FORMATIONS AUX STATISTIQUES : MANUEL DU FORMATEUR DE NIVEAU 1

Janvier 1991

SCÉNARIO 3

Echantillonnage

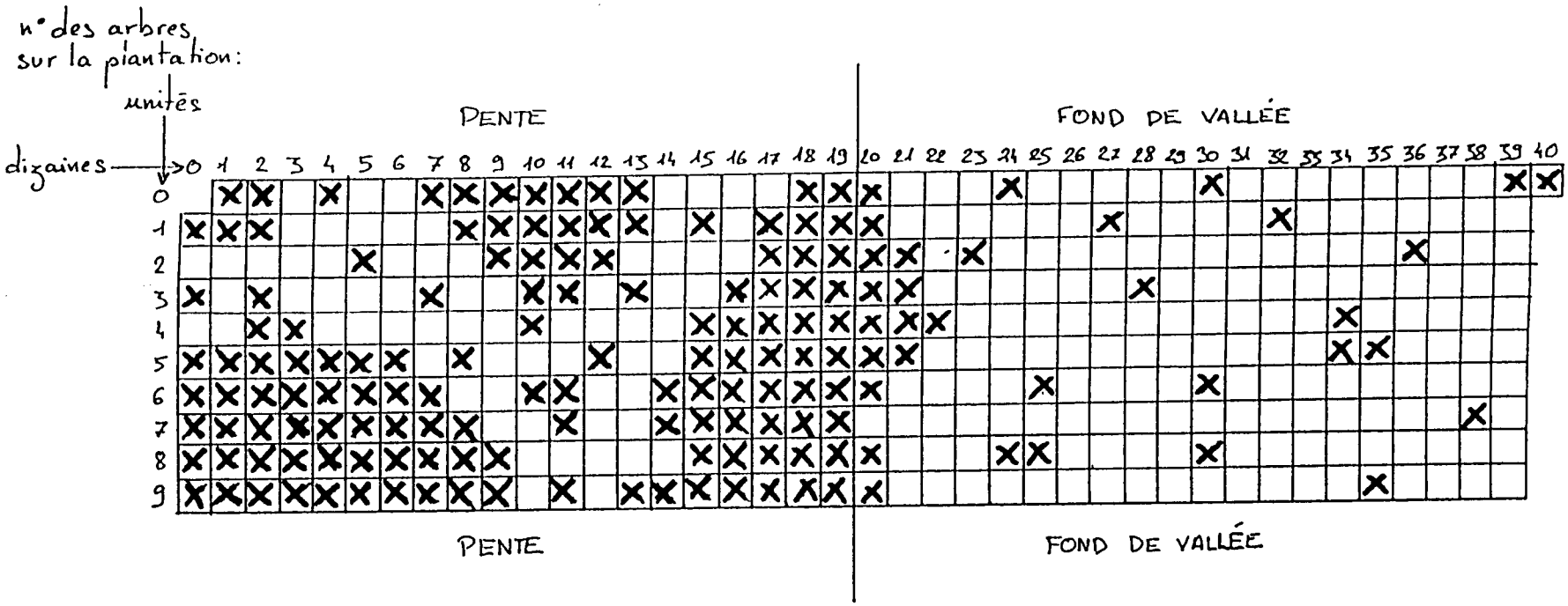
INRA

DOCUMENTS "FORMATEUR"

DÉPARTEMENT DE BIOMÉTRIE

formation permanente

plan du verger



Légende :

☐ arbre faiblement infesté

☒ arbre fortement infesté

ZONE DE PENTE :

71 arbres faiblement infestés (note de 1 à 3)
128 arbres fortement infestés (note de 4 à 5)

