Notions de base en échantillonnage

(3ième Version, Août 1996)

Jean Vaillant

Université des Antilles-Guyane
Département de Mathématiques et Informatique
Campus Fouillole - 97169 Pointe-à-Pitre

Table des matières

Table des figures

Liste des tableaux

Introduction

Introduction

Ce document a pour but de présenter les concepts de base en théorie de l'échantillonnage. Sa rédaction a été guidée par le souci de faire connaître l'apport de la statistique dans le domaine de la mise en œuvre de protocoles d'observation, ceci pour diverses disciplines. En effet, il existe une littérature importante sur les différentes techniques de prélèvement de l'information dans une population à partir de l'étude d'une fraction de cette population (on parle, selon le domaine concerné, soit de techniques de sondage, soit de plans d'échantillonnage). Cependant, la littérature en langue française concerne le plus souvent les études de type socio-économiques et les ouvrages français se rapportant à l'échantillonnage dans le domaine écologique ou agronomique sont rares. On peut noter le livre publié en 1983 par Frontier dans la série *Collection d'écologie*.

Dans ce document, nous exposons d'abord les principes élémentaires de probabilités et statistique (Chapitre 1), puis les plans d'échantillonnage classiques (Chapitre 2). Cette présentation est accompagnée d'exemples concrets mettant en relief les avantages et inconvénients de ces plans d'échantillonnage, tout en illustrant les problèmes spécifiques au phénomène étudié. Au chapitre 3, nous présentons plus en détail l'échantillonnage systématique en raison de l'intérêt qu'il présente en écologie. Le chapitre 4 s'attache plus particulièrement au problème de la disposition spatiale des individus sur lesquels s'applique un échantillonnage tandis qu'au chapitre 5, notre préoccupation est de mettre en évidence les différences de base entre procédures d'échantillonnage à taille non fixée et procédures informatives.

Nous avons essayé autant que faire se peut de mettre l'accent sur l'incidence de la procédure d'échantillonnage sur l'interprétation de l'échantillon observé et de décrire les liens entre la collecte des données et leurs traitements statistiques. Notre objectif est donc d'attirer l'attention du lecteur sur ce domaine de la statistique par trop méconnu et de promouvoir l'utilisation éclairée des méthodes d'échantillonnage en fonction des objectifs recherchés.

L'utilisateur averti de statistiques pourra éviter le chapitre 1, et en parti-

culier le § 1.2. La terminologie statistique plus spécifique à la théorie de l'échantillonnage est, par contre, présentée au § 1.3. A la fin de chaque chapitre, une bibliographie pourra être consultée. Pour les deux premiers chapitres, elle est dépouillée de tout article trop pointu et chaque référence est accompagnée d'un petit commentaire indiquant succintement son contenu.

Chapitre 1

Notions de base en probabilité et échantillonnage

Plan :

1.1 Introduction

L'objet de ce chapitre est de donner les idées essentielles pour comprendre les techniques d'échantillonnage. Par la même occasion, le vocabulaire spécifique de cette partie de la statistique sera introduit sans formalisme. Le but poursuivi n'est pas tant de donner des formules mathématiques qui pourront être trouvées dans la littérature classique (Cochran, 1977; Jessen, 1978; Gourieroux, 1981; Droesbeke *et al.*, 1987; Grosbras, 1987) que de développer l'intuition du lecteur sur les notions fondamentales. Certaines distinctions, qui peuvent sembler du ressort du spécialiste, en particulier celles procédant de la théorie des probabilités, sont cependant indispensables pour bien appréhender la démarche.

Tout d'abord une petite précision : dans les domaines économique et sociologique, on emploie plutôt le terme de sondage que celui d'échantillonnage. Nous préférons ce dernier car son usage est consacré dans le domaine écologique ; en fait, nous considérons ces 2 termes comme étant équivalents.

La statistique s'appuie sur la théorie des probabilités. Nous commencerons donc par en évoquer les notions de base avant d'aborder les problèmes relatifs à l'échantillonnage.

1.2 Notions de probabilité et de statistique

1.2.1 Les variables aléatoires

Les phénomènes comportant des éléments d'incertitude sont dits aléatoires. Une variable aléatoire peut être considérée comme la formalisation mathématique d'événements dont on ne peut prévoir avec certitude la réalisation. Deux exemples classiques sont le résultat du jet d'un dé (variant de 1 à 6) ou encore le rendement d'une variété de blé dans des conditions précisées (qui ne peut être inférieur à 0 mais dont il est difficile de donner un majorant précis : supposons 120 quintaux/ha). Les figures FIG. 1.1 et FIG. 1.2 donnent une représentation graphique de ces deux exemples.

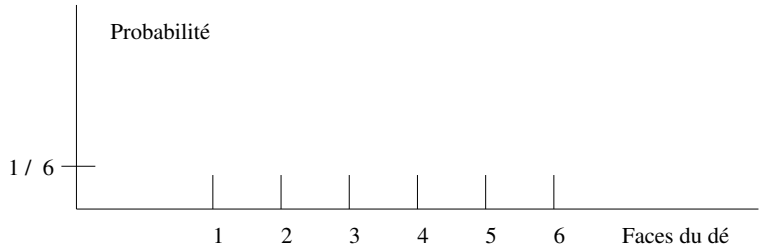


FIG. 1.1 - *Chacune des 6 faces du dé a une probabilité de 1/6 d'apparaître.*

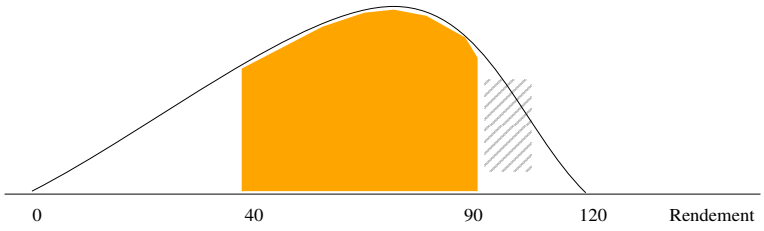


FIG. 1.2 - *Densité de probabilité de la variable aléatoire rendement supposé continu.*
La superficie totale sous la courbe vaut l'unité, celle de la région ombrée représente la probabilité d'un rendement compris entre 40 et 90 quintaux/ha.

Les 2 variables aléatoires associées à ces deux exemples sont de natures différentes : la première est discrète (elle prend des valeurs entières entre 1 et 6), la seconde est continue (elle prend n'importe quelle valeur réelle entre 0 et 120).

Dans la pratique, une variable continue est souvent observée de manière discontinue. Par exemple, un rendement n'est mesuré qu'au kilogramme près par unité de surface; une taille d'écolier n'est mesurée qu'au centimètre près. Ceci est dû à la précision limitée des appareils de mesure. Ainsi, un résultat de 4 253 kg/ha correspond en général au fait que le rendement est compris dans l'intervalle [4 252.5,4 253.5[; il s'agit donc bien d'une variable continue. Inversement,

on peut approcher une variable discrète par une variable continue lorsque le nombre de valeurs possibles pour cette variable discrète est très grand. Ces approximations peuvent être excellentes. Un exemple nous est donné par le nombre de pucerons sur un pied de maïs en période de fortes infestations, considéré comme une variable log-normale (Johnson et Kotz, 1970).

Les deux premiers exemples présentés (dé et rendement) sont des variables quantitatives, mais une variable aléatoire peut aussi être qualitative : le sexe d'un animal, la couleur d'une fleur, la catégorie socio-professionnelle d'une personne sondée.

1.2.2 Le concept de distribution de probabilité

La notion de probabilité peut se présenter de deux manières.

Dans le cas discret (comme celui du dé), il est habituel de la considérer comme le rapport du nombre de cas favorables sur le nombre de cas possibles. Ainsi, à chacune des valeurs possibles du dé est attachée une probabilité de 1/6 et la probabilité d'obtenir, par exemple, un nombre pair est $3/6 = 1/2$. Mais pour que ce calcul soit correct, il faut que les événements élémentaires (les cas) aient la même probabilité d'apparaître.

On peut aussi interpréter les probabilités comme des pondérations subjectives associées aux différentes valeurs de la variable aléatoire discrète et caractérisant le degré de *probabilité* de leur occurrence; ces pondérations doivent satisfaire quelques propriétés (axiomes du calcul des probabilités) pour mériter l'appellation de probabilités au sens mathématique. En particulier, la somme de toutes ces pondérations doit être l'unité, un événement impossible est de probabilité nulle, un événement certain est de probabilité unité.

Dans le cas continu (comme celui du rendement), on ne peut pas associer une probabilité à chaque valeur possible de la variable aléatoire. Les probabilités sont associées à des intervalles. Ces probabilités sont données par l'intégrale d'une fonction positive f appelée fonction de densité de la variable aléatoire (FIG. 1.2): la probabilité que la variable aléatoire appartienne à l'intervalle $[a, b]$ est :

$$\int_a^b f(x)dx.$$

On appelle distribution de probabilité d'une variable aléatoire discrète X l'ensemble des probabilités p_i associées aux valeurs possibles x_i de cette variable. On écrit :

$$P(X = x_i) = p_i \quad \text{où} \quad 0 \leq p_i \leq 1 \quad \text{et} \quad \sum_i p_i = 1.$$

Si X est une variable aléatoire continue, sa distribution de probabilité correspond à sa fonction de densité de probabilité f. On a :

$\int f(x)dx = 1$ où l'intégration est sur le domaine des valeurs possibles de X.

1.2.3 Les paramètres d'une distribution de probabilité

Lorsqu'une variable aléatoire est quantitative, 2 paramètres importants de sa distribution sont souvent pris en considération : son espérance mathématique (dénommée plus brièvement espérance) et sa variance. Il s'agit de 2 paramètres, c'est à dire de 2 valeurs fixes qui caractérisent, en partie, la distribution de la variable aléatoire. Si la variable aléatoire est nommée X, son espérance et sa variance se noteront respectivement $E(X)$ et $V(X)$.

Si $f(x)$ est la densité de cette variable aléatoire, alors l'espérance se définit par la formule :

$$E(X) = \int xf(x)dx,$$

l'intégration s'effectuant sur le domaine de définition de la fonction de densité f(x). Cette formule correspond exactement à celle d'un barycentre en mécanique, c'est donc une valeur centrale parmi l'ensemble des valeurs que peut prendre la variable aléatoire. Ceci se voit peut-être encore plus facilement dans le cas d'une variable discrète où l'intégration s'exprime par une sommation :

$$E(X) = \sum_i p_i x_i$$

où les x_i sont les valeurs possibles de la variable aléatoire de probabilité p_i.

Soulignons que l'espérance n'est pas une variable aléatoire, c'est un paramètre qui prend une valeur fixe, par exemple l'espérance de la variable aléatoire *résultat du jet de dé* vaut :

$$\frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = 3.5.$$

De manière générale, on définit l'espérance d'une fonction d'une variable aléatoire $\phi(X)$ par :

$$E(\phi(X)) = \int \phi(x)f(x)dx \quad \text{ou} \quad E(\phi(X)) = \sum_i p_i \phi(x_i),$$

C'est-à-dire qu'on considère $\phi(X)$ comme une variable aléatoire elle-même, ce qu'elle est effectivement. Son espérance s'interprète aussi comme une valeur centrale parmi l'ensemble des valeurs possibles de la variable aléatoire $\phi(X)$.

Pour caractériser la dispersion d'une variable aléatoire, on emploie l'espérance de *l'écart quadratique à l'espérance* ou plus simplement la variance :

$$V(X) = E[(X - E(X))^2].$$

Bien entendu, il s'agit encore d'un paramètre fixe de la distribution de la variable aléatoire X ; il est tout à fait analogue à l'inertie en mécanique. Dans le cas du jet de dé, la variance vaut :

$$\frac{1}{6}(1-3.5)^2+\frac{1}{6}(2-3.5)^2+\frac{1}{6}(3-3.5)^2+\frac{1}{6}(4-3.5)^2+\frac{1}{6}(5-3.5)^2+\frac{1}{6}(6-3.5)^2 \simeq 2.917.$$

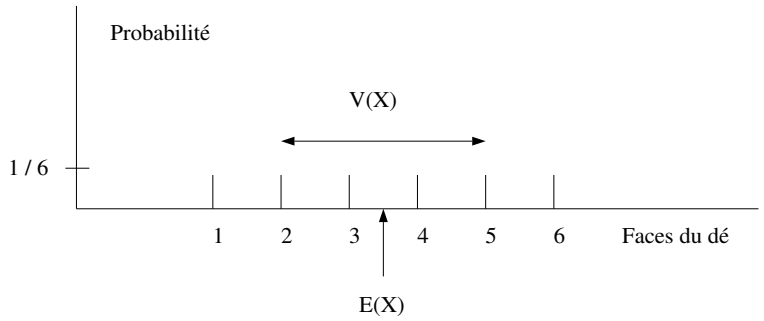


FIG. 1.3 - *Espérance et variance de la distribution du jet d'un dé.*

On peut noter que l'espérance n'est pas forcément une valeur que peut prendre la variable aléatoire : c'est le cas du dé. L'espérance n'est donc pas toujours la valeur la plus probable : c'est la problématique du *français moyen.*

D'autre part, la variance est une valeur non négative qui caractérise la dispersion de la variable aléatoire autour de son espérance. Plus $V(X)$ est importante, plus cette dispersion est grande. A la limite si $V(X) = 0$, la variable aléatoire ne prend qu'une seule valeur. Si l'espérance s'exprime dans la même unité que la variable aléatoire, la variance s'exprime dans le carré de cette unité; c'est pourquoi, on lui préfère parfois l'écart-type qui est la racine carrée de la variance.

1.2.4 Indépendance d'événements, de variables aléatoires

Une autre notion élémentaire capitale à saisir est celle de l'indépendance. Deux événements sont indépendants s'ils n'ont aucune relation de causalité, directe ou indirecte, entre eux. Dans ce sens l'indépendance est une notion opposée à celle de relation déterministe. Il n'est pas facile de donner une idée intuitive de cette notion qui soit valable dans tous les cas, aussi considérerons nous quelques exemples. Le jet de 2 dés fournit 2 variables aléatoires indépendantes, le résultat de l'un n'influe pas sur l'autre. Si l'on tire, *avec remise*, dans un

sac contenant des boules blanches et noires, la couleur de la boule tirée au *nième* tirage est indépendante des tirages précédents. Par contre, si l'on tire, *sans remise*, la probabilité du second tirage sera différente selon que le premier tirage aura été une boule noire ou une boule blanche. Ce dernier exemple n'est pas complètement théorique, car il se transpose immédiatement au domaine écologique dans le cas de prélèvement d'individus dans une population. Si la taille de la population est grande par rapport à la taille du prélèvement, on pourra faire l'hypothèse que les prélèvements successifs sont indépendants et de même loi, ce qui ne sera pas possible en cas contraire.

