i

$\Rightarrow$ Nécessité de recommencer la procédure r fois

Ce qui donne un nombre total d'unités de $n = m.r$

ESTIMATION DE LA MOYENNE ET DE SA VARIANCE

On distinguera le plan équilibré (même nombre d'individus sélectionnés dans chaque lieu) et non-équilibré

Dans le cas d'un tel plan (ce qui est souvent le cas) la moyenne et sa variance sont estimées par les expressions suivantes

$$\bar{x}_{Rank} = \frac{1}{r.m} . \sum_{i}^{m} \sum_{j}^{r} x_{i,j}$$

$$Var(\bar{x}_{Rank}) = \frac{1}{r(r-1)} \sum_{j}^{r} \frac{1}{m^2} \sum_{i}^{m} (x_{i,j} - \bar{x}_i)^2$$

$$\bar{x}_i = \frac{1}{r} \sum_{j}^{r} x_{ij}$$

EXEMPLES

La taille des marres d'une région

La reproduction des saumons est liée à la taille des marres

Sélection aléatoire sur la taille des segments de rivières

Les pâturages

La surface d'un terrain contaminé en plomb avec utilisation d'un XRF sur le terrain

Deux paramètres sont nécessaires pour dimensionner son plan :

Le rapport C et la précision relative RP :

$$C = \frac{\text{Cout de la mesure au laboratoire}}{\text{Cout du rangement des unités sur place}}$$

$$RP = \frac{\text{Variance}_{\text{aléatoire}}}{\text{Variance}_{\text{RSS}}}$$

On peut montrer que :

$$1 < RP < \frac{m+1}{2}$$

RSS = Plan aléatoire

Gain obtenu avec un RSS

PARAMÈTRES INITIAUX À DÉTERMINER

Pour bien dimensionner un plan RSS, il faut :

1. Avoir une idée de la loi statistique de la distribution de la caractéristique étudiée

2. Le nombre d'échantillons $\mathbf{n_0}$ voulus (en général dimensionné initialement sur la base d'un plan aléatoire)

3. Une estimation du coefficient de variation **CV** (essais pilote avec 10, mesure, bibliographie, expertise, ….)

4. Se fixer un nombre **m** de classes (évalué au jugement en prenant en compte les contraintes du « terrain »

DÉTERMINATION DE RP OU DE M À PARTIR DU CV

Valeur du RP pour une loi Lognormale (proche de la loi normale pour des CV petits)

Taille m	CV = 0.1	CV = 0.3	CV = 0.5	CV = 0.8
2	1.5	1.4	1.4	1.3
3	1.9	1.8	1.7	1.5
4	2.3	2.2	2.0	1.8
5	2.7	2.6	2.3	2.0
6	3.1	2.9	2.6	2.2
7	3.6	3.3	2.8	2.4
8	3.9	3.6	3.1	2.5
9	4.3	3.9	3.3	2.7
10	4.7	4.3	3.6	2.9

(pour un plan équilibré)

A partir des informations précédente, on détermine le nombre de cycle r :

$$r = \frac{n_0}{m} . \frac{1}{RP}$$

Ce qui permet d'ajuster le nombre d'échantillons : $n = r.m$

EXEMPLE

- A partir de l'hypothèse d'un échantillon aléatoire on détermine qu'il faut 25 échantillons de sol
- Un étude analogue a montré que la CV ≈ 0,4
- On décide de se fixer m = 5

Dans l'hypothèse d'une distribution normale, on détermine à partir de la table que RP ≈ 2,45

On en déduit : $r = \frac{n_0}{m} . \frac{1}{RP} = \frac{25}{5} . \frac{1}{2,45} = 2,04$ arrondi à 3

La taille réelle de l'échantillon est donc : n=3.5 = 15

(alors que pour le même résultat, il aurait fallu 25 échantillon avec un plan aléatoire)