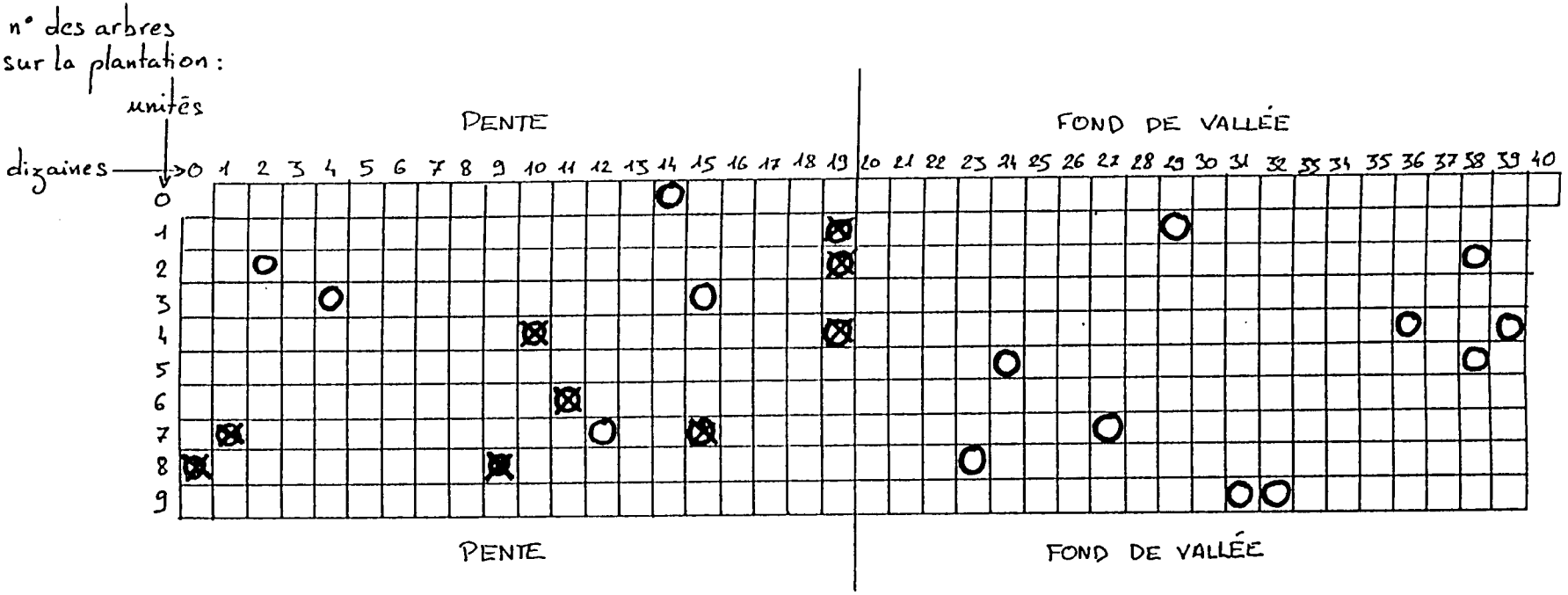
ZONE DE FOND DE VALLÉ :

168 arbres faiblement infestés (note de 1 à 3)
33 arbres fortement infestés (note de 4 à 5)

ENSEMBLE DU VERGER :

239 arbres faiblement infestés (note de 1 à 3)
161 arbres fortement infestés (note de 4 à 5)