La situation d'indépendance est une situation beaucoup plus simple à traiter car la probabilité conjointe des réalisations de plusieurs variables aléatoires est, dans ce cas, simplement le produit des probabilités de chacune prise séparément. Dans le cas des dés, si l'on note $X(i)$ le résultat du jet du dé i, on a :

$$P[X(1) = 1 \text{ et } X(2) = 1] = P[X(1) = 1].P[X(2) = 1] = 1/36.$$

1.2.5 Quelques distributions de probabilité

Beaucoup de distributions sont utilisées en échantillonnage (Johnson et Kotz, 1969, 1970; Hastings et Peacock, 1974). Nous présenterons deux distributions continues : la distribution normale (ou de Laplace-Gauss) et la distribution uniforme, et trois distributions discrètes : la distribution binomiale, la distribution de Poisson et la distribution binomiale négative.

1.2.5.1 La distribution normale

L'importance de la loi normale tient au fait que c'est une loi approchée par de nombreux phénomènes naturels. En particulier si beaucoup de petites causes indépendantes s'additionnent, elles induisent une distribution normale. Cette distribution ne dépend que de 2 paramètres, on peut la caractériser par son espérance et sa variance ; elle est symétrique (FIG. 1.4).

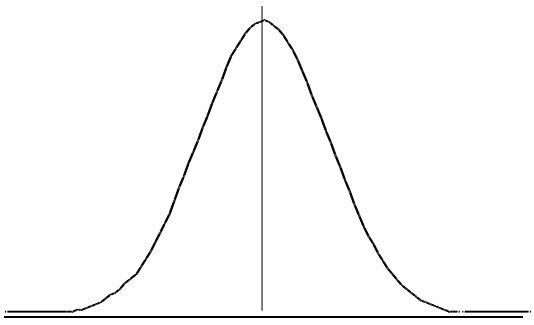


FIG. 1.4 - *Distribution normale.*

1.2.5.2 La distribution uniforme

La loi uniforme donne la même probabilité à tout intervalle de même longueur sur la portion de droite qu'elle couvre ; on peut donc la qualifier de distribution *au hasard* dans le sens où aucun point n'est plus probable qu'un autre. La distribution du jet de dé est aussi une distribution uniforme mais sur un ensemble de points. Dans le cas de la distribution de la figure FIG. 1.5 (définie sur $[a, b]$), l'espérance vaut $(a + b)/2$ et la variance vaut $(b - a)^2/12$. Cette loi se généralise aisément à une portion de plan, c'est en fait elle qu'on utilise implicitement lorsqu'on dit tirer au hasard un point.

FIG. 1.5 - *Distribution uniforme.*

1.2.5.3 La distribution binomiale

La distribution binomiale est une distribution importante en échantillonnage. Elle correspond au nombre de fois qu'on obtient pile en jetant N pièces de monnaie en l'air. C'est le même problème que d'examiner un ensemble d'éléments pouvant prendre 2 états (par exemple mâle ou femelle) en supposant que les probabilités associées aux deux états ne varient pas d'un individu à l'autre et surtout qu'il y a indépendance (voir ci-après). La distribution binomiale dépend de 2 paramètres : N le nombre d'essais (ou en l'occurrence le nombre d'individus examinés) et p la probabilité associée à l'un des 2 états. En général N est connu et p est le paramètre inconnu, mais N peut aussi ne pas être connu. Bien évidemment il s'agit d'une distribution discrète qui prend ses valeurs sur l'ensemble des entiers $0, 1, \ldots, N$. Si, comme dans le cas des couleurs de fleurs, le nombre des états est plus grand que 2, on aboutit à une distribution multinomiale dont la distribution binomiale est un cas particulier.

Si X suit une binomiale de paramètres N et p, on vérifie que :

$$P(X = n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

et

$$E(X) = Np, \qquad V(X) = Np(1-p).$$

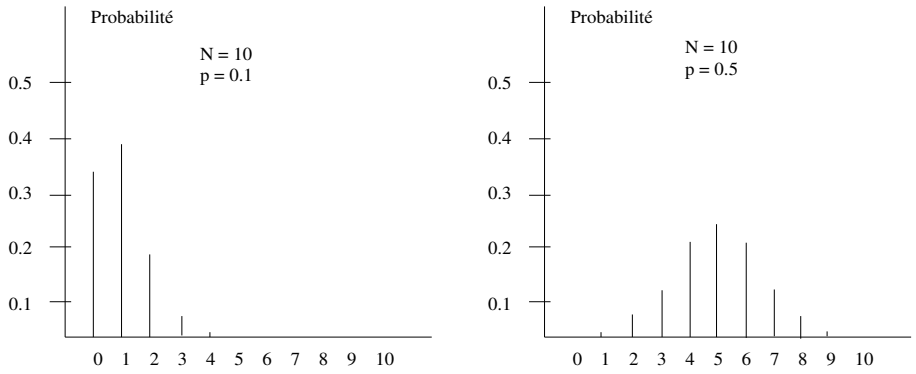


FIG. 1.6 - *Distribution binomiale.*

1.2.5.4 La distribution de Poisson

La distribution de Poisson joue pour les variables aléatoires discrètes un rôle semblable à celui de la distribution normale pour les variables continues. En particulier, elle est la limite vers laquelle tend la loi binomiale lorsque N tend vers l'infini alors que le produit Np (i.e. son espérance) reste constant. C'est elle qu'on retrouve dans les répartitions *au hasard* de points dans une portion de plan (processus Poissonniens). Elle ne dépend que d'un paramètre, classiquement noté λ. Si elle prend ses valeurs sur les entiers positifs (donc jusque l'infini), en réalité elle ne charge d'une manière non-négligeable qu'un ensemble fini de points comme le montre les 2 exemples de la figure FIG. 1.7.

Si X suit une distribution de Poisson de paramètre λ, alors :

$$P(X = n) = \frac{e^{-\lambda}\lambda^n}{n!} \qquad \text{pour } n \geq 0$$

et

$$E(X) = V(X) = \lambda.$$

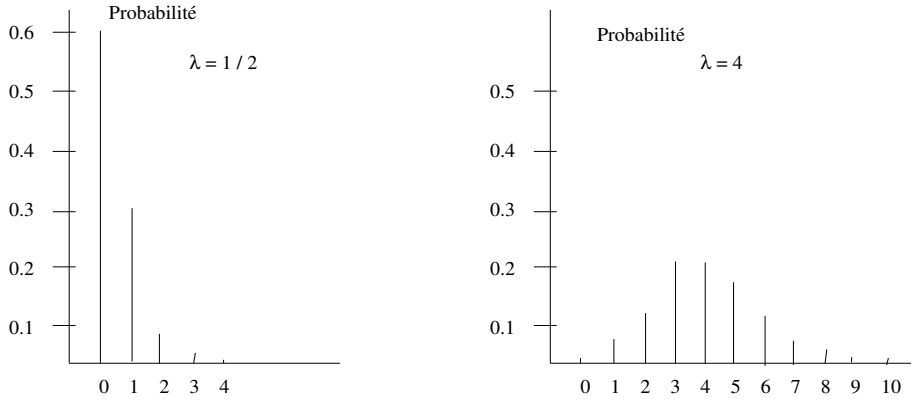


FIG. 1.7 - *Distribution de Poisson.*

1.2.5.5 La distribution binomiale négative

La distribution binomiale négative peut comme la distribution binomiale être présentée par la succession de piles et de faces engendrés par le jet d'une pièce de monnaie. Mais cette fois-ci la variable aléatoire n'est pas le nombre de piles obtenues pour N jets de la pièce, c'est le nombre de jets aboutissant à des faces avant l'obtention d'un nombre fixé à l'avance de piles, disons k. La loi binomiale négative dépend donc aussi de 2 paramètres : p (la probabilité d'obtenir pile lors d'un jet) et k le nombre de fois que le résultat pile doit être obtenu.

On se réfère souvent à cette distribution dans le cas de répartitions agrégatives de points dans le temps ou dans l'espace (Taylor, 1984) et le mécanisme d'obtention de fréquences de cette loi est alors tout autre que celui décrit plus haut (Pielou, 1977; Southwood, 1978).

Si X suit une distribution binomiale négative de paramètre p et k, alors :

$$P(X = n) = \frac{(n+k-1)!}{(k-1)!n!} p^k (1-p)^n$$

et

$$E(X) = k(1-p)/p, \qquad V(X) = k(1-p)/p^2.$$

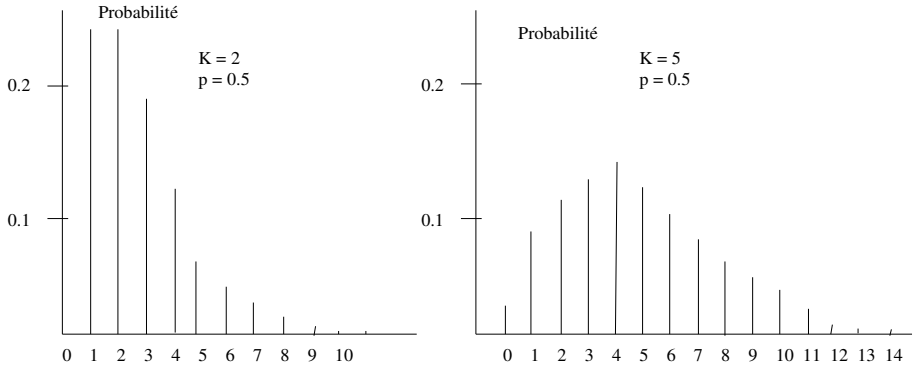


FIG. 1.8 - *Distribution binomiale négative.*

1.2.6 Quelques concepts liés à la statistique

1.2.6.1 Les estimateurs

Il faut bien insister sur le fait que ni l'espérance, ni la variance ne sont des variables aléatoires. Par contre ce sont des paramètres en général inconnus (contrairement au cas du dé qui nous a servi de premier exemple) et le travail du statisticien consiste à les estimer du mieux qu'il peut en fonction de l'information dont il dispose. Pour ce faire, il crée des estimateurs qui eux sont

des variables aléatoires. Aussi un estimateur a-t-il à son tour une espérance et une variance qui caractérisent sa distribution et vont donc servir pour juger de sa pertinence à estimer le paramètre choisi. Par exemple, dans le cas de l'observation de la réalisation n d'une loi binomiale de paramètre N connu et de paramètre p inconnu, n/N sera l'estimateur de p. D'autre part, la moyenne empirique obtenue à partir d'observations indépendantes d'une même loi normale sera l'estimateur de l'espérance de cette loi normale.

Il convient donc de bien distinguer *l'espérance* (qui est un paramètre d'une distribution de probabilité) de *la moyenne* que nous emploierons pour l'estimer (Dagnélie, 1973). Cette dernière est la moyenne arithmétique de plusieurs variables aléatoires qui suivent cette distribution. La moyenne, fonction de plusieurs variables aléatoires, est elle-même aléatoire et a donc une espérance. Certains auteurs pour mieux appuyer la distinction emploient la locution *moyenne empirique*. Malheureusement, le vocabulaire parallèle n'existe pas pour la variance. Nous emploierons variance quand il s'agira du paramètre et variance empirique lorsqu'il s'agira de l'estimation classique de la variance à savoir :

$$\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2,$$

où

$$\overline{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$$

est la moyenne de l'échantillon dont on dispose, et N la taille de cet échantillon.

En général, le contexte ne laisse pas de doute sur la variance qui est en cause, néanmoins, au risque de paraître nous répéter, nous insistons sur l'importance de la distinction : elle revient à séparer le modèle probabiliste utilisé du traitement statistique.

1.2.6.2 Biais, variance et précision

Si l'espérance de l'estimateur est la valeur du paramètre qu'on veut estimer, on le dit sans biais. Le biais d'un estimateur se définit comme la différence entre le paramètre et l'espérance de l'estimateur. Mais un estimateur peut être sans biais et très mauvais : en effet sa variabilité autour de l'espérance peut être très grande (voir FIG. 1.9). C'est pour cela, qu'en général on essaie d'obtenir des estimateurs sans biais qui soient de variance minimale.

Les estimateurs classiques de l'espérance et de la variance d'une distribution sont, lorsqu'on dispose de plusieurs de ses réalisations, la moyenne et la variance empirique. Mais d'autres estimateurs sont possibles : par exemple la médiane pour estimer l'espérance. Il est aussi important de distinguer le paramètre à estimer de son estimateur, que l'objectif qu'on poursuit, du moyen employé pour essayer de l'atteindre.

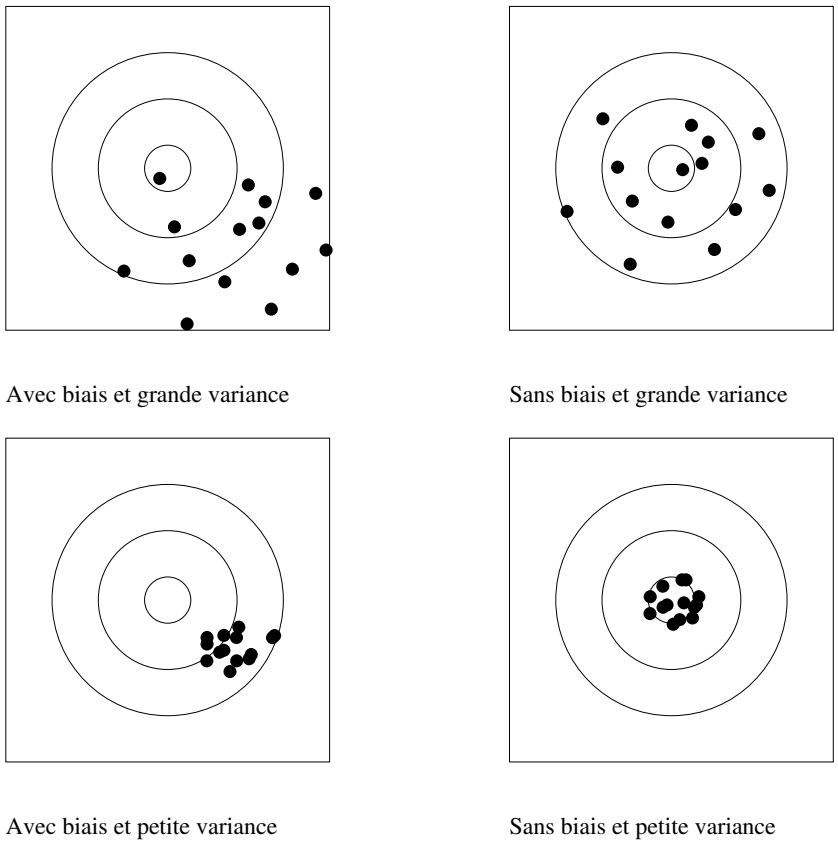


FIG. 1.9 - *Biais et variance d'un estimateur.*

La vraie valeur est au centre de la cible, les points d'impact représentent la distribution de probabilité de l'estimateur (Jessen, 1978).

