FORMATIONS AUX STATISTIQUES : MANUEL DU FORMATEUR DE NIVEAU 1

Janvier 1991

SCÉNARIO 3

Echantillonnage

INRA

DOCUMENTS "STAGIAIRE"

DÉPARTEMENT DE BIOMÉTRIE

formation permanente fp

Scénario 3

ECHANTILLONNAGE

1 Présentation générale

1.1 Les étapes

1. Notions de probabilité
   équiprobabilité
   estimation
   précision de l'estimation
   biais de l'estimation
2. Prise de conscience de l'intérêt de l'échantillonnage.
3. Description des 3 principales stratégies d'échantillonnage, avec leurs avantages et inconvénients.
4. (*facultative*) Application des méthodes d'échantillonnage en histologie (peut être remplacé par l'éxemple de l'efficacité du vaccin contre la polio, ou un exemple d'échantillonnage en forêt).
5. Evaluation : exercice d'échantillonnage (verger de clémentiniers).

1.2 Matériel

1 dés à 6 faces.
Projecteur de transparents.
15 transparents.
Feutres à transparents.

1.4 La forme

- présentation orale et sur transparents.
- exercice par petits groupes : réalisation d'un échantillonnage.
- distribution d'un extrait du poly. : échantillonnage (de J. Vaillant).

1.5 Durée

3 heures.

1.6 Prérequis

Pas.

1.7 Evaluation

Par petits groupes pour l'exercice d'échantillonnage (étape 5).

TABLE DE NOMBRES AU HASARD

Pour tirer des nombres au hasard, à l'aide de cette table, il faut choisir, par un moyen quelconque, un point de départ, c'est-à-dire une ligne et une colonne, et prélever les chiffres au fur et à mesure de leur rencontre, dans un sens fixé à l'avance.

Exemple :

Si l'on veut tirer au hasard 5 objets dans une liste de 100, numérotés de 0 à 99, on commence par choisir :

- un point de départ : 15e ligne, 41e colonne par exemple,
- un sens de lecture : de gauche à droite (on peut aussi lire de droite à gauche, de haut en bas ...)

On obtient 56626 77422 ...

Si on prend des nombres de deux chiffres, on va donc prélever les objets numérotés 56, 62, 67, 74 et 22.

NOMBRES AU HASARD

	5	10	15	20	25	30	35	40	45	50
	13407	62899	78937	90525	25033	56358	78902	47008	72468	57949
	50230	63237	94083	93634	71652	02656	57532	60307	91519	48916
	84980	62458	09703	78397	66179	46982	67619	39254	90763	74055
	22116	33546	17545	31321	65772	86506	09811	82848	92211	51178
5	68645	15068	56898	87021	40115	27524	42221	86293	67592	06430
	26518	39122	96561	56004	50260	68648	85596	83979	09041	62350
	36493	41655	27871	71329	69212	57932	65281	57233	07732	58439
	77402	12994	59892	85581	70823	53338	34405	67080	16558	00854
	83579	97154	40341	84741	08967	73287	94952	59008	95774	44927
10	71802	39355	02981	89107	79788	51330	37129	31898	34011	43304
	57494	72484	22676	44311	15355	05348	03582	66183	68392	86844
	73364	36416	93128	10297	11419	82937	84369	88273	96010	09843
	14499	83965	75403	18002	45068	54257	18085	92525	60911	39137
	40747	03084	07734	86940	88722	85717	73810	79556	84853	66547
15	42237	59122	92855	62097	81276	06318	81607	00555	56526	77422
	32934	60227	56707	44858	36081	79981	01291	68707	45427	82145
	05754	14284	73059	80830	17231	42935	48472	18782	51646	37564
	32705	94879	93188	66049	25988	46656	35355	13800	83745	40141
	22190	27559	95568	53261	21575	98943	43518	42110	93402	93997
20	81516	15641	94921	95970	63505	22007	29956	38144	62556	07854
	26099	65501	69870	84445	58248	21282	56938	54729	67757	66412
	71874	61992	80001	21430	02305	59741	34252	15157	27545	14522
	05774	29589	42245	41903	69179	96652	91818	60812	47631	50609
	37294	93028	56850	63380	05912	29830	37512	15593	73198	99257
25	33912	37995	78967	57201	65915	73998	54269	07147	84313	51938
	63510	61475	25950	23804	54972	72055	19403	53755	04251	96022
	01570	41701	30262	54647	05077	29354	95704	75925	21811	88274
	24159	77787	38973	82178	46802	90245	01505	23906	95559	05755
	92834	52941	88301	22127	23459	40229	74578	21859	98645	72388
30	16178	60053	59284	16279	48003	44634	08523	32752	40472	05470
	81808	32950	80550	98391	62243	19578	39551	18398	35918	43543
	28528	52072	04954	52609	86508	68017	11120	28638	72850	03550
	62249	65757	12273	91251	96983	15052	83651	77582	81728	52157
	84541	99891	01585	96711	29712	02877	70955	59593	26828	98011
35	89052	39051	99811	69831	47234	93253	47385	17452	18874	74210
	43545	69232	00384	10858	21769	14093	06258	46450	97850	23490
	61518	19275	40744	22482	12424	98601	19089	53165	41835	28205
	68135	05545	04029	47945	19526	27122	42515	55043	23912	61105
	74005	34558	93779	96120	01695	47720	88545	73520	40050	90553
40	54437	88525	07943	81795	31709	13358	04525	64835	92133	44221
	01950	94752	89926	84764	19159	95355	98213	17704	47400	30637
	02404	42408	57981	43684	55467	47030	42545	43920	11199	35521
	59253	71535	26149	35629	87127	45581	00185	01041	48852	98897
	20471	13914	99330	37938	69649	57954	97149	41528	78854	80727
45	65946	60765	74084	22484	49514	89820	41310	19722	07045	28808
	00939	47818	75949	44707	49105	05777	31958	79942	98351	10265
	49952	29123	45950	67578	13524	03023	18046	75287	74989	58152
	17328	70732	46319	25950	19037	02831	35558	82712	05590	64941
	19420	70215	90476	76400	51553	12158	14568	15555	37595	94559
50	19121	41190	49145	05373	00755	17817	22757	76115	76977	94570