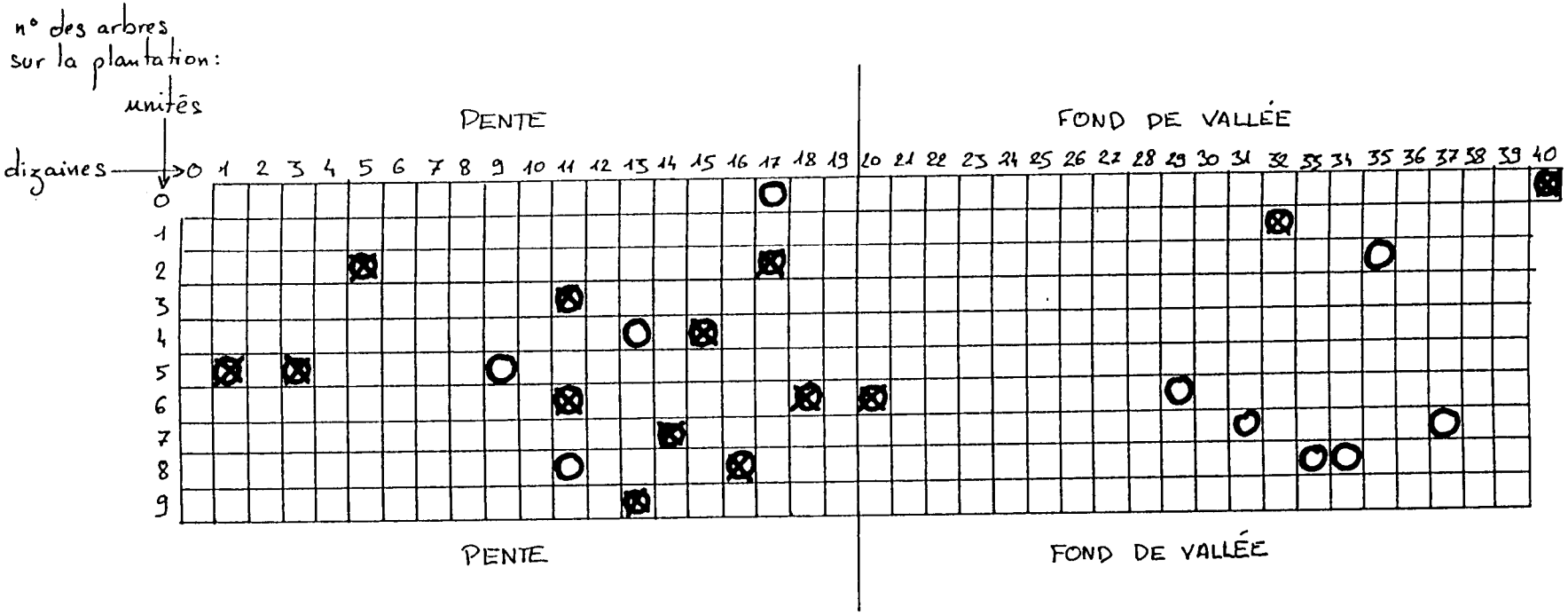
tirage aléatoire simple



Légende :

arbre non échantillonné □
arbre échantillonné faiblement infesté ○
arbre échantillonné fortement infesté ⊗

Exemple de tirage aléatoire simple ⟶ décision de ne pas traiter

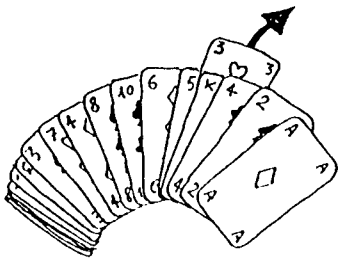


Légende :

arbre non échantillonné
arbre échantillonné faiblement infesté
arbre échantillonné fortement infesté

Exemple d'échantillon aléatoire simple $\longrightarrow$ décision de traiter l'ensemble

Notion de probabilité



Jeu de 52 cartes :

quatre AS
2
3
.
.
10

quatre valets
dames
rois

Tirer au sort
au hasard
1 carte

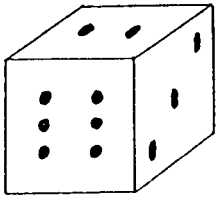
Quelle chance
probabilité
de tirer 1 AS de ♡, 1 ◇?

Transparent 1 :

On dispose d'un jeu de 52 cartes, constitué de 4 As,..., 4 rois.

Quelle est la probabilité de tirer un AS de ♡?
Le jeu contient 52 cartes, et ne contient qu'un AS de ♡, c'est donc 1 / 52.
Quelle est la probabilité de tirer un ◇?
Le jeu contient 52 cartes, dont 13 ◇ (AS, 2, 3,..., valet, dame, roi).
C'est donc 13 / 52.

Notion d'équiprobabilité



1 dé - 6 faces
1 lancer de dé

Quelle **probabilité** d'obtenir le 1 ?
Quelle **probabilité** d'obtenir le 6 ?

⇒ Chaque numéro a la **même probabilité** d'être obtenu.

numéro	1	2	3	4	5	6
probabilité	1/6	1/6	1/6	1/6	1/6	1/6

Probabilité = 1 / 6

Transparent 2 :

On dispose d'un dé, à 6 faces, supposé non pipé.

- 1 / 6 (le dé a 6 faces, un seul est marqué d'un 1).
- 1 / 6

C'est le raisonnement (et surtout l'hypothèse "dé non pipé" qui nous amène à dire que la probabilité d'obtenir 1, 2, 3, 4, 5 ou 6, est la même; elle vaut 1 / 6.

On voudrait expérimenter, et estimer la probabilité d'obtenir 1, 2, 3, 4, 5 et 6.

Notion d'estimation

1 dé - 6 faces

6 lancers

numéro obtenu	1	2	3	4	5	6
fréquence observée	1	1	1	1	1	1

1 expérience : 6 lancers

numéro obtenu	1	2	3	4	5	6
fréquence observée						
probabilité estimée						

$\Downarrow$

estimer la probabilité d'obtenir 1, 2, 3,. . . ,6.