Pour apprécier la qualité d'un estimateur, on utilise généralement l'écart-quadratique moyen. Si $\phi(X)$ est l'estimateur d'un paramètre θ, il se définit comme $E[(\phi(X)-\theta)^2]$, c'est à dire la dispersion de cet estimateur autour de la véritable valeur du paramètre. On peut développer l'écart-quadratique moyen suivant la formule suivante :

$$E[(\phi(X)-\theta)^2] = (E[\phi(X)]-\theta)^2 + V[\phi(X)].$$

L'écart quadratique moyen est l'indice de qualité d'un estimateur. Plus il est faible, plus l'estimateur est dit précis. Dire qu'un estimateur possède une bonne précision signifie dans la pratique que son écart quadratique est au dessous d'une valeur seuil de référence.

La formule de l'écart quadratique moyen montre que l'erreur qu'on commet se décompose en deux parties : le carré du biais et la variance de l'estimateur. Le biais correspond à l'écart entre la valeur à estimer et l'espérance de

l'estimateur, la variance correspond à la variabilité de l'estimateur. La distinction entre ces 2 composantes d'erreur est importante dans certains cas : par exemple, dans une expérimentation où l'on effectue des mesures répétées avec un même instrument, le biais peut correspondre à une erreur systématique alors que la variance peut peut-être diminuer avec un nombre de répétitions plus important.

A titre d'exemple, considérons la mesure de la température dans un champ à 15 cm du sol. Si le thermomètre utilisé majore systématiquement la température mesurée d'un demi-degré, il engendre un biais systématique. Alors que la variance, liée par exemple à l'hétérogénéité du champ peut être diminuée par des mesures supplémentaires. Si le biais est important par rapport à la variance, il est inutile d'augmenter le nombre de mesures, mieux vaut prendre un instrument plus exact.

1.3 Echantillonnage

1.3.1 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, ...

1.3.2 Procédure d'échantillonnage

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

Si l'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.

1.3.3 Populations finies, infinies, fixes et aléatoires

Deux concepts importants sont à distinguer, le premier est lié à la taille de la population, le second est lié au point de vue selon lequel on examine la population.

Dans la figure FIG. 1.10, le nombre d'unités statistiques de la population est très restreint, on dit être dans le cas de population finie.

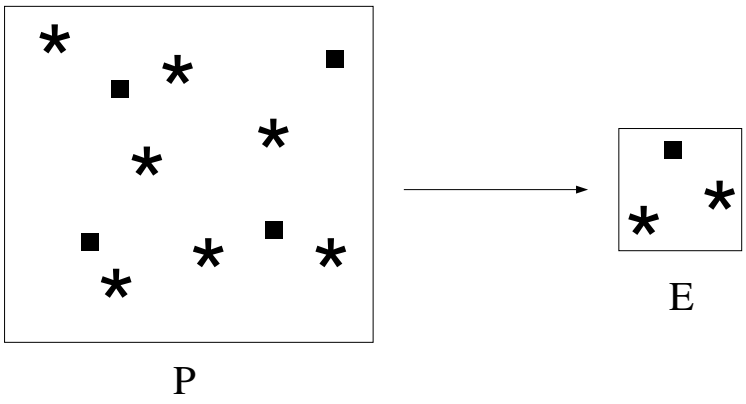


FIG. 1.10 - *Echantillonnage : tirage d'une fraction de la population P, appelée échantillon E.*

Les populations statistiques sont en général finies mais lorsque le nombre d'unités statistiques est très grand (par exemple, dans un champ de plusieurs hectares, on compte des centaines de milliers de plants), on dit être dans le cas de population infinie. Ceci correspond alors à une approche asymptotique et les calculs mathématiques sont simplifiés.

Si l'on ne s'intéresse pas à la population qu'on échantillonne mais à un ensemble de populations dont la population échantillonnée n'est qu'un représentant, on dit alors être dans le cas de super-population. La population étudiée correspond quasiment à un premier niveau d'échantillonnage. On préfèrera à super-population l'expression de population aléatoire qui s'oppose mieux à celui de population fixe. Ce dernier terme traduit le fait qu'on échantillonne la population conditionnellement aux valeurs prises par la variable examinée sur les unités statistiques. Ces valeurs sont donc considérées fixes et inconnues, et non pas des réalisations de variables aléatoires issues d'un modèle probabiliste. C'est dans ce cadre que s'applique la notion de représentativité (voir ci-dessous). Au contraire le cas de population aléatoire suppose une modélisation probabiliste comme par exemple celle d'un processus Poissonnien ou d'une distribution normale.

1.3.4 Espérance et variance d'une population

Enfin, il est commode d'utiliser les termes d'espérance et de variance d'une population de taille N pour désigner les fonctions suivantes du caractère étudiée :

$$E(P) = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \text{et} \qquad V(P) = \frac{1}{N}\sum_{i=1}^{N}(X_i - E(P))^2$$

où X_i est la valeur observée sur l'unité statistique i.

Dans le cas d'une population fixepopulation fixe, $E(P)$ et $V(P)$ sont des paramètres fixes à estimer tandis que dans le cas de population aléatoire, ils sont des paramètres aléatoires à prédire.

1.3.5 Recherche d'une procédure d'échantillonnage

La mise au point d'une procédure d'échantillonnage (FIG. 1.11) :

1. s'apparente dans la formalisation statistique à celle de la construction d'un dispositif expérimental. La différence essentielle est que dans le cas des dispositifs expérimentaux on peut contrôler *a priori* les facteurs influant sur les résultats.

2. n'a guère de sens si elle ne se fait pas dans le cadre de contraintes, en général un coût limité. Sinon, on imagine aisément que la procédure optimale serait toujours l'observation exhaustive de la population.

L'utilisation des fonctions de coût est en général assez délicate car leur détermination réaliste est souvent difficile. Comment chiffrer le déplacement dans une parcelle par rapport au temps d'observation? Comment intégrer le risque de non-respect des consignes qu'introduit une procédure plus précise mais trop sophistiquée? En général on travaille à taille d'échantillon constant, mais c'est supposer implicitement que la fonction de coût ne dépend que du nombre d'observations! Une autre possibilité est de rechercher la taille minimale pour obtenir une précision donnée. Finalement il faut toujours que le biologiste ait le dernier mot dans le choix final car dans sa tête d'expert il peut intégrer des considérations plus fines, impossibles à mettre en équation. Mais pour pouvoir faire la part des choses, il est nécessaire qu'il ait suffisamment assimilé les principes qui guident les choix proposés par le statisticien.

La mise au point d'une procédure d'échantillonnage ne serait pas complète si l'on ne lui associait pas l'analyse statistique des résultats. Dans le cas de populations finies on retrouvera les méthodes classiques type analyse de la variance, les problèmes liés à l'estimation et aux tests d'hypothèses.

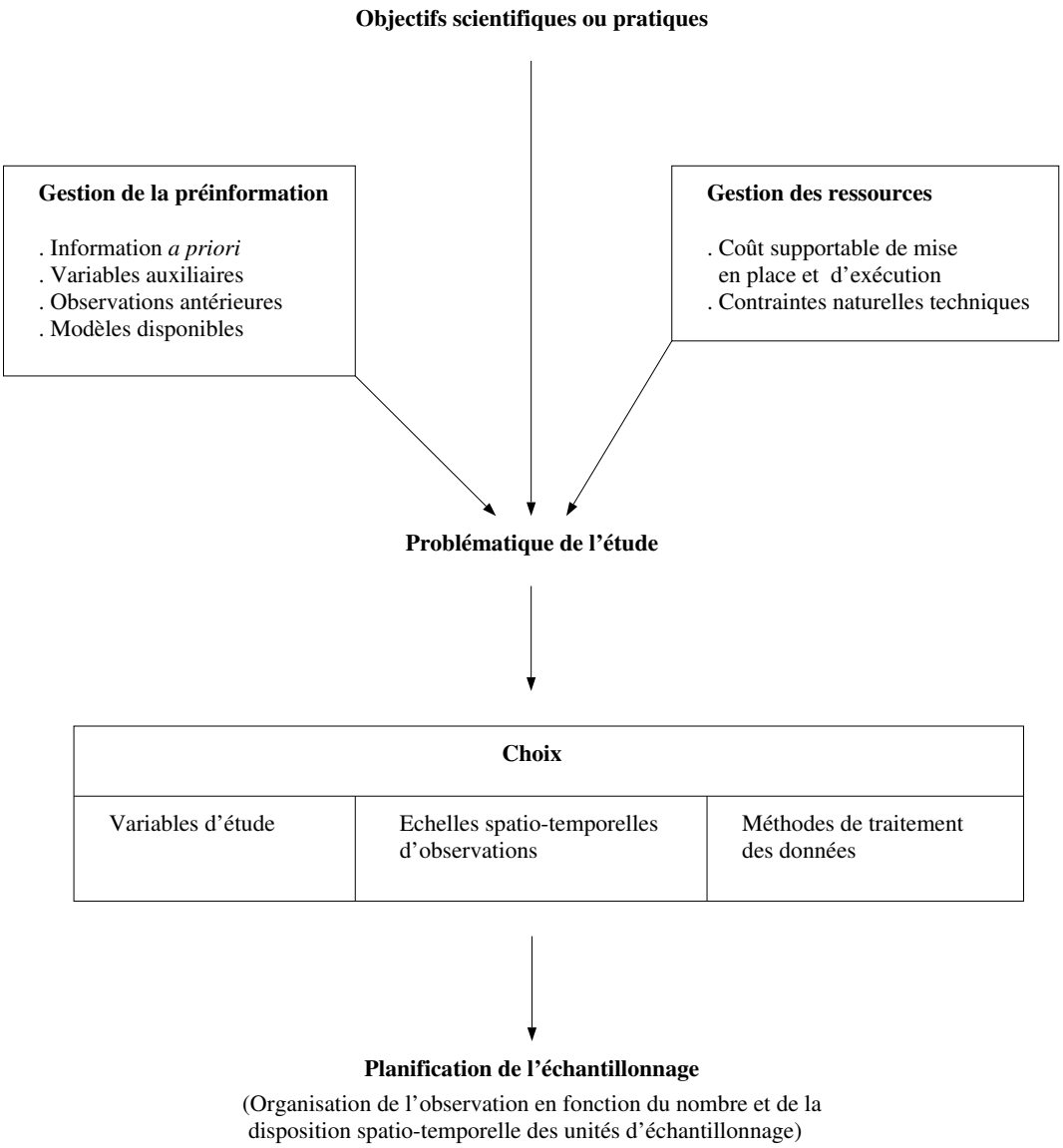


FIG. 1.11 - *Recherche d'une procédure d'échantillonnage, schéma décisionnel.*

1.3.6 Représentativité d'un échantillon

Une notion importante est celle de la représentativité d'un échantillon. Un échantillon est dit représentatif d'une population si tout élément de la population a une probabilité non nulle et connue d'appartenir à l'échantillon. Dans le cas des champs de maïs évoqués ci-dessus, cela signifie qu'on ne pourra déduire des informations sur l'ensemble des champs que si a priori tous les champs ont une chance de participer à l'échantillon. Réciproquement si nous ne disposons que d'information sur un (ou quelques) champ(s) particulier(s), nous ne pourrons pas tirer de conclusions sans faire de suppositions supplémen-

taires (par exemple une modélisation probabiliste). En d'autres termes, le fait d'avoir des éléments de la population ayant une probabilité nulle d'être inclus dans tout échantillon, conduit, dans le cadre population fixe, à des échantillons non représentatifs. Par contre, ce n'est pas forcément le cas dans le cadre superpopulation puisque, on considérera généralement que les individus de probabilité d'inclusion nulle présenterons certaines similarités avec ceux de probabilité d'inclusion non nulle. Notons que la notion d'échantillon représentatif est également employé dans d'autres sens : stratification, méthodes des quotas (Gourieroux, 1981; Grosbras, 1987).

1.3.7 Population cible

Il faut prendre garde à ne pas confondre la population statistique et la population cible (voir les exemples du § 1.3.1). La population statistique est l'ensemble dont on extrait un échantillon représentatif : c'est sur elle que portent les inférences statistiques. Ainsi, l'ensemble des parcelles d'une forêt est la population statistique dans l'exemple du § 1.3.1) alors que, *par définition, la population cible est l'ensemble sur lequel doivent porter les conclusions de l'étude.* Il peut s'agir des arbres de la forêt sur lesquels on étudie la présence d'une maladie; mais, il peut s'agir également de la population des larves d'un certain parasite des arbres. La distinction fondamentale entre population statistique et population cible tient à l'impossibilité technique parfois d'extraire un échantillon représentatif de la population cible. On peut tirer au hasard de manière équiprobable les parcelles de la forêt mais on ne sait pas le faire pour les larves de parasite.

1.4 Bibliographie

Cochran, W. G. (1977) Sampling techniques. Wiley. 3*ième* édition. 428 pages.

Ouvrage de référence. Il présente la démonstration mathématique de nombreux résultats. Toutes les procédures classiques d'échantillonnage sont passées en revue. Leurs avantages et inconvénients sont décortiqués.

Dagnélie, P. (1973) Théorie et méthodes statistiques (2 volumes). Presses agronomiques de Gembloux. 378+463 pages

Ouvrage général de présentation des méthodes statistiques usuelles.

Droesbeke, J-J., Fichet, B., et Tassi, P. (1987) Les sondages. Economica, ASU. 310 pages.

Ouvrage résultant des Journées d'Etude de l'Association des Statisticiens Universitaires, organisées en 1986. Présentation mathématique de

la théorie des sondages avec des exemples d'applications de nature socio-économique.

GOURIEROUX, C. (1981) Théorie des sondages. Economica, ESA. 272 pages.

Livre discutant les problèmes de mise en oeuvre des enquètes par sondages en socio-économie mais présentant surtout les liens mathématiques entre théorie des sondages et mathématique statistique.

GROSBRAS, J. (1987) Méthodes statistiques des sondages. Economica, ESA. 331 pages.

Présentation d'une vaste panoplie d'outils permettant de guider la réflexion dans les problèmes de collecte d'information.

HASTINGS, N. A. J. AND PEACOCK, J.B. (1974) Statistical Distributions. Wiley. 130 pages.

Manuel présentant les lois de probabilité usuelles et leurs principales caractéristiques.

JESSEN, R. J. (1978) Statistical survey techniques. Wiley. 520 pages.

Ouvrage de référence avec beaucoup d'exemples numériques.

JOHNSON, N. L. ET KOTZ, S. (1969) Discrete distributions. Houghton Mifflin. 328 pages.

Ouvrage de référence concernant les lois de probabilité discrètes les plus utilisées. Etudes approfondies de leur propriétés distributionnelles.

JOHNSON, N. L. ET KOTZ, S. (1970) Continuous univariate distribution-1. Houghton Mifflin. 300 pages.

Même chose que pour le livre précédent mais pour les lois continues unidimensionnelles.

PIELOU, E. C. (1977) Mathematical Ecology. Wiley. 384 pages.

Ouvrage présentant les outils théoriques de modélisation en dynamique des populations et en mesure de diversité écologique, ainsi que les modèles d'interaction et de ségrégation inter-spécifiques.