EXERCICE ECHANTILLONNAGE

Un verger de clémentiniers comprend 400 arbres. On peut définir une zone de pente (199 arbres) et une zone de fond de vallée (201 arbres).

On veut déterminer le degré d'infestion de ce verger par les pucerons, pour savoir s'il faut traiter sachant qu'il est recommandé de traiter quand 50% des arbres présentent une forte infestation (c'est à dire note 4 ou 5 sur l'échelle de 1 à 5).

De plus, il faut réaliser l'étude en deux heures, il faut 5 mm par arbre pour réaliser la notation et une seule personne peut le faire.

- A partir du plan ci-joint, où les arbres sont identifiés par leur numéro :
  1 à 199 : arbres de la zone de pente.
  200 à 400 : arbres de la zone de fond de vallée.

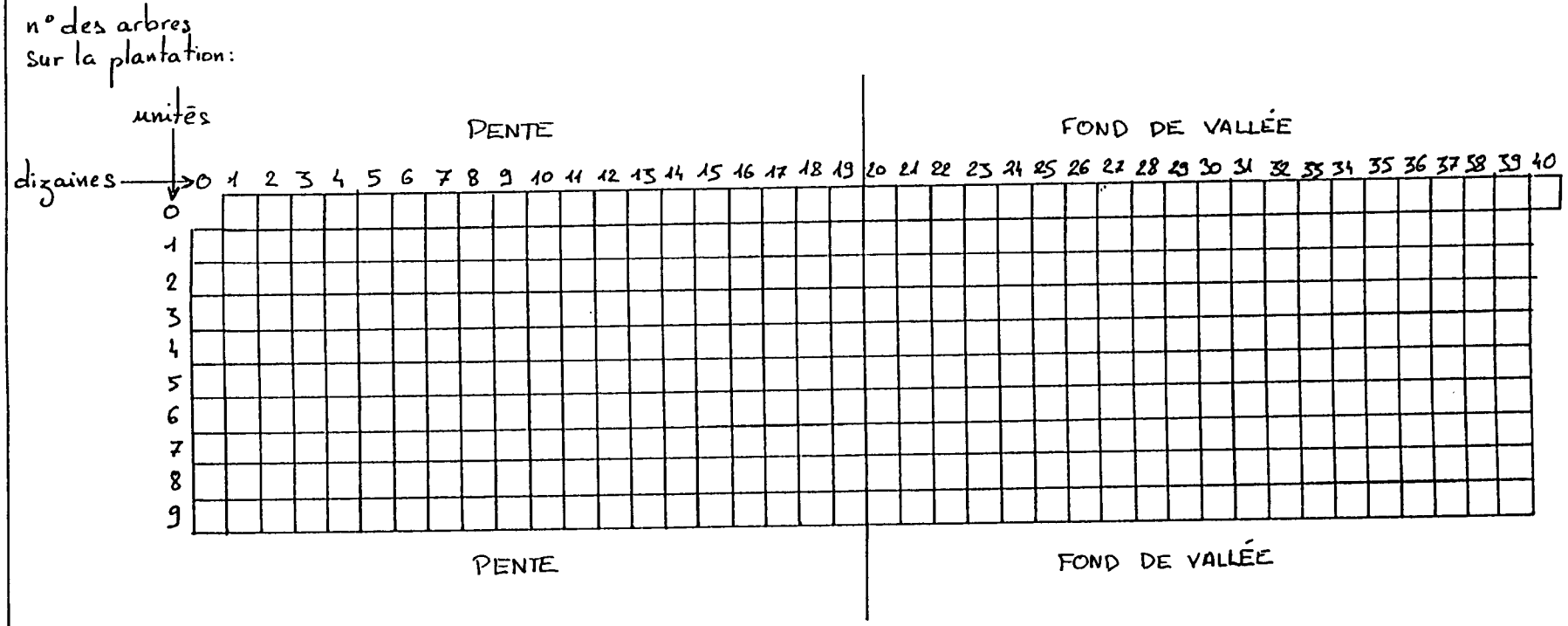
  Réaliser l'un des 3 trois types d'échantillonnage suivant :