Probabilité estimée $\neq$ probabilité réelle (1 / 6)

Transparent 3 :

Pour cela, il faudrait lancer le dé 6 fois.
Obtiendrait-on ce résultat? Certainement pas!
Ce n'est pas impossible, mais c'est très peu probable.

On expérimente : 6 lancers
Remplir la ligne *fréquence observée* au fur et à mesure.
On peut estimer la probabilité d'obtenir 1, par le nombre de fois où on a effectivement obtenu 1, divisé par 6 (le nombre de tirages).

Les probabilités estimées sont différentes des probabilités réelles (égales à 1 / 6).

Notion de précision d'un estimateur
Notion de précision d'une estimation

Comment améliorer cette situation ?

Numéro obtenu	1	2	3	4	5	6
6 lancers						
60 lancers						
60 000 lancers	1/6	1/6	1/6	1/6	1/6	1/6

$\Downarrow$

On "attrape" la probabilité réelle.

6 lancers $\longrightarrow$ 1 estimateur
($\uparrow$ *plus précis*)
60 lancers $\longrightarrow$ 1 estimateur
($\uparrow$ *plus précis*)
60 000 lancers $\longrightarrow$ 1 estimateur

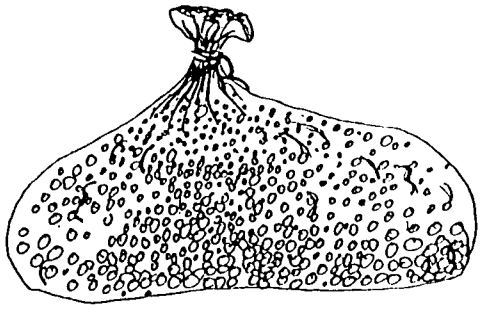
Transparent 4 :

On a vu que la probabilité estimée diffère de la probabilité réelle, 1 / 6.
Comment améliorer notre estimation?
$\Longrightarrow$ augmenter le nombre de lancers.

Dans le cas de 60 000 lancers, la probabilité estimée est très proche de la probabilité réelle (1 / 6).
On attrape la probabilité réelle.
L'estimation est très précise.

L'estimation faite avec 60 lancers est plus précise que l'estimation faite avec 6 lancers.
L'estimation faite avec 60 000 lancers est plus précise que l'estimation faite avec 60 lancers.
A 6 lancers, on associe un estimateur.
A 60 lancers, on associe un estimateur.
A 60 000 lancers, on associe un estimateur.
Le deuxième est plus précis que le premier,
le troisième est plus précis que le deuxième.

Notion de biais de l'estimation



← grains légers
← grains lourds

Problème : estimer le poids de 1000 grains

Tirer 1000 grains

en haut	au fond
grains légers	grains lourds

Estimation biaisée

Pourquoi?

Les grains n'ont pas la même probabilité d'être tirés.

en haut :	en bas :
plus forte probabilité	plus forte probabilité

Transparent 5 :

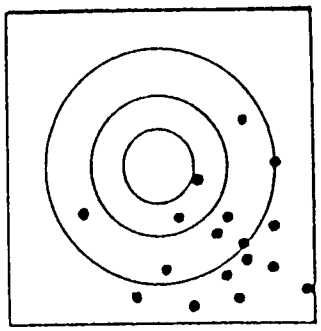
On dispose d'un sac de grains. Le sac est manipulé, secoué
$\Longrightarrow$ les grains lourds se retrouvent en bas.

Le problème posé : estimer le poids de 1000 grains.
2 possibilité au moins :

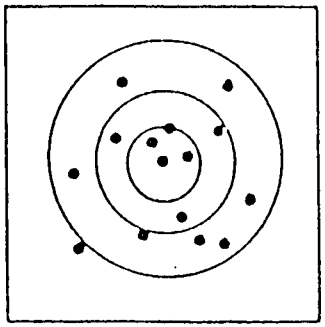
- ouvrir le sac et tirer 1000 grains.
- percer le fond du sac et tirer 1000 grains.

1er cas $\Longrightarrow$ grains légers $\Longrightarrow$ poids de 1000 grains sous-estimé.
2ème cas $\Longrightarrow$ grains lourds $\Longrightarrow$ poids de 1000 grains surestimé.
Dans les deux cas, on obtient une estimation biaisée.
Pourquoi?
Parce que chaque grain n'a pas pas eu la même probabilité d'être tiré :
Dans le 1er cas, les grains du haut ont eu une plus grande probabilité d'être tirés.