SAPORTA, G. (1978) Théories et Méthodes de la statistique. Institut Français du Pétrole. 386 pages.

Ouvrage exposant, de façon relativement élémentaire, les fondements mathématiques de la théorie statistique et les méthodes statistiques classiques.

SOUTHWOOD, T. R. E. (1978) Ecological methods with particular reference to the study of insect populations. Chapman and Hall, London. 524 pages.

Présentation d'une panoplie de méthodes statistiques et pratiques pertinentes dans l'étude de populations écologiques, avec une attention particulière pour les insectes.

TAYLOR, L. R. (1984) Assessing and interpreting the spatial distributions of insect populations. Ann. Rev. Entomol., 29, 321-357.

Présentation des divers outils statistiques et probabilistes d'investigation des répartitions spatiales de populations d'insectes à partir de données de dénombrement.

Chapitre 2

Plans d'échantillonnage classiques

Plan :

2.1 Introduction

Il est souvent nécessaire de prélever un échantillon de la population qu'on 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 avec une probabilité non nulle connue.

Il est clair que s'il s'agit d'étudier une population dont certains éléments ne

sont pas observables avec la technique de prélèvement utilisée, l'échantillon tiré ne sera pas représentatif. Il faudra, ou bien adopter une définition plus réaliste de la population étudiée, ou bien améliorer sa technique de prélèvement, ou faire des hypothèses de similarité (discussion du § 1.3.6).

Un corps de théorie, la théorie de l'échantillonnage, encore appelée théorie des sondages, est né de ce besoin de conclure sur le tout à partir de la partie. De cette théorie découle naturellement certaines techniques ou plans d'échantillonnage, dont le but est d'estimer certains paramètres d'une population statistique avec le maximum de précision et le minimum d'effort (voir, par exemple, Chaudhuri et Vos, 1988).

Si ces techniques ont connu un grand succès en sciences humaines, en particulier dans les enquêtes économiques et sociologiques, elles sont demeurées mal exploitées ailleurs. Pourtant leur connaissance par le biologiste peut lui permettre de minimiser le coût de collecte de ses données ou d'en tirer un meilleur parti en optimisant la précision de ses estimations. Un autre avantage, indirect celui-là, est d'amener le biologiste à accroitre la cohérence interne et la rigueur de sa démarche en lui donnant des fondements plus solides tant dans la phase de collecte que dans celle de traitement des données (FIG. 1.11).

Dans ce chapitre nous allons examiner un certain nombre de plans d'échantillonnage parmi les plus connus. Afin de simplifier cette présentation d'ensemble, nous nous placerons dans le cas d'une population finie et fixe (voir ces notions au § 1.3.1).

2.2 Echantillonnage non-aléatoire

Exemple 1 :

Soit à étudier le personnel de l'entreprise où nous travaillons. Si nous connaissons très bien cette entreprise, nous pouvons être tenté de choisir certains individus parce que nous pensons qu'ils sont représentatifs de leur service. Cette procédure est appelée *échantillonnage raisonné.*

Il présente de sérieux dangers. Il est clair que nous risquons d'introduire un fort biais personnel dans notre choix. Ce choix va refléter notre connaissance du personnel (sommes-nous bien sûr de connaitre tout le monde ?) et notre jugement, ici très subjectif. Nous nous exposons donc à introduire des biais (simples ou subtils) en procédant de la sorte.

Exemple 2 :

Soit à étudier les étudiants présents à un enseignement. On peut constituer un échantillon constitué des élèves du premier rang.

Là encore danger, les étudiants du premier rang peuvent être un peu dur

d'oreille, ou myope, ou mal comprendre le français, ou plus intéressé par le cours que ceux du fond de l'amphi, *etc.* Toutes ces caractéristiques peuvent être plus ou moins corrélées aux variables à étudier (ou qui le seront, bien que non prévues au début de l'étude).
La même critique vaudrait si nous nous laissions guider uniquement par des considérations de facilité d'exécution, de rapidité, d'économie, ou simplement parce que nous ne voyons pas de critiques à formuler à l'encontre de la procédure suivie.

Remarque : Il ne s'agit pas de condamner sans nuance le procédé de l'échantillonnage non-aléatoire (l'expression échantillonnage raisonné est également utilisé de façon équivalente). Il revient à introduire une information préalable sur la population étudiée. Mais, il s'agit de faire en sorte que cette information initiale soit utilisée pour orienter correctement l'étude statistique, sans introduire de biais et d'artefacts. Ce paradoxe se résoud aisément si on envisage une hiérarchie de niveaux : certains niveaux supérieurs faisant l'objet de sélections raisonnées et les niveaux inférieurs de tirages aléatoires.

Un autre type d'échantillonnage non-aléatoire est l'*échantillonnage exhaustif* ou *recensement*. Il consiste tout simplement à observer toutes les unités de la population statistique. Bien-sûr, ceci nécessite un effort d'échantillonnage qui n'est que rarement envisageable.

2.3 Plans d'échantillonnage aléatoire à un niveau

2.3.1 Echantillonnage aléatoire simple

2.3.1.1 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. Les tirages peuvent être avec ou sans remise. Dans la pratique, ils sont pour la plupart sans remise : on prélève au hasard n éléments (ou unités d'échantillonnage) dans une population qui en comporte N. Dans ce cas, la probabilité pour un élément quelconque d'être inclus dans l'échantillon 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). Le nombre d'échantillons différents pouvant être tirés de la population est égal au nombre de combinaisons de n éléments tirés parmi N soit $N!/n!(N-n)!$.

NB. Nous employerons en synonymie les termes *individu*, *élément* et *unité statistique*.

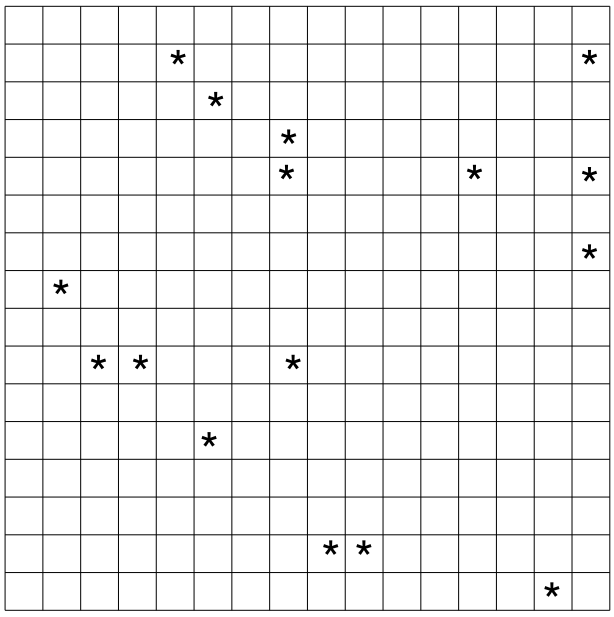


FIG. 2.1 - *Echantillonnage aléatoire simple de 16 unités de surface sur 256.*

2.3.1.2 Application pratique

Cette application n'est pas nécessairement immédiate. En effet si l'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 l'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 informatique *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 tire alors n nids 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.

2.3.1.3 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. Il se traduit en effet parfois par une moyenne d'échantillon pour laquelle l'écart quadratique moyen est en général plus élevé que celui d'autres plans. On dit que sa préci-

sion est faible. C'est le prix qu'il faut payer pour ne pas utiliser d'informations complémentaires sur la population, informations obtenues par des expériences antérieures ou concommitantes.

Ce plan présente un certain nombre d'avantages mathématiques. Ses estimateurs sont en général sans biais (FIG. 1.9) et facilement calculables. A quelques exceptions près, tous les tests d'hypothèse (paramétriques et non-paramétriques) sont directement applicables, ainsi que les méthodes d'analyse multidimensionnelles.

Pour appliquer ce plan, il faut connaître 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 exemple, 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. Ainsi, en saisissant à la main et au hasard, des blattes dans leur conteneur, on croira pratiquer un échantillonnage aléatoire simple alors que les individus les plus faibles ont une probabilité plus grande d'être échantillonné.

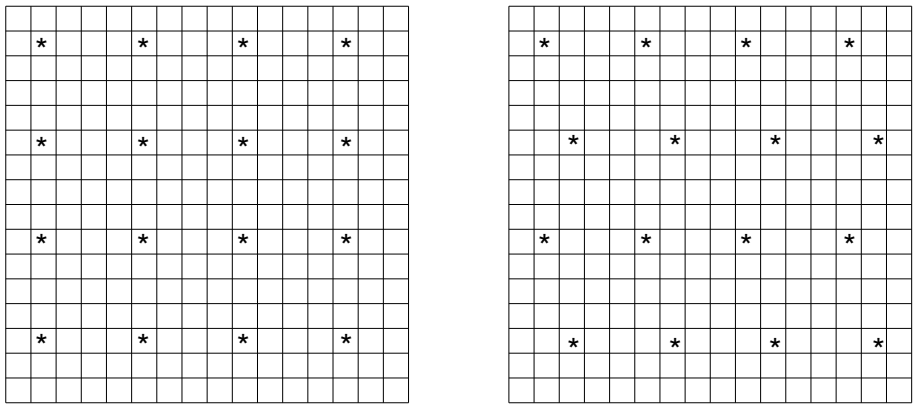
2.3.2 Echantillonnage systématique

2.3.2.1 Définition

L'échantillonnage aléatoire simple n'est pas toujours facile à réaliser car il oblige à trouver chacun des éléments sélectionnés, où qu'il puisse se trouver. On préfère donc, dans certains cas, se simplifier la tâche en partant d'un élément tiré au hasard et en prélevant ensuite des éléments régulièrement espacés suivant un *pas* choisi généralement en fonction du coût global d'échantillonnage.

Dans ce plan, seul le premier 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 k. On a donc $N = nk$ (quand N n'est pas un multiple de n, l'approche est différente (Iachan, 1982). Ensuite, on tire au hasard un nombre parmi $1, \ldots, k$. Soit i le nombre tiré, l'échantillon est alors constitué du *ième* élément de chacun des sous-ensembles.

Le nombre d'échantillons différents qui peuvent être tirés de la population est beaucoup plus petit avec ce plan qu'avec le plan aléatoire simple : il est de $k = N/n$ seulement.



(a) (b)

FIG. 2.2 - *Echantillonnage aléatoire systématique d'unités de surface (16 unités tirées sur 256 : (a) de façon ordinaire, (b) en quinconce).*

Le plan systématique maximise la distance entre unités observées : il est donc meilleur si l'autocorrélation entre unités est positive et décroit avec la distance (§ 3.5.2.2).

2.3.2.2 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; k 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 k. Il s'applique également lorsque la population est à 2 dimensions, comme un champ, ou à 3 dimensions, comme un lac. Par exemple, les n sous-ensembles ont alors une surface de k mètres carrés ou un volume de k mètres cubes. Les unités échantillonnées sont alors disposés de manière régulière, à raison d'une par sous-ensemble.

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 sous-ensemble.

2.3.2.3 Avantages et inconvénients

Ce qui distingue principalement le plan systématique des autres plans classiques, c'est la régularité des points échantillonnés. Quelles conséquences cela

a t-il? Par rapport à l'échantillon que fournit un sondage aléatoire simple, celui obtenu par un sondage systématique est :

1. plus représentatif si les éléments de la population sont autocorrélés positivement ou bien présentent une tendance linéaire. Rappelons qu'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. En effet ce dernier comporte, par le fait du hasard, des relevés proches et d'autres éloignés, ce qui entraine des redondances et des défauts d'information dans l'échantillon. Un exemple simple d'autocorrélation spatiale et temporelle est fournie par les conditions météorologiques car les relevés effectués en 2 stations rapprochées sont très corrélées. De nombreux phénomènes écologiques présentent également une autocorrélation positive. Le plan systématique est supérieur dans ce cas parce qu'il impose une distance minimum entre les éléments et évite, par conséquent, la collecte de données redondantes.

2. plus précis si le caractère étudié varie linéairement avec le numéro d'ordre des éléments dans la série (§ 3.5.2.1). Ce phénomène peut apparaitre en écologie dans un gradient d'altitude, de distance à la côte, de hauteur de végétation, dans l'évolution temporelle d'un phénomène.

3. équivalent si la caractéristique étudiée se répartit purement au hasard. L'avantage du plan systématique résulte alors de la commodité de sa préparation et de son exécution sur le terrain.

4. 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. Une importante erreur systématique peut apparaitre. Or les phénomènes cycliques sont fréquents dans la nature. Il est facile de tenir compte des rythmes circadiens, lunaires, saisonniers ou annuels. Par contre d'autres régularités naturelles peuvent passer inaperçues : la perte de précision dans le cas où il existe des variations périodiques insoupçonnées est l'inconvénient majeur de ce plan.

2.3.3 Echantillonnage avec probabilités inégales

Tout d'abord, il convient de souligner la différence fondamentale existant dans la littérature entre *probabilité de tirage (ou de sélection)* d'une unité et *probabilité d'inclusion dans l'échantillon* d'une unité. La distinction vient du fait

suivant : les n unités d'un échantillon ne sont, en général, pas choisies simultanément mais par n tirages successifs d'une unité et une seule. Ainsi, la probabilité de tirage d'une unité se réfère à la chance que cette unité a d'être retenue pour un tirage donné tandis que la probabilité d'inclusion concerne la chance qu'a une unité, une fois les n tirages effectués, d'être dans l'échantillon. Voyons concrètement ce qu'il en est sur les deux exemples suivants :

Echantillonnage aléatoire simple avec remise de taille n dans une population de N unités :

la probabilité de sélection d'une unité lors du premier, du deuxième, ..., du *nième* tirage est $1/N$. Par contre, la probabilité d'inclusion dans l'échantillon est

$$1 - (1 - \frac{1}{N})^n.$$

Echantillonnage aléatoire simple sans remise de taille n dans une population de N unités :

la probabilité de sélection d'une unité est $1/N$ lors du premier tirage, $1/(N-1)$ (ou 0 si elle a été tirée précédemment) lors du deuxième tirage, ..., $1/(N-n+1)$ (ou 0 si elle a été tirée précédemment) lors du *nième* tirage. Mais la probabilité d'inclusion dans l'échantillon est simplement n/N.

2.3.3.1 Définition

L'échantillonnage avec probabilités inégales de sélection des unités, consiste à prélever de façon aléatoire un échantillon de taille n parmi N unités caractérisées par des probabilités de tirage inégales, non-nulles et toutes connues. Un cas particulier important est celui du tirage d'une unité avec probabilité proportionnelle à son importance. Dans ce cas, la probabilité p_i de sélection de la *ième* unité, de taille t_i, est

$$p_i = t_i/T$$

où T est la somme des tailles des unités.