  - aléatoire simple
  - systématique
  - stratifié

  Pour cela, il faut choisir le nombre d'arbres à échantillonner, et faire le tirage de leurs numéros.

- Quel est le résultat de votre étude?
  Quelle décision prenez-vous quant au traitement éventuel?

plan du verger



EXTRAIT DE :
PRATIQUE DE L'ÉCHANTILLONNAGE NOTIONS DE BASE ET ILLUSTRATIONS

GROUPE ECHANTILLONNAGE DE VERSAILLES
J. Vaillant, J.B. Denis et J.P. Rospars

Octobre 1989

Introduction

Un chercheur est parfois amené à prélever un échantillon de la population qu'il étudie. Ceci se produit lorsque les individus qui forment cette population sont trop nombreux, trop inaccessibles ou d'une façon générale trop coûteux à prélever, pour être examinés individuellement. On conçoit que, si certaines précautions sont prises, on pourra tirer des conclusions sur la population en se fondant sur l'échantillon qu'on en extrait. Pour que cette généralisation soit valide il faut que l'échantillon soit représentatif de la population, condition qui est remplie lorsque tout individu de la population à étudier peut figurer dans l'échantillon. En terme plus technique on pourra dire : un individu quelconque doit avoir une probabilité non-nulle, et si possible connue, de figurer dans l'échantillon.

Echantillonnage

Définitions

Population et Unités Statistiques

On emploie souvent le terme de populations en échantillonnage. Il s'agit de l'ensemble des unités statistiques auxquelles on s'intéresse. Cette population est statistique et ne fait pas forcément référence à une population biologique. Ce peut être l'ensemble des français ayant l'âge de voter. Ce peut être l'ensemble des pontes de pyrale d'un champ. Ce peut être l'ensemble des arbres d'une forêt, l'ensemble des champs de colza d'une région agricole,...

Procédure d'Echantillonnage

L'échantillonnage consiste à observer, sur un sous-ensemble de la population, certaines caractéristiques dans le but d'en déduire des valeurs concernant l'ensemble de la population échantillonnée.

Si on reprend les exemples précédents, il pourrait s'agir du parti pour lequel l'individu a voté, si la ponte est parasitée ou pas, s'il s'agit d'un conifère ou d'un feuillu, le rendement en quintaux par hectare.

L'échantillonnage consiste à définir la procédure de recueil des unités statistiques (individus de la population) qui seront observées. Cette procédure dépendra de l'objectif poursuivi et des contraintes techniques. Dans l'objectif est incluse la définition de la population à laquelle on s'intéresse. Cette définition précise est fondamentale, ce n'est pas la même chose de s'intéresser aux pieds de maïs d'un champ ou à ceux de l'ensemble des champs d'une région.

Echantillonnage aléatoire simple

Définition

Cette technique est la plus simple (au moins en principe) et la plus connue. Elle consiste à reconstituer les conditions de tirage d'une boule dans une urne. On prélève au hasard et de façon indépendante n éléments (ou unités d'échantillonnage) dans une population qui en comporte N. Dans ce cas la probabilité de tirage d'un élément quelconque est la même pour tous les éléments : elle est égale à n/N. Si l'échantillon est réduit à 1 élément cette probabilité est $1/N$; s'il est de taille N, la probabilité de tirage est 1, c'est-à-dire que tout individu est tiré (il s'agit alors d'un recensement).