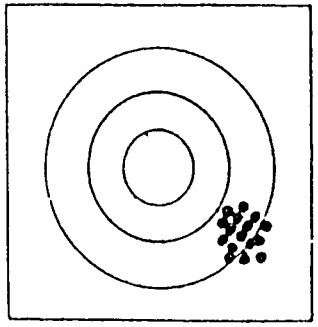
Notion de précision de l'estimateur
Notion de biais de l'estimateur



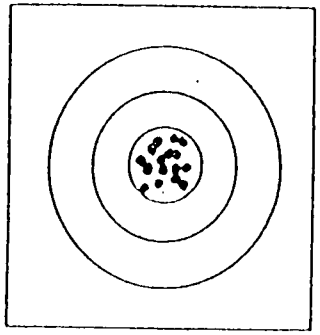
1o) estimateur :
précis - peu précis
biaisé - sans biais



2o) estimateur :
précis - peu précis
biaisé - sans biais



3o) estimateur :
précis - peu précis
biaisé - sans biais



4o) estimateur :
précis - peu précis
biaisé - sans biais

Transparent 6 :
Chaque schéma représente un estimateur (on définissait 1 estimateur avec 6 lancers, un autre estimateur avec 60 lancers,...).
Au centre de chaque cible, se trouve la vraie valeur (inconnue) à estimer.
Chaque point représente une estimation :
par exemple, l'estimation de la probabilité d'obtenir un 1, lorqu'on a fait 6 lancers de dés.

Quels sont les estimateurs :
- précis - peu précis
- biaisé - sans biais? (rayer)

Réponses :

1o)
- peu précis
- biaisé

2o)
- peu précis
- sans biais

3o)
- précis
- biaisé

4o)
- précis
- sans biais

Dans le cas d'un estimateur biaisé, on a un écart systématique entre la vraie valeur et les estimations.

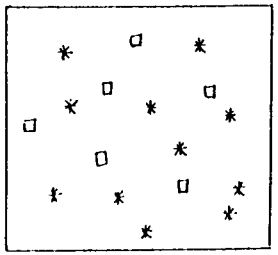
On cherche des estimateurs sans biais, les plus précis possible.

Population = ensemble d'unités statistiques

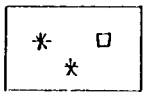
Procedure d'échantillonnage :

- extraire 1 échantillon de la population
- observer certaines caractéristiques

$\Longrightarrow$ en déduire des valeurs concernant l'ensemble de la population



Population



Echantillon

Transparent 7 :

On appelle population statistique, l'ensemble des unités statistiques auxquelles on s'intéresse.
La population statistique ne fait pas forcément référence à une population biologique : cela peut être :

- l'ensemble des électeurs d'un pays.
- l'ensemble des ponts de pyrale d'un champ.
- l'ensemble des arbres d'une forêt.

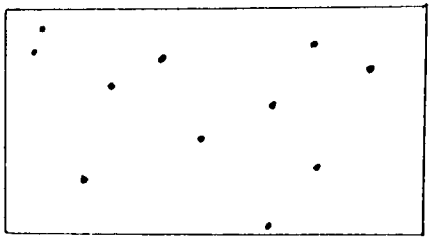
...

$\Longleftarrow$ tout individu de la population peut figurer dans l'échantillon.

Echantillonnage aléatoire simple

Prélèvement d'unités statistiques :

- au hasard
- avec la même probabilité
- de façon indépendante



exemple d'échantillonnage aléatoire simple

1 champ

Avantages :

- conception simple
- estimateurs no biaisés

Inconvénients :

- nécessité d'identifier chaque élément de la population
- protocole de sélection des unités pas toujours aisé
- précision faible

Transparent 8 :

Cela consiste en un prélèvement d'unités statistiques :

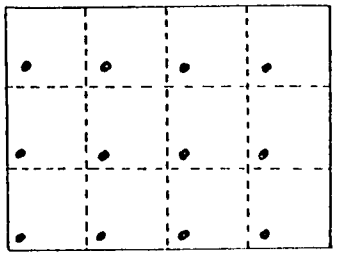
- au hasard
- avec la même probabilité pour toutes les unités
- de façon indépendante

La nécessité d'identifier chaque élément de la population peut poser des problèmes :
par exemple dans le cas d'un échantillonnage d'animaux sauvages, on ne sait pas toujours les identifier tous.
Le protocole de sélection des unités n'est pas toujours aisé :
par exemple, dans le cas de notre sac de grains, il ne suffit pas d'ouvrir le sac et de "tirer au hasard" les grains. Ce ne serait pas un échantillonnage aléatoire simple.
La précision est faible, car on n'utilise pas l'information qu'on pourrait avoir sur la population.