Si l'échantillonnage est sans remise, cette expression de p_i n'est valable que pour le premier tirage. Ensuite, elle change en fonction des unités sélectionnées aux tirages précédents. Par contre, si l'échantillonnage s'effectue avec remise, l'expression de p_i reste inchangée au cours des tirages et la probabilité d'inclusion de l'unité i dans l'échantillon est

$$1 - (1 - p_i)^n.$$

2.3.3.2 Application pratique

Exemple : Soit à estimer le nombre de cheptels bovins atteints par l'hypoder-

mose bovine dans une région, à partir d'enquêtes par comptages de varrons. On peut décider de tirer un cheptel pour examen en fonction du nombre de bovins qu'il comprend. Plus un cheptel sera important, plus il aura des chances d'être inclus dans l'enquête. On aura alors un échantillonnage à probabilités inégales, l'inégalité des probabilités provenant de l'inégalité des tailles de cheptel.

Pour appliquer le plan à probabilités proportionnelles à la taille, il faut connaître le nombre d'unités N, leur taille cumulée T et la taille t_i des unités sélectionnées aléatoirement. Ceci ne pose pas de difficultés particulières si les éléments de la population sont des intervalles de temps, de distance, de surface, de volume. Dans les autres cas, cela peut être contraignant.

Si les tailles t_i ne sont pas connues, on peut les estimer sur un échantillon aléatoire simple de grande taille. On pratique alors ce qu'on appelle le double échantillonnage. Le second échantillon, avec remise et probabilités proportionnelles, est un sous-échantillon du premier.

2.3.3.3 Avantages et inconvénients

L'échantillonnage avec probabilités proportionnelles à la taille est fortement recommandé lorsque :

1. la taille ou l'importance t des éléments de la population varie considérablement
2. la variable étudiée Y est fortement corrélée à t
3. le coût unitaire de la mesure de Y est indépendant de t.

Dans ces conditions, la précision d'estimation de la moyenne de la caractéristique étudiée est souvent très nettement supérieure à celle obtenue par l'aléatoire simple. D'une façon plus générale, c'est le cas dès lors que les probabilités de tirage sont proportionnelles à une variable auxiliaire X corrélée positivement à Y.

En revanche, il faut savoir que la majorité des programmes informatiques de traitements statistiques ne sont pas conçus pour ce plan.

2.4 Plans d'échantillonnage à plusieurs niveaux

Les méthodes d'échantillonnage simple et systématique correspondent à une situation bien particulière : celle où on ne tient compte d'aucune information *structurale* sur les éléments de la population. C'est oublier l'intérêt intrinsèque que présente une telle information. En effet, les éléments de la population peuvent être, dans bien des cas, classés et regroupés de façon hiérarchique.

Examinons les principales méthodes d'échantillonnage utilisant ces informations.

2.4.1 Echantillonnage stratifié

2.4.1.1 Définition

L'échantillonnage stratifié consiste tout d'abord à subdiviser une population hétérogène en H 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 :

1. au premier niveau on effectue un recensement (celui des strates)

2. au second niveau on effectue un échantillonnage simple (celui des individus dans une strate)

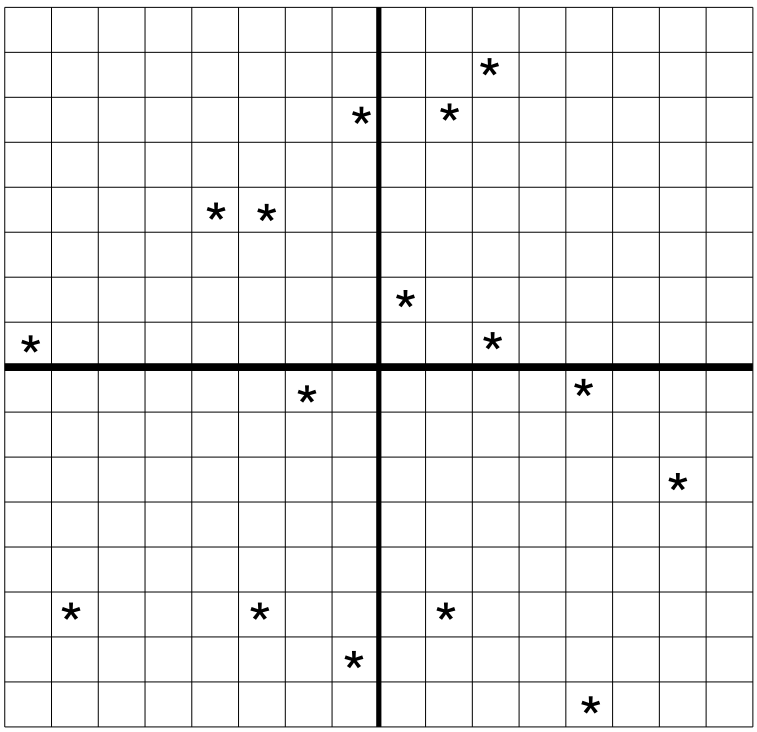


FIG. 2.3 - *Echantillonnage stratifié d'unités de surface (4 strates de 64 unités, 4 observations par strate).*

NB. Pour certains auteurs, l'échantillonnage utilise, par définition, un échantillonnage aléatoire simple au second niveau. D'autres étendent la définition à l'utilisation de n'importe quel plan à ce second niveau.

2.4.1.2 Application pratique

a. *Construction des strates.*

Il faut, pour commencer, choisir un critère de stratification. Le meilleur critère serait, bien sûr, la variable étudiée Y. On découperait sa distribution de fréquences en classes (les strates) ce qui aurait évidemment pour effet d'accroitre l'homogénéité au sein des strates (la variance intra-strate serait réduite). Mais comme on ne connait pas au départ les valeurs prises par la variable Y sur la population statistique (on s'est justement donné pour tâche de les mesurer sur certaines unités), on utilise, en pratique, une autre variable appelée *stratificateur*. Cette variable doit être en corrélation aussi étroite que possible avec la variable étudiée. Le stratificateur peut être une variable quantitative ou qualitative. En outre on peut utiliser simultanément plusieurs critères de classification (double, triple stratification).

On doit ensuite fixer le nombre de strates. La précision s'accroit avec le nombre de strates. Cependant au delà d'un certain nombre de strates le rendement (rapport précision-coût) de l'échantillonnage décroît très vite.

La fixation des limites de strates n'est pas immédiate dans le cas d'un stratificateur quantitatif. Il s'agit de minimiser la variabilité des estimateurs. On doit enfin déterminer le nombre d'éléments N_h de la strate h pour $h = 1, 2, \ldots, H$.

b. *Détermination de l'effectif des échantillons.*

L'effort d'échantillonnage pouvant varier d'une strate à l'autre, plusieurs stratégies sont utilisables :

La première idée venant à l'esprit est de sélectionner dans chaque strate un nombre n_h d'éléments proportionnel à l'importance N_h de cette strate dans la population. C'est *l'allocation proportionnelle*. Ce faisant, on ne tient pas compte du fait que les variabilités existant au sein des différentes strates peuvent grandement différer d'une strate à l'autre.

L'allocation dite *optimale* tient compte de ce phénomène. Elle consiste à prendre un échantillon de taille d'autant plus élevée que la variabilité à l'intérieur de la strate est grande et que son effectif N_h est élevé. Si l'on veut tenir également compte du coût unitaire de l'échantillonnage, on privilégie les strates à coût faible.

Enfin, il existe une solution de compromis qui consiste à fixer un nombre minimum d'éléments dans chaque strate puis à compléter en fonction des règles de l'allocation optimale. C'est uniquement dans le cas où l'on ignore les valeurs

des variances intra-strates qu'on opte pour l'allocation proportionnelle.

c. *Choix du plan d'échantillonnage à l'intérieur des strates.*

Dans le cas de la définition large de ce plan, on peut, en principe, choisir n'importe quel plan ou combinaison de plans puisque l'échantillonnage d'une strate est indépendante de celui d'une autre. Cependant, dans la majorité des cas, pour ne pas compliquer les choses, on a intérêt à pratiquer un échantillonnage aléatoire simple. On est ainsi ramené à la définition restreinte.

2.4.1.3 Avantages et inconvénients

Un échantillonnage avec stratification correctement faite est toujours plus efficace qu'un sondage aléatoire simple (§ 2.4.2.2). Même rudimentaire, la stratification peut entrainer des gains de précision appréciables. L'allocation optimale fournit des résultats d'une précision supérieure (au pire égale) à l'allocation proportionnelle, qui fournit elle-même des résultats plus précis que l'échantillonnage aléatoire simple. On aura donc intérêt à stratifier une population chaque fois qu'on pourra.

Encore faut-il connaître avec précision la répartition de la population entre les différentes strates. Une erreur d'évaluation du poids relatif N_h/N des strates entraîne en effet un biais qui reste constant quel que soit l'effectif n_h des échantillons.

Dans le cas de l'allocation proportionnelle, la majorité des tests d'hypothèses et les méthodes multidimensionnelles peuvent être directement appliquées sur l'échantillon agrégé d'effectif n. Des restrictions apparaissent dans le cas de l'allocation optimale.

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 ...), lorsque certaines catégories de la population doivent être sur-représentées (éléments rares, au rôle important, *etc.*) ou lorsqu'on désire maximiser la plage de variation d'un facteur pour étudier ses effets. Prenons le premier cas : soit à réaliser 12 prélèvements par jour. Pour des raisons de coût on ne peut en faire que 2 par nuits. On est ainsi amener à considérer une strate *journée* et une strate *nuit*.

2.4.2 Echantillonnage par grappes

2.4.2.1 Définition

L'échantillonnage par grappes a une ressemblance superficielle avec l'échantillonnage stratifié en ce sens qu'ici encore la population est divisée en sous-populations. Mais, première différence avec le plan précédent, quelques-unes

seulement des sous-populations sont sélectionnées. En outre, seconde différence, les sous-populations sélectionnées sont intégralement étudiées alors qu'on effectue un échantillonnage non exhaustif à l'intérieur de chaque strate. On procède donc, à l'inverse du sondage stratifié, de la façon suivante :

1. premier niveau : échantillonnage simple

2. second niveau : recensement

L'information préliminaire qu'on utilise dans ce plan d'échantillonnage concerne la structure de la population. Les N unités d'échantillonnage de la population, appelées grains, sont regroupées en unités plus grandes, appelées grappes. L'échantillonnage en (ou par) grappes se réalise en trois étapes. On commence par définir les grappes (elles correspondent parfois à une entité naturelle). Puis on tire certaines grappes selon la méthode du sondage aléatoire simple (ou systématique pour certains auteurs). Enfin on examine (mesure) tous les grains des grappes tirées.

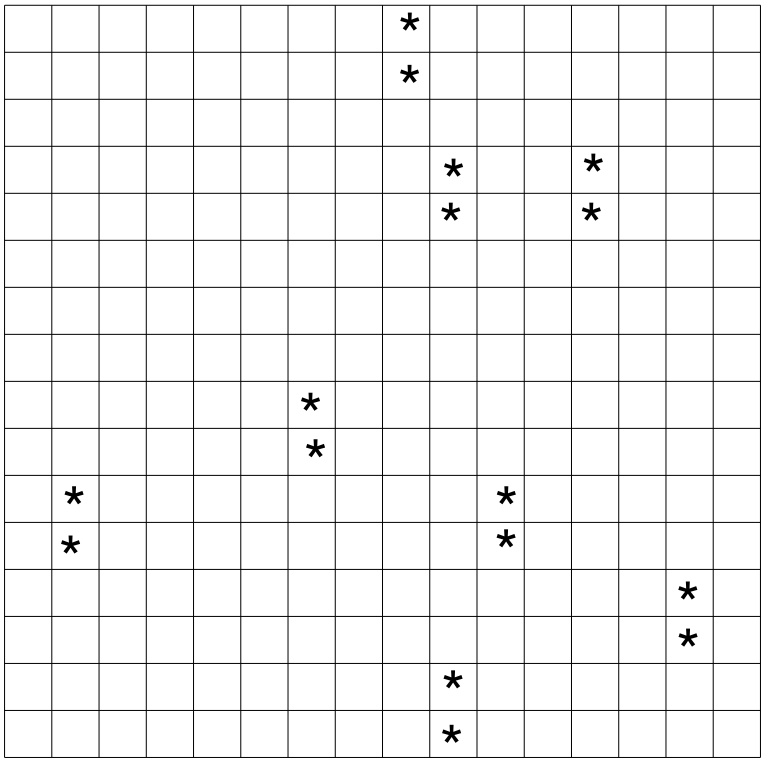


FIG. 2.4 - *Echantillonnage en grappe d'unités de surface (8 grappes tirées parmi 128, 2 grains recensés par grappe).*

2.4.2.2 Comparaison avec les autres méthodes

a. *Simple*

Si l'on agrège les observations effectuées sur les grains d'une même grappe à une valeur unique, alors le sondage par grappe se ramène au sondage simple puisqu'on ne tient plus compte de l'information apportée individuellement par chaque grain.

b. *Systématique*

L'échantillonnage systématique de taille $n = N/k$ peut être considérer comme l'échantillonnage d'une grappe unique choisie parmi k grappes de même taille n (§ 3.3).

c. *Stratifié*

Lors de la subdivision d'une population en sous-population (strate ou grappe, que nous désignerons ici sous le terme générique de classe) la variabilité totale de la population (traduisant les écarts entre la moyenne de la population et chaque élément de cette population) peut être décomposée en plusieurs termes : la variabilité inter-classe (traduisant les écarts entre la moyenne de chaque classe et la moyenne de la population) et les variabilités des classes (traduisant les écarts entre la moyenne d'une classe et chaque élément de cette classe). On montre que la variabilité totale notée SC est simplement la somme de la variabilité inter-classe SC_{inter} et d'une variabilité, dite intra-classe SC_{intra}, associée à la moyenne (pondérée par les effectifs de chaque classe) des variances des classes.

$$SC = SC_{inter} + SC_{intra}$$

Supposons que le découpage de la population soit tel que la variabilité intra-classe soit nulle. Cela signifie que toutes les variances de classe sont nulles et, par conséquence de leur définition, que les écarts entre les valeurs des éléments et la moyenne de chaque classe sont nuls. Autrement dit si la variabilité intra-classe est nulle, les éléments de la même classe présentent tous la même valeur. Il suffit alors pour connaître la moyenne de la population de tirer un élément dans chaque classe. Dans ce cas l'échantillonnage stratifié avec tirage d'un seul élément par strate (ou classe) est suffisant pour parfaitement estimer le paramètre étudié de la population (par exemple, son espérance).

Si, en revanche, le découpage est tel que la variabilité inter-classe est nulle, il s'ensuit immédiatement que chaque classe présente la même valeur moyenne. Il suffit donc pour connaître la moyenne de la population de tirer une classe et une seule et de l'étudier intégralement. Dans ce cas, l'échantillonnage par grappe avec tirage d'une seule grappe est suffisant.

Bien entendu, ce sont *2 cas limites* mais qui montrent bien le caractère complémentaire de ces 2 méthodes. On peut en retenir que si on peut découper la population en strates à grande homogénéité interne mais très différentes entre

elles on se trouve dans les conditions idéales d'application de l'échantillonnage stratifié. Au contraire si on peut la découper en grappes à grande hétérogénéité interne mais très semblables entre elles, on se place dans les meilleures conditions d'utilisation de l'échantillonnage en grappes (TAB. 2.1).