Application pratique

Cette application n'est pas nécessairement immédiate. En effet si on veut assimiler la population à une urne avec tirage équiprobable, il faut, être capable d'identifier chaque élément de la population, c'est-à-dire le distinguer individuellement de tous les autres. Cette condition est en particulier remplie si on peut dresser la liste complète des éléments de la population : dans ce cas on les numérote, puis on en tire un échantillon de taille n à l'aide d'une table de nombres au hasard ou d'un programme ad hoc.

Exemple 1: Soit à étudier une colonie d'oiseaux comportant une centaine de nids (N). On peut en faire la carte et les numéroter. On en tire alors n pour étude approfondie. Pas de difficultés dans ce cas.

Exemple 2 : Soit à étudier un champ de maïs. Il peut être difficile d'établir une liste complète des pieds. Cependant les pieds sont identifiables ce qui permet de se mettre dans les conditions d'application du plan.

Exemple 3 : Soit à étudier la population de larves de moustique d'une mare. Pour dénombrer cette population on peut prélever des échantillons d'eau à l'aide d'une louche et compter le nombre de larves par louche. Pour se conformer au plan on tire au hasard la position de chaque prélèvement d'eau dans la mare.

Avantages et inconvénients

Le plan d'échantillonnage simple est en principe universel. Son caractère passe-partout résulte du fait qu'il n'exige aucune information préalable sur la population-cible étudiée. C'est aussi bien sûr son point faible. On dit que sa précision est faible

Pour appliquer ce plan il faut connaitre ou dresser la liste complète et sans répétition des éléments de la population. Or, dans beaucoup de cas, il s'avère difficile, voire impossible, de dresser une telle liste. Dans un échantillonnage d'animaux, par ex., il peut être très difficile de voir ou de capturer tous les individus d'une aire ou d'un volume donné pour des raisons de visibilité, de mobilité etc. Il en résulte un biais qui ne peut pas être corrigé par un procédé statistique, mais dont l'importance et surtout le sens (sur ou sous-estimation) peuvent être plus ou moins évalués.

D'autre part, selon la caractéristique étudiée, il peut s'avérer nécessaire que les éléments échantillonnés soient semblables entre eux en un certain sens (en taille, en poids, en accessibilité ...) pour des raison d'effort d'échantillonnage, mais surtout pour que la probabilité d'être dans l'échantillon soit constante d'un élément à un autre dans la population.

Echantillonnage systématique

Définition

Ce plan est une variante de l'échantillonnage aléatoire simple dans lequel seul le 1er élément de l'échantillon est tiré au hasard. Pour cela on divise la population statistique en n (n est la taille de l'échantillon) sous-ensembles de taille t. On a donc $N = nt$ (quand N n'est pas un multiple de n, l'approche est différente (Iachan, 1982; cf. p.51)). L'échantillon est constitué du ième élément de chacun des sous-ensembles, le nombre i étant tiré au hasard.

Sondage systématique = éléments régulièrement espacés

Application pratique

Ce plan s'applique de manière assez naturelle lorsque la population statistique étudiée est ordonnée. C'est le cas lors d'une étude d'évolution dans le temps ; t est alors la période de prélèvement. C'est encore le cas, dans l'espace cette fois, pour le lit d'une rivière, la lisière d'une forêt, un talus, une falaise etc. Chaque sous-ensemble a alors une certaine longueur t. Il s'applique également lorsque la population est à 2 dimensions, comme un champ, ou à 3 dimensions, comme un lac par ex. Les n sous-ensembles ont alors une surface de t mètres carrés ou un volume de t mètres cubes. Dans le cas d'une surface les points échantillonnés sont disposés de manière régulière.

Pour appliquer le plan systématique, comme pour appliquer le plan aléatoire simple, il faut être capable d'identifier les éléments de la population. C'est indispensable pour ne retenir qu'un seul élément par période t.

II.3.2.c. Avantages et inconvénients

Ce qui en fait distingue le plan systématique du plan aléatoire simple c'est la régularité des points échantillonnés. Quelles conséquences cela a-t-il ? Par rapport à l'échantillon que fournit un sondage ordinaire, celui obtenu par un sondage systématique est :

- plus représentatif si les éléments de la population sont autocorrélés ou bien présentent une tendance linéaire. Il y a autocorrélation positive si 2 éléments de la population sont d'autant plus semblables qu'ils sont plus proches l'un de l'autre (dans le temps ou dans l'espace). Le plan systématique est alors préférable à l'aléatoire simple.