Echantillonnage systématique

- n : taille de l'échantillon
- population : n sous-ensembles de même taille t
- tirer au hasard i, entier, entre 1 et t
- prendre la i-ème unité de chaque sous-ensemble

$\Longrightarrow$ échantillon



n = 12

Avantages :

- commodité
- approprié à l'étude de répartition spatial
- évite de prendre des individus proches

Inconvénients :

- il faut découper la population en sous-ensembles de même taille
- problèmes si variations périodique

Transparent 9 :

Pour pratiquer un échantillonnage systématique, il faut :

- fixer la taille n de l'échantillon
- diviser la population en n sous-ensembles de même taille t
- tirer au hasard un entier compris entre 1 et t $\Longrightarrow$ i
- prendre la i-ème unité de chaque sous-ensemble

Les unités qui constituent l'échantillon sont régulièrement espacés.

- commodité de préparation et d'exécution : une fois qu'on s'est donné n et t, on peut échantillonner.
- approprié à l'étude de répartition spatiale.
- la procédure évite de prendre des individus proches, qui se ressemblent; elle évite donc la redondance d'information (ex : relevés météo faits *sur des stations rapprochées* $\Longrightarrow$ redondant).
- découper la population en sous-ensembles de même taille peut poser des problèmes dans certains cas :
  ex : étude d'une population sauvage, qui se déplace.
- problèmes si le pas du sondage est voisin de la période : l'échantillonnage systématique est moins bon (efficace) que l'échantillonnage aléatoire simple.

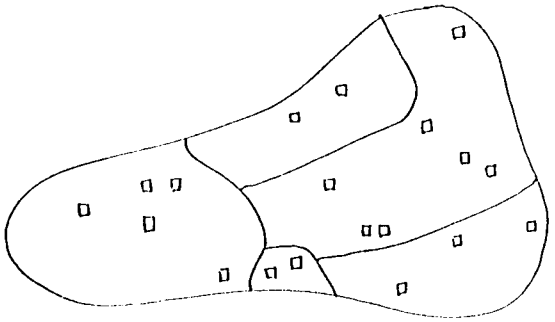
Echantillonnage aléatoire stratifié

région à étudier — domaines homogènes

$\Longrightarrow$

population — strates

Dans chaque strate, échantillonnage aléatoire simple.



Avantages :

- gain en précision,
  surtout si :
  - variabilité inter-strate forte
  - taille d'échantillon dans chaque strate bien choisie.

Inconvénients :

- l'information pour la stratification peut être inadéquate.

Transparent 10 :

La région à étudier (ou la population) est divisée en domaines (ou strates) aussi homogènes que possible (vis à vis du caractère étudié).
Dans chaque strate, on pratique un échantillonnage aléatoire simple (dont la taille est déterminée en fonction de l'*importance* de la strate vis à vis du phénomène étudié; importance donc au nombre ou à autre chose).

- la stratification apporte un gain de précision, surtout si les strates sont bien choisies, c'est à dire si la variabilité inter-strate est forte (et si la taille de l'échantillon de chaque strate est bien choisie).
- si les strates sont mal choisies, la stratification est inadéquate, les strates ne sont pas homogènes, et la variabilité inter-strate est faible.

Donc attention à une mauvaise stratification.
Cependant, même une stratification rudimentaire peut entrainer un gain de précision appréciable ($\Rightarrow$ préférable à l'échantillonnage aléatoire simple, si on est sûr d'une stratification, même rudimentaire).

Si on ne sait rien

$\Longrightarrow$ **échantillonnage aléatoire simple**
estimateur sans biais,
mais de faible précision

$\Diamond$

Si étude de répartition spatiale

$\Longrightarrow$ **échantillonnage systématique**

$\Diamond$

Si on a une connaissance de la population,
Si on peut la diviser en strates homogènes

$\Longrightarrow$ **échantillonnage aléatoire stratifié**
(plus précis que l'échantillonnage aléatoire simple)

Transparent 11 :

Tableau récapitulatif, simplifié, pour résumer.