Variance intra-classe	**Variance inter-classe**	**Plan d'échantillonnage à utiliser**
Faible	Forte	Stratifié
Forte	Faible	En grappes

TAB. 2.1 - *Choix entre plans stratifié et en grappes selon le type de variabilité.*

2.4.2.3 Application pratique

Exemple 1 : Biologie de la reproduction de la chouette Effraie. L'auteur a visité systématiquement les clochers de la Côte d'Or pour y recenser les nichées. Une fraction des clochers abritant une nichée a été tirée au hasard. Tous les jeunes (grains) des nichées sélectionnées (grappes) ont été pesés.

Exemple 2 : Etude de la population de bovins de moins d'un an dans une région. Pour étudier une caractéristique quelconque de cette population (ex: attaque par un parasite), on choisit au hasard des cheptels et on détermine la valeur de la caractéristique étudiée sur toutes les bovins de moins d'un an des cheptels sélectionnés.

a. *Définition des grappes.* Les grappes peuvent être des regroupements naturels d'éléments naturels (nids de jeunes) ou artificiels (bassin versant divisé en parcelles-échantillons), ou bien des regroupements artificiels d'éléments naturels (piéges renfermant des insectes) ou artificiels (parcelles-échantillons composées de placettes).

b. *Choix de la taille des grappes.* Dans un certain nombre de cas, où l'on peut faire varier la taille des grappes (surface de prélèvement, période d'observation, regroupement artificiel d'unités voisines), des méthodes empiriques normalisées ont été mises au point (Cochran, 1977).

Les grappes doivent, dans la mesure du possible, être constituées du même nombre de grains et former des unités fonctionnelles facilitant le travail sur le terrain. Pour un nombre fixé de grains à échantillonner, on préfère prendre

une taille de grappe d'autant plus grande que le déplacement entre grappes sera coûteux. Un choix rationnel entre deux tailles de grappe peut être effectué en appliquant le principe du moindre coût à précision fixée ou de la moindre variabilité à coût fixé. Cependant, l'information sur la variabilité intra-grappe n'est pas toujours disponible.

Notons que dans le cas de grappes naturelles, les tailles de grappe ne sont généralement pas contrôlées et que, par conséquent, le nombre de grains dans l'échantillon n'est donc pas connu d'avance (voir discussion au § 5.2.1)

c. *Tirage des grappes.* Pour certains auteurs, les grappes doivent être choisies par tirage aléatoire simple (probabilités égales). Pour d'autres, la définition s'étend à d'autres types de tirage, en particulier systématique et à probabilité proportionnelle à la taille (ou à une estimation de la taille).

2.4.2.4 Avantages et inconvénients

a. *Le problème de la liste exhaustive des éléments*

Soit à faire une étude d'exploitations agricoles. Le tirage en grappe consiste, par exemple, à tirer au sort, non parmi les exploitations agricoles, mais parmi les communes, puis à faire l'étude complète des exploitations des communes échantillonnées. On voit immédiatement l'avantage de ce sondage : il n'est pas nécessaire de posséder la liste complète des exploitations pour l'ensemble du territoire mais seulement pour les communes échantillonnées. C'est un avantage qui est souvent décisif et qui fait qu'on préfère ce plan aux autres dans la mesure où ils exigent une liste complète des éléments de la population.

L'échantillonnage en grappe s'impose donc lorsqu'il est impossible d'inventorier les éléments (grains) de la population mais seulement des regroupements de grains (grappes). C'est souvent le cas en écologie car il n'est pas toujours possible d'énumérer tous les arbres d'une forêt ou toutes les plantes d'un champ pour procéder à un tirage aléatoire simple ou systématique de certains d'entre eux, mais l'inventaire d'arbres ou de plantes d'une petite parcelle peut être envisagé.

b. *Le problème de la dispersion géographique*

Les plans examinés précédemment (non en grappe) ne peuvent éviter une dispersion géographique des individus composant l'échantillon. Ils peuvent donc, dans certains cas, coûter cher en frais de transport. On préfère alors prélever non pas des éléments pris un à un mais des *grappes* d'individus géographiquement voisins.

Cet avantage en coût se paye parfois en perte en précision parce que des éléments proches auront tendance dans de nombreux cas à se ressembler davantage que des éléments éloignés. On s'éloigne ainsi de l'optimum qui est une hétérogénéité des grappes aussi grande que possible. Dans ce cas d'autocorrélation

positive à l'échelle de la grappe, le plan en grappes entraîne une perte substantielle de précision par rapport à l'échantillonnage stratifié. Il est introduit pour des raisons essentiellement pratiques de coût de mise en oeuvre : ce qui est optimisé, c'est la précision par unité de coût (rapport précision/coût).

Par rapport aux échantillonnages aléatoires simple et systématique, l'échantillonnage par grappe est :

- moins efficace lorsque les grains d'une grappe se ressemblent beaucoup, à cause de la redondance des informations recueillies dans cette grappe.

- équivalent si le regroupement en grappe n'a aucun rapport avec la caractéristique étudiée.

- plus efficace si l'on trouve dans chaque grappe des individus très dissemblables : on évite alors les redondances dans l'information.

c. *Analyse des résultats*

Certains tests d'hypothèses, comme l'analyse de la variance à un critère de classification, ne sont pas directement applicables. L'interprétation des analyses multidimensionnelles nécessite certaines précautions à cause, notamment, de probables autocorrélations intra-grappes.

2.4.3 Echantillonnage par degré

Il ne peut s'appliquer que lorsque la population est constituée en système hiérarchisé d'unités d'échantillonnage.

L'échantillonnage par degré regroupe un ensemble de plans :

- L'échantillonnage à 2 degrés (ou du deuxième degré) consiste à ne pas examiner tous les grains des grappes tirées, mais seulement certains d'entre eux. Là encore ces grains sont tirés de manière aléatoire. Ainsi, dans l'exemple des exploitations agricoles, on n'étudie pas toutes les exploitations des communes sélectionnées. On en fait la liste exhaustive et on en tire au sort un certain nombre pour étude complète.

On appelle les grappes *unités primaires* et les grains *unités secondaires.*

- On peut bien entendu continuer en définissant un échantillonnage à 3 degrés et ainsi de suite. Remarquons, pour généraliser, que les unités primaires, secondaires *etc.* ne sont pas obligatoirement de même taille. Quand c'est le cas, il vaut mieux parfois utiliser l'échantillonnage à probabilité de sélection proportionnelle à la taille.

Exemple : Etude des pêches. unités primaires : les ports de pêche. unités secondaires : barques du port i. unités tertiaires : poissons de la barque j du port i.

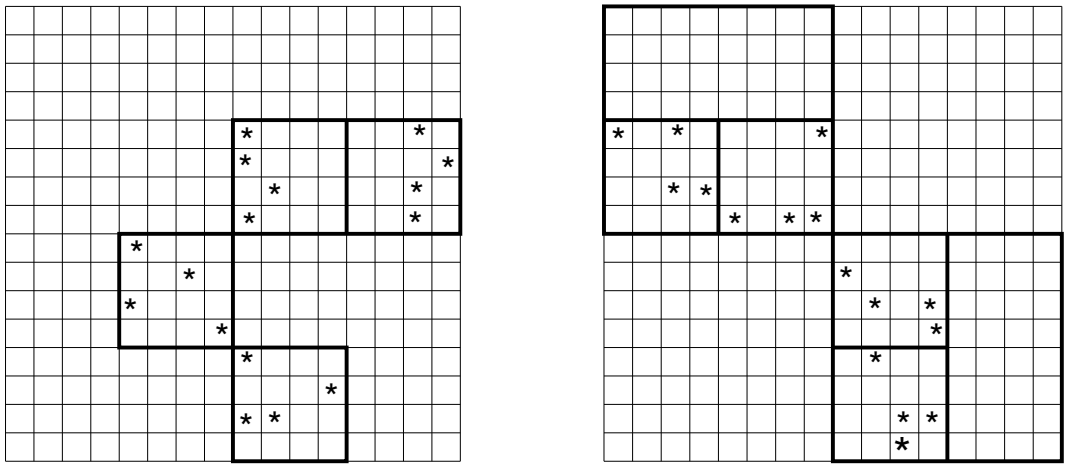


FIG. 2.5 - *Echantillonnage par degré d'unités de surface.*

- 2*ième* degré (à gauche) : 4 unités primaires tirées sur un total de 16. Au sein de chaque unité primaire (de taille 16), 4 unités secondaires sont tirées.

- 3*ième* degré (à droite) : 2 unités primaires tirées sur un total de 4. Au sein de chaque unité primaire (de taille 64), 2 unités secondaires de taille 16 sont tirées et, au sein de chacune d'elles, 4 unités tertiaires sont tirées.

Ce plan est particulièrement adapté lorsqu'on s'intéresse spécifiquement aux différents niveaux hiérarchiques.

Ce plan s'impose également pour les études faunistiques qui portent sur un ensemble d'espèces de tailles très différentes. En effet les espèces de petite taille sont en général trop nombreuses pour être étudiées sur les places-échantillons prévus pour les espèces de grande taille.

La répartition de l'effort d'échantillonnage entre les différents niveaux dépend des coûts d'échantillonnages (c_1 pour une grappe, c_2 pour un grain) et de la variabilité des données aux différents niveaux (variances inter-grappe s_1 et intra-grappe s_2). Des relations, fonctions de ces paramètres, donnant le nombre optimum de grains par grappe peuvent être, selon les cas, mises au point (voir par exemple, Droesbeke *et al.*, 1987).

Les logiciels les plus répandus ne permettent pas de calculer directement les divers estimateurs de ce plan. Mais la programmation que nécessite ces calculs est loin d'être inacessible aux non-numériciens.

2.5 Expression synthétique d'estimateurs usuels

Soit une population statistique de N unités numérotées de 1 à N.

Soit $\mathcal{Y}$ la caractéristique à étudier, prenant la valeur Y_i sur l'unité i, $i =$

$1, \ldots, N$.

Les Y_i sont bien-sûr inconnues avant échantillonnage, et le but de l'échantillonnage est d'estimer la moyenne $\overline{Y}$ des Y_i.

Plaçons nous dans le cadre $\mathcal{C}$ le plus courant, défini de la façon suivante :

$\mathcal{C}$: Plan d'échantillonnage à taille fixée n sans remise, sans probabilité d'inclusion nulle, et modèle de population fixe.

Notons α_i la probabilité d'inclusion de l'unité i dans l'échantillon, et α_{ij} la probabilité d'inclusion simultanée de i et j dans l'échantillon.

Alors, l'estimateur de Horvitz-Thompson Y_{HT} (voir, par exemple, Chaudhuri et Vos, 1988) fournit une estimation sans biais de $\overline{Y}$:

$$Y_{HT} = \sum_{i \in s} \frac{Y_i}{N \alpha_i}, \qquad s \quad \text{étant l'échantillon observé.}$$

Cette formule synthétique résume toutes les formules d'estimateurs sans biais présentées dans la littérature, dans le cadre C défini plus haut [1]. En particulier, si les probabilités d'inclusion sont toutes égales (elles valent alors n/N), Y_{HT} correspond tout simplement à la moyenne d'échantillon :

$$\sum_{i \in s} \frac{Y_i}{n}.$$

La variance de Y_{HT} peut s'exprimer de deux façons :

$$V_1 = \sum_{i=1}^{N} \frac{Y_i^2 (1 - \alpha_i)}{N^2 \alpha_i} + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{Y_i Y_j (\alpha_{ij} - \alpha_i \alpha_j)}{N^2 \alpha_i \alpha_j}$$

$$\text{ou} \qquad V_2 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (\alpha_i \alpha_j - \alpha_{ij}) \left(\frac{Y_i}{\alpha_i} - \frac{Y_j}{\alpha_j} \right)^2$$

A partir d'un échantillon observé s, on a également de manière synthétique l'expression des estimateurs sans biais de V_1 et V_2. On obtient respectivement :

$$\hat{v}_1 = \sum_{i \in s} \frac{Y_i^2 (1 - \alpha_i)}{N^2 \alpha_i^2} + \sum_{i,j \in s, j \neq i} \frac{Y_i Y_j (\alpha_{ij} - \alpha_i \alpha_j)}{N^2 \alpha_{ij} \alpha_i \alpha_j}$$

et

$$\hat{v}_2 = \frac{1}{2N^2} \sum_{i,j \in s, j \neq i} \frac{(\alpha_i \alpha_j - \alpha_{ij})}{\alpha_{ij}} \left(\frac{Y_i}{\alpha_i} - \frac{Y_j}{\alpha_j} \right)^2 \qquad \text{(estimateur de Yates-Grundy).}$$

1. Le lecteur pourra consulter les tableaux fournis au chapitre 2 dans Frontier (1983).

2.6 Bibliographie

CHAUDHURI, A. AND VOS, J. W. E. (1988) Unified theory and strategies of survey sampling. North-Holland. 414 pages.

Ouvrage de référence parmi les plus récents. Le livre est construit en deux parties. La première concerne les problèmes d'inférence statistique et d'échantillonnage spécifiques aux populations finies. La seconde passe au crible les différents plans d'échantillonnage discutés dans la littérature et les méthodes d'estimation de paramètres en population finie qui y correspondent.

COCHRAN, W. G. (1977) Sampling techniques. Wiley. 3*ième* édition. 428 pages.

Ouvrage de référence. Il présente la démonstration mathématique de nombreux résultats.

CORMACK, R. M., PATIL, G. P. AND ROBSON, D. S. (1979) Sampling biological populations. Statistical Ecology, vol 5. I.C.P.H. 392 pages.

Présentation de diverses méthodes d'échantillonnage spécifiques aux études statistiques en écologie : transects, lignes d'interception, observations de quadrats, capture-recaptures, ...

DEROO, M. ET DUSSAIX, A. M. (1980) Pratique et analyse des enquêtes par sondage. P.U.F. 302 pages.

Cet ouvrage traite de nombreux problèmes, principalement les sondages, mais aussi les tests d'hypothèses, les plans d'expériences, les analyses multidimensionnelles *etc.* Les développements mathématiques demeurent simples. La présentation du sujet est relativement originale. Le public visé est d'abord celui qui s'occupe des études de marché.

DESABIE, J. (1966) Théorie et pratique des sondages. Dunod. 481 pages.

Ouvrage de référence, notamment pour les sondages à plusieurs degrés.

DROESBEKE, J-J., , B., ET TASSI, P. (1987) Les sondages. Economica, ASU. 310 pages.

Ouvrage résultant des Journées d'Etude de l'Association des Statisticiens Universitaires, organisées en 1986. Présentation mathématique de la théorie des sondages avec des exemples d'applications de nature socio-économique.

FRONTIER, S. *et al.* (1983) Stratégies d'échantillonnage en écologie. Masson-P.U.L. 494 pages.