Une autre forme de régularité de la population est celle où le caractère étudié varie linéairement avec le numéro d'ordre des éléments dans la série. Ce phénomène peut apparaître en écologie dans un gradient d'altitude, de distance à la côte, de hauteur de végétation, dans

l'évolution temporelle d'un phénomène etc. Dans ce cas encore le sondage aléatoire simple est moins précis.

- équivalent si les éléments échantillonnés sont disposés de manière aléatoire. L'avantage du plan systématique résulte alors de la commodité de sa préparation et de son exécution sur le terrain.

- moins représentatif si les éléments de la population apparaissent selon une séquence qui engendre des variations périodiques du caractère étudié, et si le pas du sondage est voisin de cette période. L'échantillonnage systématique est alors moins efficace que l'aléatoire simple.

Echantillonnage stratifié

Définition

L'échantillonnage stratifié consiste à subdiviser une population hétérogène en k sous-populations ou strates plus homogènes, non-chevauchantes et collectivement exhaustives. La somme des effectifs des strates est alors égale à l'effectif N de la population. Ensuite, au sein de chacune des strates, on tire un échantillon en appliquant un échantillonnage aléatoire simple. On distingue donc 2 niveaux :

- au 1er niveau on effectue un recensement (celui des strates)
- au 2nd niveau on effectue un échantillonnage simple (celui des individus dans une strate)

Avantages et inconvénients

Un échantillonnage avec stratification (correctement faite, voir ci-dessous) est toujours plus efficace qu'un sondage aléatoire simple. Même rudimentaire, la stratification peut entrainer des gains de précision appréciables.

Ce plan s'impose lorsque l'effort d'échantillonnage ne peut être maintenu constant pour des raisons pratiques (nuits, fins de semaines, vacances ...) ou liées au phénomène étudié (pullulation, pic de migration ...).