Seule la première partie (Théorie de l'échantillonnage écologique) et plus particulièrement le chapitre 2 (de B. Scherrer, Techniques de sondage en

écologie) correspond à la matière traitée dans les ouvrages spécialisés. La seconde partie est une analyse de cas concrets. Les exemples choisis souffrent d'un certain déséquilibre (6 chapitres sur les milieux aquatiques, 1 seul sur le milieu terrestre).

GOURIEROUX, C. (1981) Théorie des sondages. Economica, ESA. 272 pages.

Livre discutant les problèmes de mise en oeuvre des enquètes par sondages en socio-économie mais présentant surtout les liens mathématiques entre théorie des sondages et mathématique statistique.

GROSBRAS, J. (1987) Méthodes statistiques des sondages. Economica, ESA. 331 pages.

Présentation d'une vaste panoplie d'outils permettant de guider la réflexion dans les problèmes de collecte d'information.

HAJEK, J. (1981) Sampling from a finite population. Marcel Dekker. 247 pages.

Ouvrage de référence nécessitant une bonne base en mathématiques.

HERNIAUX, G. (1971) Initiation aux sondages. Masson. 162 pages.

Ce tout petit livre est en réalité une introduction élémentaire aux probabilités et aux statistiques (apparemment destinée à l'origine aux étudiants d'IUT). Il est rédigé de manière claire, pédagogique et mnémotechnique. Très utile pour ceux qui veulent connaître les concepts de base de la statistique sans se perdre dans les mathématiques.

JESSEN, R. J. (1978) Statistical survey techniques. Wiley. 520 pages.

Ouvrage de référence avec beaucoup d'exemples numériques.

KONIJN, H. S. (1973) Statistical theory of sample survey design and analysis. North Holland-Elsevier. 429 pages.

Ouvrage de référence. L'aspect mathématique est trés développé et les démonstrations des principaux résultats de la théorie de l'échantillonnage sont données.

KRISHNAIAH, P. R. AND RAO, C. R. (1988) Sampling. Handbook of Statistics, vol 6. North-Holland. 594 pages.

Ouvrage présentant les aspects théoriques et méthodologiques de la théorie de l'Echantillonnage sur leurs angles les plus récents. Les différents chapitres du livre sont signés par des spécialistes de renommée internationale.

PATIL, G. P., PIELOU, E. C. AND WATERS, W. E. (1981) Sampling and modeling biological populations and population dynamics. Statistical Ecology, vol 2. I.C.P.H. 420 pages.

Présentation de concepts et de méthodologies de statistique mathématique pour l'étude de systèmes écologiques. Illustrations à partir de nombreux exemples biologiques.

RAJ, D. (1972) The design of sample surveys. Mc Graw-Hill. 390 pages.

Présentation qualitative (aucun développement mathématique). Les chapitres 11 à 20 présentent des applications concrètes en agriculture, démographie, emploi, santé *etc.*

STUART, A. (1976) Basic ideas of scientific sampling. Charles Griffin and Co. (Londres). Seconde édition. 106 pages.

Petit livre d'introduction à l'échantillonnage, totalement dépourvu de mathématiques. Assez complet.

SUDMAN, S. (1976) Applied sampling. Academic Press. 249 pages.

Présentation complète mais dépourvue de tout appareil mathématique. Les exemples sont sociologiques et démographiques.

Chapitre 3

A propos de l' échantillonnage systématique

Plan :

3.1 Introduction

L'échantillonnage systématique d'individus dans une population est dans bien des cas plus simple et plus pratique que d'autres méthodes de sondage. En particulier, la procédure de récolte des données est facilitée par le fait que seul le premier individu à sonder est tiré au hasard. Il en est de même pour la programmation de l'observation et la répartition des tâches entre observateurs. De plus, pour certains modèles de superpopulation, ce plan de sondage permet un gain substantiel en précision. Cependant, il possède des inconvénients certains : difficulté d'estimation de la variance des estimateurs dans le cadre de population fixe (Madow et Madow, 1944; Madow, 1949; Raj, 1965; Iachan, 1980c; Wolter, 1984); possibilités d'imprécision importantes, dans certaines situations de périodicités ou gradients par exemple (Cochran, 1946; Bellhouse

et Rao, 1975; Bellhouse, 1981b; Scherrer, 1983).

Ce mode d'échantillonnage a souvent été comparé à l'échantillonnage aléatoire simple et à l'échantillonnage stratifié sous divers types d'hypothèses sur la population étudiée (Cochran, 1946; Madow, 1953; Milne, 1959; Payandeh, 1970). Il se trouve que l'optimalité de ce mode de sondage est fortement liée à la structure de la population (Cochran 1977, Bellhouse 1981a, Iachan 1983). Certains auteurs, par contre, se sont attachés à trouver des variantes de l'échantillonnage systématique permettant de conserver en partie les avantages pratiques qui lui sont reconnus tout en lui enlevant ses inconvénients théoriques (Hartley, 1966; Singh *et al.*, 1968; Singh et Singh, 1977; Zinger, 1980; Iachan, 1983). Dans ce qui suit, nous nous proposons donc de mettre en évidence les avantages et dangers de l'échantillonnage systématique à l'aide des résultats théoriques présentés dans la littérature mais aussi à partir d'exemples biologiques illustrant bien diverses caractéristiques de cette procédure d'échantillonnage.

3.2 Notations et Définition

On se placera toujours dans ce qui suit dans le cadre de population de taille finie.

La taille de la population sera notée N, celle des échantillons n et celle des strates k. Ici, on entend par strates les sous-ensembles de la population discutés au § 2.4.2.

On notera y_i la valeur de la caractéristique pour le i^e individu, $i = 1, \ldots, N$

$$\text{et} \quad \overline{Y} = \sum_{i=1}^{N} y_i/N \qquad \text{la moyenne de la population.}$$

Pour plus de clarté, on utlilisera parfois, dans ce qui suit, la notation à deux indices suivante :

$$y_{lj} = y_{l+(j-1)k} \quad (l = 1, \ldots, k; \quad j = 1, \ldots, n)$$

et on supposera que $N = nk$ (FIG. 3.1).

Un échantillon systématique est tiré en choisissant au hasard un individu parmi les k premiers puis en choisissant chaque k^{ieme} individu, par la suite (FIG. 3.2). Soit h l'indice du premier individu tiré ($1 \leq h \leq k$), l'échantillon observé sera donc :

$$C_h = \{y_{hj}; \quad j = 1, \ldots, n\}.$$

On voit donc que l'échantillonnage systématique nécessite le découpage de la population en n strates $\{y_1, \ldots, y_k\}, \{y_{k+1}, \ldots, y_{2k}\}, \ldots$ et que la procédure de sondage met en relief le découpage de la population en k sous-ensembles qui sont les échantillons réalisables $C_1, \ldots, C_k$.

Dans le cas $N \neq nk$, l'approche est quelque peu différente (voir par exemple, Iachan, 1982). Ainsi, si l'on a des unités statistiques alignées, on les placera virtuellement sur un cercle, l'une d'entre elles sera tirée au hasard puis chaque k^{ie} unité dans un sens donné de la circonférence du cercle, sera ensuite choisie afin de composer l'échantillon (FIG. 3.3).

1	2	...	k	$k+1$	$k+2$	...	$2k$	...	...	$N-k$	$N-k+1$	...	N
$1,1$	$2,1$	...	$k,1$	$1,2$	$2,2$	...	$k,2$	...	...	$1,n$	$2,n$	...	k,n

FIG. 3.1 - *Deux types de notations indicielles pour N $(= nk)$ unités alignées. 1^e ligne : unités notées en séquence ; 2^{de} ligne : notation à deux indices.*

|		•	|		•	|	...	...	|		•	|
$1,1$		$h,1$	$k,1$		$h,2$	$k,2$	...	...	$1,n$		h,n	k,n	

FIG. 3.2 - *Echantillonnage systématique unidimensionnel.*

On a N $(= nk)$ unités, | correspond à la première unité de chacune des n strates et les • correspondent aux unités échantillonnées.

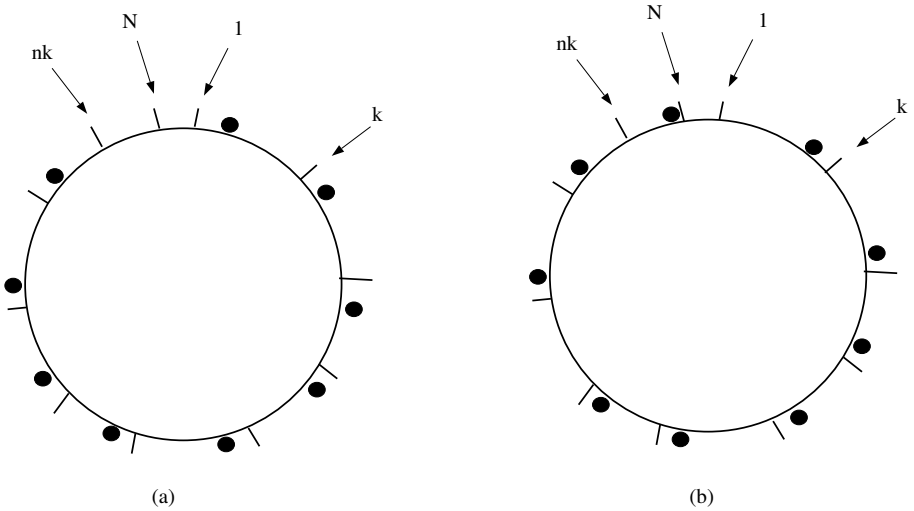


FIG. 3.3 - *Choix des unités à échantillonner quand $N \neq nk$:*
(a) pas d'unité observée entre la nk^{ieme} et la N^{ieme},
(b) une unité est observée entre la nk^{ieme} et la N^{ieme}. Les • correspondent aux unités échantillonnées.

3.3 Liaison avec l'échantillonnage en grappe

L'échantillonnage systématique peut être assimilé à un échantillonnage en grappe où l'on ne prélève qu'une seule grappe C_h. En effet, une fois qu'on a déterminé le point de départ de l'échantillonnage c'est-à-dire tiré au hasard un individu dans la première strate, les $n-1$ autres individus de notre échantillon sont alors déja définis. Le terme grappe est ici pris au sens large dans la mesure où habituellement on l'utilise pour caractériser des individus proches géographiquement ou se succédant sur une liste. On voit donc que l'échantillonnage systématique revient donc à choisir au hasard une grappe C_h parmi les k grappes $C_1, \ldots, C_k$ de même taille n. Ceci permet de mieux comprendre l'impossibilité d'estimer la variance d'échantillonnage en population fixe c'est-à-dire sans supposition distributionnelle sur la population étudiée.

3.4 Estimation de la moyenne de la population

On estime la moyenne $\bar{Y}$ de la population par la moyenne de l'échantillon notée y_{sy}. Cet estimateur est sans biais et sa variance est :

$$Var(y_{sy}) = \frac{1}{k}\sum_{l=1}^{k}(\overline{Y_{l.}} - \overline{Y})^2$$

où

$$\overline{Y_{l.}} = \frac{1}{n}\sum_{j=1}^{n} y_{lj}.$$

$Var(y_{sy})$ est, comme on l'a déja dit, impossible à estimer si l'on n'a pas d'information *a priori* sur la caractéristique étudiée permettant de se placer dans le cadre de modèle de superpopulation. Cependant, une variante du sondage aléatoire systématique consistant à tirer au hasard non plus un individu dans la première strate mais m individus ($1 \leq m \leq k$) permet d'estimer la variance de y_{sy} avec cependant peu de précision dans la pratique car m est en général faible (Gautschi, 1957).

3.5 Comparaison entre sondages aléatoire simple, systématique, et stratifié

3.5.1 Modèle de population fixe

Intuitivement, l'échantillonnage systématique semblerait plus précis que l'échantillonnage aléatoire simple dans la mesure où il évite les redondances

d'information dues aux éventuelles proximités entre individus tirés. En fait, cette opinion est basée sur une idée préconçue sur l'allure, les caractéristiques de la population. L'efficacité du sondage systématique par rapport à l'aléatoire simple ou le stratifié (dans tout ce chapître, il s'agira du sondage stratifié avec un tirage par strate) dépend, évidemment, des propriétés de la population étudiée. Une connaissance de la structure de la population est donc requise pour comparer l'échantillonnage systématique à d'autres procédures de sondage. Dans le cadre du modèle de population fixe, on a le résultat suivant : Si la variance intra-grappes (telles que définies plus haut c'est-à-dire intra échantillons possibles) est supérieure à la variance de la population alors le sondage systématique est meilleur que le sondage aléatoire simple.

3.5.2 Modèle de superpopulation

Notons $E_M(X)$ et $Var_M(X)$ (respectivement $E_P(X)$ et $Var_P(X)$) l'espérance et la variance d'une variable quelconque X quand l'aléa est dû au modèle de superpopulation M (respectivement à la procédure de sondage P).

Quand on considère la variance de la moyenne d'échantillon $\overline{y}$, pour un modèle de superpopulation M et une procédure de sondage P à taille fixe n, non informative et où chaque individu a la même probabilité d'inclusion n/N (c'est le cas des trois procédures que l'on compare), alors on a le résultat suivant :

$$Var_{M,P}(\overline{y}) = E_M(Var_P(\overline{y})) + Var_M(E_P(\overline{y}))$$

or $E_P(\overline{y}) = \overline{Y}$ (car les probabilités d'inclusion des individus sont égales)

donc $Var_M(E_P(\overline{y})) = Var_M(\overline{Y})$ est indépendant du plan P.

Il revient donc au même de prendre $E_M(Var_P(\overline{y}))$ au lieu de $Var_{M,P}(\overline{y})$ comme critère de comparaison de nos trois procédures d'échantillonnage. Le critère qu'on utilise souvent, par commodité, est donc $E_M(Var_P(\overline{y}))$. Le plan d'échantillonnage sera considéré comme étant d'autant meilleur que ce critère sera faible. On le notera :

C_{alea} pour l'échantillonnage aléatoire simple,

C_{syst} pour l'échantillonnage systématique,

C_{stra} pour l'échantillonnage stratifié aléatoire.

3.5.2.1 Individus non corrélés

Soit $y_i, \quad i = 1, \ldots, N$ la réalisation de la variable aléatoire étudiée sur le i^e individu. On pose :

$E_M(y_i) = \mu_i$ ou encore $E_M(y_{lj}) = \mu_{lj}$ en cas de notation double indice.

La non autocorrélation signifie que

$$E_M[(y_i - \mu_i)(y_j - \mu_j)] = 0 \quad \forall\, i \neq j$$

D'autre part, notons

$$\sigma_i^2 = Var_M(y_i), \quad \bar{\mu}_{l.} = \sum_{j=1}^{n} \mu_{lj}/n, \quad \bar{\mu}_{.j} = \sum_{l=1}^{k} \mu_{lj}/k, \quad \bar{\mu} = \sum_{j=1}^{n}\sum_{l=1}^{k} \mu_{lj}/N,$$

respectivement la variance de la caractéristique pour le i^e individu et son espérance pour la moyenne de: l'échantillon l, la strate j, toute la population.