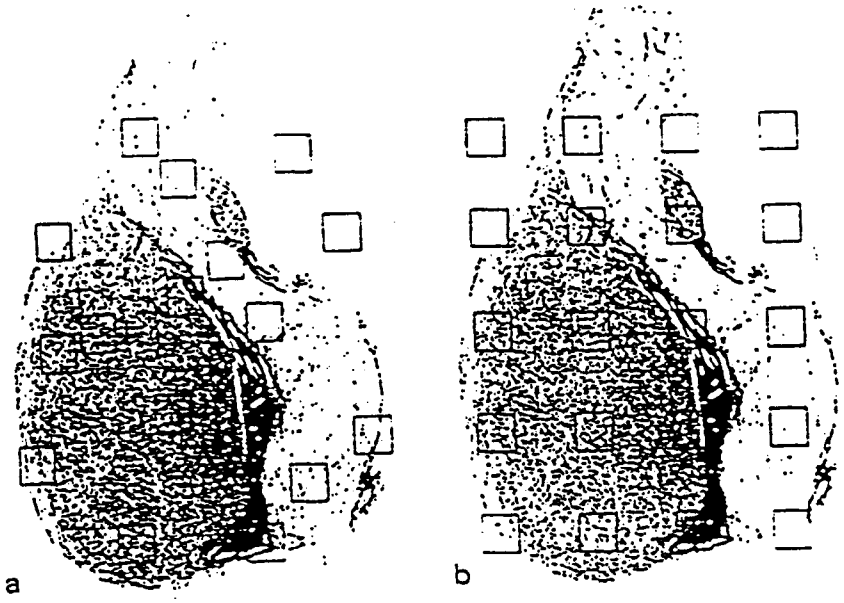


fig.1 : microscopie optique : coupe d'une hypophyse

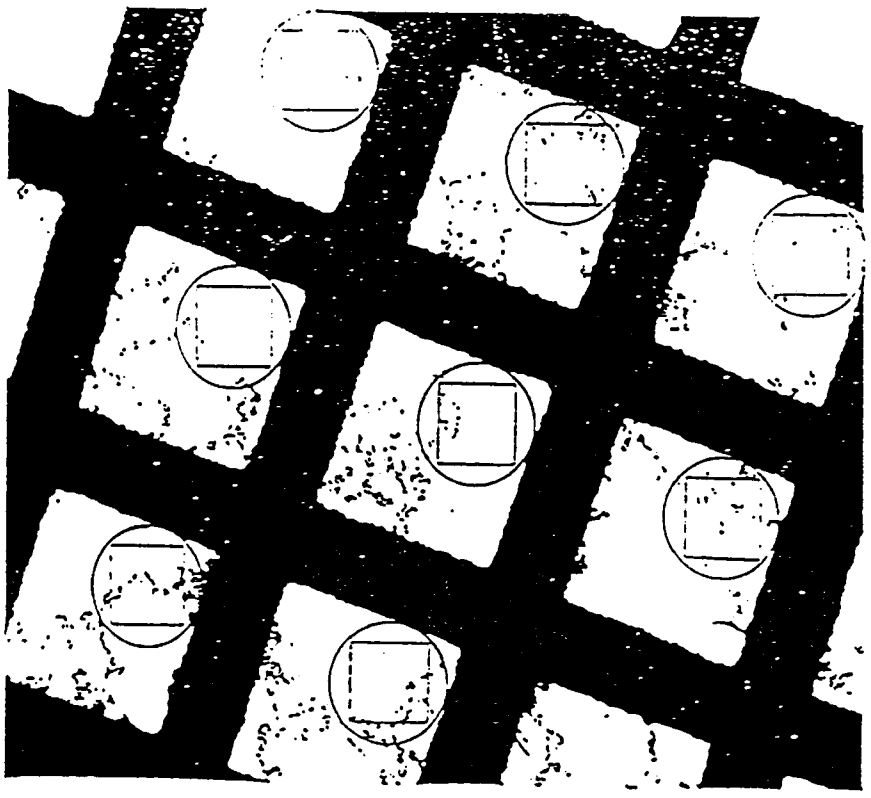


fig.2 : microscopie électronique : échantillonnage systématique

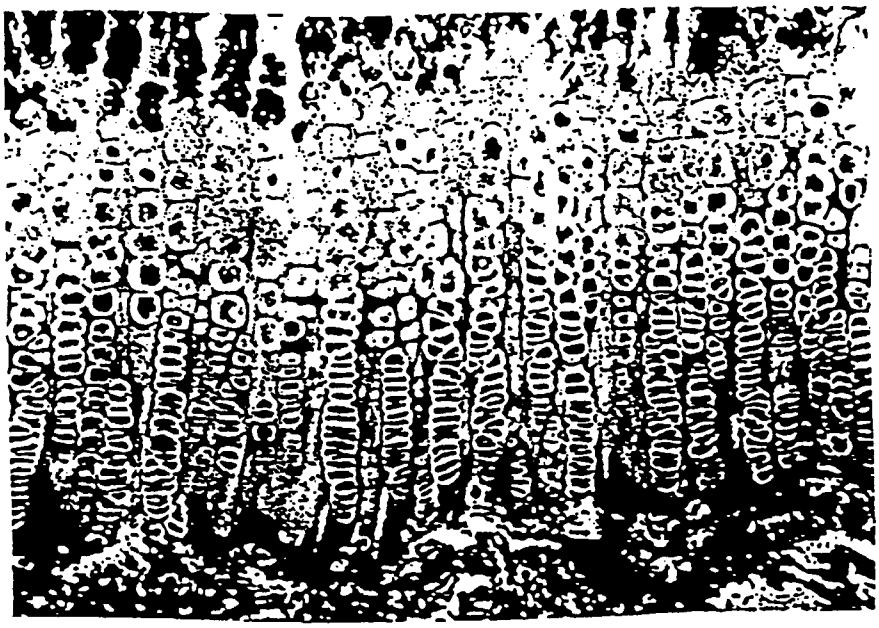


fig.3 : coupe dans le cartilage d'un os

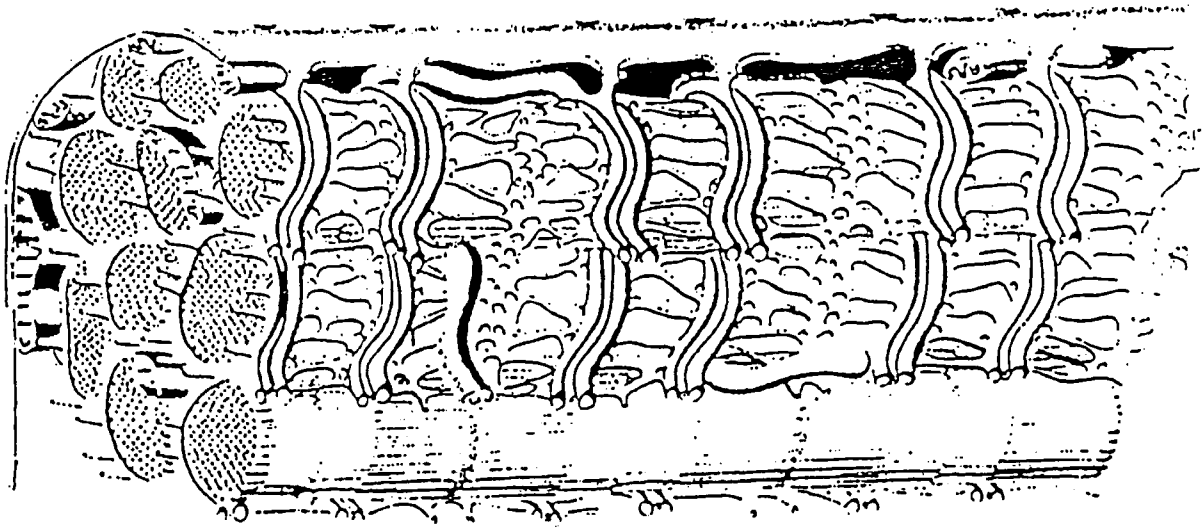


fig.4 : muscle du squelette

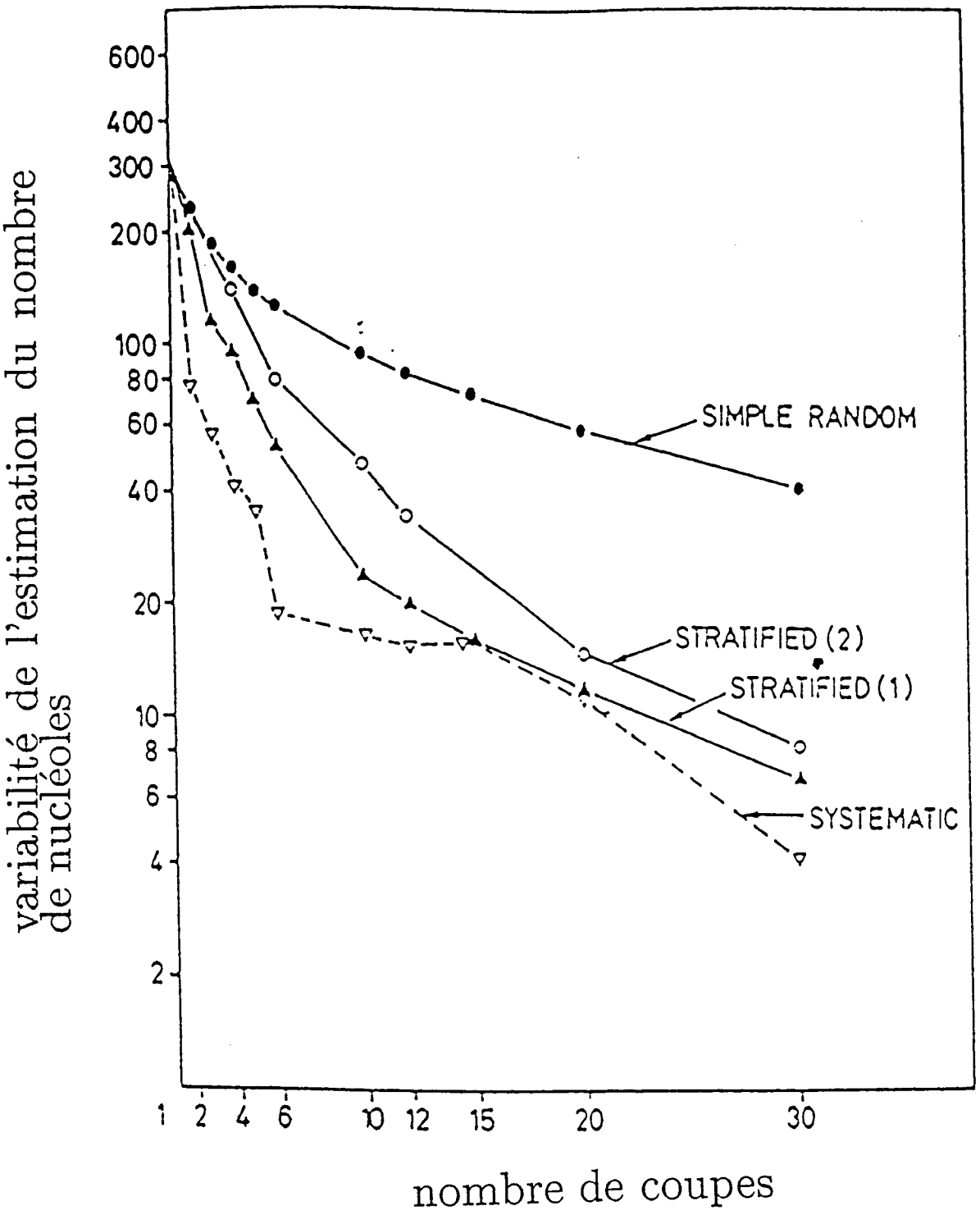


fig.5 : comparaison de la précision de quatre types d'échantillonnage

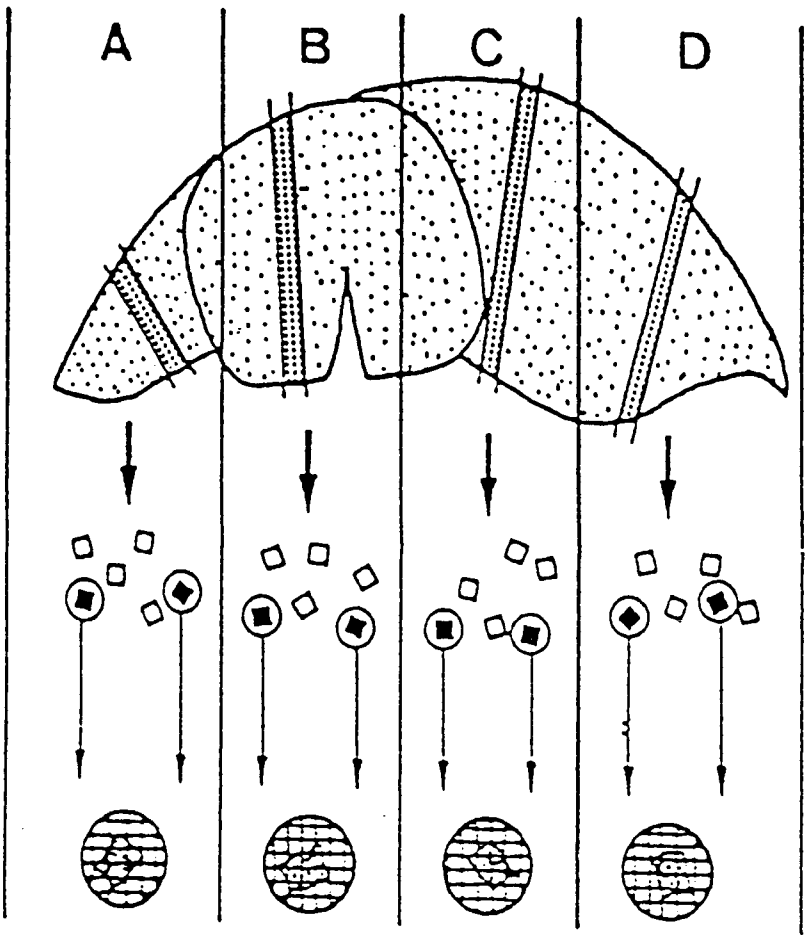


Fig. 6 : un organe, schématisation du mode d'échantillonnage

Scénario 3

VOCABULAIRE

Biais (bias) : le biais constitue une déviation par rapport à la vraie valeur qu'on veut trouver.

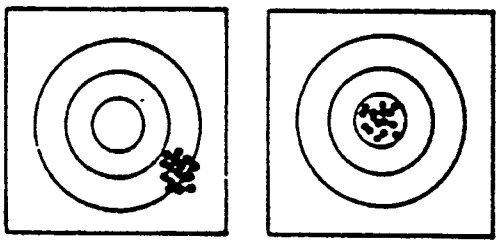


Figure a Figure b

Exemple : si le centre de la cible est la valeur qu'on veut trouver, la figure a présente un biais mais pas la figure b.

Echantillon (sample) : fraction représentative d'une **population.**

Echantillonnage aléatoire simple (random sampling) : il consiste à prélever au hasard et de façon indépendante une fraction de **population.** Cette fraction est appelée **échantillon.**

Echantillonnage aléatoire stratifié (stratified random sampling) : la population étudiée est divisée en strates. Dans chaque strate est réalisé un tirage aléatoire simple.

Echantillonnage systématique (systematic sampling) : les individus de la **population** sont prélevés selon un pas régulier ou une unité de quadrillage régulière.

Population (population) : une population est un ensemble d'individus ou d'unités.

Précision d'une estimation : une estimation a une bonne précision si elle n'est pas très différente de la valeur cherchée.