On a les résultats suivants:

$$C_{syst} = (1 - \frac{n}{N}) \sum_{i=1}^{N} \frac{\sigma_i^2}{N} \quad + \quad \frac{1}{k} \sum_{l=1}^{k} (\bar{\mu}_{l.} - \bar{\mu})^2$$

$$C_{alea} = (1 - \frac{n}{N}) \sum_{i=1}^{N} \frac{\sigma_i^2}{N} \quad + \quad \frac{N-n}{Nn} \sum_{l=1}^{k}\sum_{j=1}^{n} (\mu_{lj} - \bar{\mu})^2/(N-1)$$

$$C_{stra} = (1 - \frac{n}{N}) \sum_{i=1}^{N} \frac{\sigma_i^2}{N} \quad + \quad \frac{1}{Nn} \sum_{l=1}^{k}\sum_{j=1}^{n} (\mu_{lj} - \bar{\mu}_{.j})^2/N$$

On voit donc que:

1. Dès lors que les $y_i \quad i = 1, \ldots, N$ sont non corrélées et ont la même espérance μ, alors:

   $$C_{alea} = C_{sys} = C_{stra}.$$

   C'est en particulier le cas quand la population est en ordre aléatoire (équiprobabilité des permutations d'individus).

2. En cas de gradient linéaire, c'est-à-dire $\mu_i = \alpha + \beta_i$, alors

   $$C_{stra} \leq C_{sys} \leq C_{alea}$$

   avec inégalité stricte dès que $n > 1$, ce qui est bien-sûr toujours le cas.

Il est à noter que de nombreuses procédures *quasi-systématiques* ont été proposées dans la littérature (voir Bellhouse et Rao, 1975) dans le cas de populations possédant un gradient linéaire. L'idée est de conserver l'aspect pratique du sondage systématique tout en gagnant en précision d'estimation.

3.5.2.2 Populations autocorrélées

Certains auteurs ont démontré que lorsque les individus d'une population possédaient même espérance, même variance, et une certaine forme de corrélation mutuelle (corrélogramme positif décroissant et convexe, ce qu'on peut interpréter comme une forte similarité entre observations sur des individus proches, cette similarité diminuant rapidement entre individus de plus en plus éloignés les uns des autres), alors l'échantillonnage systématique en une dimension était meilleur que l'échantillonnage stratifié lui même meilleur que l'échantillonnage aléatoire simple.

$$C_{sys} \leq C_{stra} \leq C_{alea}.$$

Par contre, quand on passe à deux dimensions, les choses se compliquent un peu (Quenouille, 1949; Payandeh, 1970; Bellhouse, 1981a et 1981b).

3.6 Estimation de la variance d'échantillonnage

On a vu précédemment qu'aucun estimateur valable de la variance de la moyenne $\overline{y}_{sy}$ n'existait dans le cadre de population fixe. Cependant, avec davantage d'information sur la population, on peut obtenir un estimateur de la variance de $\overline{y}_{sy}$. Par exemple, sous l'hypothèse de population en ordre aléatoire, un estimateur sans biais de la variance de $\overline{y}_{sy}$ est :

$$\frac{N-n}{Nn}\sum_{i=1}^{n}(y_i - \overline{y}_{sy})^2/(n-1)$$

C'est en fait l'estimateur de la variance d'échantillonnage pour un échantillon aléatoire simple : Pour une population en ordre aléatoire, tous les échantillonnages à taille fixée n sont équivalents et donc équivalents à l'échantillonnage aléatoire simple.

Sous certains autres modèles, il est également possible d'obtenir des estimateurs de la variance de $\bar{y}_{sy}$. Le problème est alors celui de la robustesse de ces estimations par rapport au modèle en question. Comme on l'a déja dit précédemment, en choisissant non plus une grappe au hasard parmi les k grappes $C_1, \ldots, C_k$ mais m grappes ($1 \leq m \leq k$) toujours au hasard (et sans remise), on peut estimer la variance d'échantillonnage par :

$$\frac{k-m}{m(k-1)}\sum_{l=1}^{m}(\bar{y}_{l.} - \bar{y}_{sy})^2/(m-1)$$

où $\bar{y}_{l.}$ est la moyenne de la grappe l.

La précision est faible quand m est petit puisqu'on n'a que $m-1$ degrés de liberté pour effectuer cette estimation .

La question qu'on est en droit de se poser avec l'échantillonnage systématique avec plusieurs points de départ (on l'appellera systématique multi-origine) est sa précision par rapport à l'échantillonnage systématique classique (uni-origine), pour des tailles d'échantillons égales. Gautschi (1957) prouve les résultats suivants : Le systématique uni-origine est plus précis que le systématique multi-origine en cas de population non autocorrélée ayant un gradient linéaire et en cas de population ayant une fonction d'autocorrélation positive, décroissante et convexe (la précision étant la même si cette fonction est constante).

3.7 Quelques exemples

Les exemples présentés ci-dessous sont tirés de l'article de Scherrer (1983).

Exemple 1

Douglas(1977), voulant étudier la dynamique d'une population de campagnols, a défini deux aires d'échantillonnage de 150m x 375m. Il y a pratiqué l'échantillonnage systématique car cette procédure facilitait les opérations d'installation et de repérage des pièges et permettait, qui plus est, de déduire l'étendue du domaine vital des individus. Ainsi, sur chacune des aires concernées, l'auteur a placé une grille de 10 colonnes et 25 lignes espacées de 15m les unes des autres, puis il a placé 250 pièges à petits mammifères de façon systématique dans chaque case de la grille.

Exemple 2

L'hirondelle de rivage niche en colonie dans les falaises de sables où chaque couple creuse un terrier pour y construire son nid. Pour des raisons mécaniques, les orifices des tunnels sont souvent alignés au niveau des couches stratigraphiques de sable fin et cet agencement facilite généralement le dénombrement et la sélection des cavités. Aussi, pour étudier le taux d'éclosion des oeufs ou tout autre paramètre relatif à la nidification, est-il facile d'appliquer l'échantillonnage systématique : Il est plus commode de sélectionner un trou sur 5,7 ou 10 que de classer par ordre croissant n nombres aléatoires différents compris entre 1 et N et de visiter les nids correspondant à ces numéros.

Exemple 3

Pour vérifier si l'orignal cherche à diversifier son régime alimentaire en sélectionnant dans son domaine vital des espèces peu disponibles, des nutritionnistes ont utilisé l'échantillonnage systématique pour mesurer la disponibilité de nourriture. Sur un site de 25 hectares, aux limites arbitraires, 89 quadrats de 7,5 m^2 ont été disposés régulièrement tous les $\sqrt{25.10^4/89} = 53$m environ, le long de lignes parallèles distantes de 53m environ. Etant donné la superficie de la parcelle, un intervalle d'environ 50m séparait chaque parcelle-échantillon sur lesquelles les données relatives à la disponibilité étaient recueillies.

3.8 Résumé

L'échantillonnage systématique est attrayant car de mise en oeuvre et d'exécution relativement simple. Il est dans certains cas plus précis que l'échantillonnage stratifié (à un tirage par strate) et que l'échantillonnage aléatoire simple. Ses défauts sont le biais en cas de population présentant certaines périodicités et, surtout, l'impossibilité d'estimer la variance d'échantillon quand on n'a pas d'information a priori sur la population permettant une modélisation probabiliste. D'une façon générale, il est délicat de recommander cette procédure d'échantillonnage plutôt qu'une autre quand on ne dispose pas de renseignement fiable sur la population à sonder. On peut toutefois conseiller ce mode de sondage si l'on a de fortes présomptions d'être dans les situations suivantes :

1. population distribuée ou ordonnée en majeure partie de façon aléatoire. L'échantillonnage systématique sera dans ce cas appliqué pour sa commodité pratique. On pourra même estimer sans biais la précision de l'échantillonnage.

2. population à autocorrélation positive, décroissante et convexe. Cette forme de liaison entre individus est assez courante dans la réalité. Dans ce cas, l'échantillonnage systématique est meilleur que le stratifié ou l'aléatoire simple puisqu'il permet d'éviter la redondance d'information due à l'observation d'individus trop proches les uns des autres.

3.9 Bibliographie

BELLHOUSE, D. R. (1981a): Area estimation by point-counting techniques. Biometrics 37, 303-312.

BELLHOUSE, D. R. (1981b): Spatial sampling in the presence of a trend. J. Statist. Plan. Infer., 5, 365-375.

BELLHOUSE, D. R. AND RAO, J. N. K. (1975): Systematic sampling in the presence of a trend. Biometrika 62, 694-697.

COCHRAN, W. G. (1946): Relative accuracy of systematic and stratified random samples of a certain class of populations. Ann. Math. Statist., 17, 164-177.

COCHRAN, W. G. (1977): Sampling techniques. Chap. 8. Wiley. 3*ième* édition.

DOUGLAS, R.J. (1977): Population dynamics, home ranges and habitat association of yellow-checked vole in North-east territories. Can. Field Nat., 91, 237-243.

GAUTSCHI, W. (1957): Some remarks on systematic sampling. Ann. math. Statist. 28, 385-394.

HARTLEY, H. O. (1966) : Systematic sampling with unequal probabilities and without replacement. JASA, 61, 739-748.

IACHAN R. (1980a) : An asymptotic theory of systematic sampling I. University of Wisconsin Tech. Report 616.

IACHAN R. (1980b) : An asymptotic theory of systematic sampling II. University of Wisconsin Tech. report 617.

IACHAN R. (1980c) : An asymptotic theory of systematic sampling III. Systematic sampling variance estimators. University of Wisconsin Tech. report 618.

IACHAN R. (1982): Systematic sampling : A critical review. Intern. Stat. Review 50, 293-303.

IACHAN R. (1983): Multiple random start systematic sampling. University of Wisconsin Tech. Report.

MADOW, W. G., AND MADOW, L. H. (1944) : On the theory of systematic sampling I. Ann. Math. Statist., 15, 1-15.

MADOW, W. G. (1949) : On the theory of systematic sampling II. Ann. Math. Statist., 20, 333-354.

MADOW, W. G. (1953) : On the theory of systematic sampling III. Comparison of centered and random start systematic sampling. Ann. Math. Statist., 24, 101-106.

MILNE, A. (1959) : The centric systematic area-sample treated as a random sample. Biometrics, 15, 271-297.

PAYANDEH, B. (1970) : Relative efficiency of two-dimensional systematic sampling. Forest Sci., 16, 271-276.

QUENOUILLE, M. H. (1949): Problems in plane sampling. Ann. Math. Statist., 20, 355-375.

RAJ, D. (1964) : The use of systematic sampling with probability proportionate to size in a large scale survey. JASA, 59, 251-255.

RAJ, D. (1965) : Variance estimation in randomized systematic sampling with probability proportionate to size. JASA, 60, 278-284.

SCHERRER, B. (1983): Techniques de sondage en écologie. Chap. 2. Collection d'écologie 17, Masson.

SINGH, D., JINDAL, K. K. AND GARG, J. N. (1968): On modified systematic sampling. Biometrika, 55, 541-546.

SINGH, D. AND SINGH, P. (1977): New systematic sampling. J. Statist. Plan. Infer., 1, 163-177.

WOLTER,, K. M. (1984): An investigation of some estimators of variance for systematic sampling. JASA, 79, 781-790.

ZINGER, A. (1964): Systematic sampling in Forestry. Biometrics, 20, 553-565.

ZINGER, A. (1980): Variance estimation in partially systematic sampling. JASA 75, No 369, 206-211.

Chapitre 4

Echantillonnage en vue d'etudier des répartitions spatiales d'individus en écologie

Plan :

4.1 Introduction

La formalisation mathématique d'un problème d'échantillonnage permet parfois de mieux situer ce problème, en particulier par rapport aux résultats sur l'admissibilité ou l'optimalité de stratégies d'échantillonnage (Cassel *et al.*, 1977) obtenus dans des situations classiques (par exemple unités statistiques échangeables, indépendantes ...). Il existe en effet une série d'ouvrages sur la théorie de l'échantillonnage en liaison avec la théorie de la décision statistique (entre autres Cassel *et al.*, 1977; konijn, 1973); mais l'incidence des objectifs poursuivis sur la procédure de sondage à employer n'est pas en général facile à appréhender et les méthodologies statistiques à employer sur les données

récoltées sont sous-jacentes aux protocoles de sondage utilisés. De plus, en écologie, le coût moyen d'échantillonnage intervient souvent dans le choix d'une stratégie. Il se trouve alors que le critère de précision de l'échantillon qui est généralement à la base des comparaisons de procédures de sondage dans la littérature, ne donne pas entièrement satisfaction. L'utilisation de fonctions de perte (cf. Cochran, 1977; § 4.10 et § 5A.4), tenant compte du coût espéré d'échantillonnage, peut être une solution mais elle n'est pas toujours possible. Certains facteurs importants sont, en effet, souvent méconnus (efficacité de l'échantillonneur par exemple). Nous verrons dans ce qui suit ce qui se passe précisément dans le cas d'études de répartition spatiale d'individus.

En fait, il y a différentes manières d'aborder l'échantillonnage d'une population statistique dans le but d'étudier une ou plusieurs de ses caractéristiques. Rappelons tout d'abord l'importante distinction à faire entre population statistique et population biologique :

- la population statistique est l'ensemble des unités d'échantillonnage (on dit plus couramment unités statistiques) parmi lesquelles certaines seront choisies pour former l'échantillon. Il pourra s'agir d'unités naturelles (plantes, arbres, nids, mares, ...) ou artificielles (unités de surface d'un champ, placettes délimitées arbitrairement, pièges ...),

- la population biologique ou population cible est l'entité écologique à laquelle on s'intéresse, et pour laquelle on désire évaluer certains paramètres, par exemple ceux liés à sa répartition spatiale ou à son évolution dans le temps.

Le tableau TAB. 4.1 nous présente différents cas de figures possibles. Il est à noter que, dans certaines situations, population biologique et population statistique peuvent donc coïncider.

D'autre part, il existe en théorie de l'échantillonnage une terminologie concernant ces deux types de population :

Pour la population statistique, on peut parler de population finie ou infinie selon que les unités statistiques soient en nombre fini ou pas, choisir un modèle de population fixe ou un modèle de population aléatoire (ou superpopulation). Comme on l'a vu au chapitre 1, le modèle de population fixe revient à considérer les valeurs de la caractéristique étudiée comme étant fixées mais inconnues pour chaque unité statistique tandis que pour le modèle de *superpopulation* , ce sont des réalisations de variables aléatoires. Le modèle de superpopulation fournit donc souvent un cadre plus intéressant pour l'étude de répartitions d'individus en écologie. C'est en particulier le cas pour l'étude des infestations en pontes de pyrale (*Ostrinia nubilalis*) d'un champ de maïs, comme nous le verrons par la suite.

Population statistique	**Unité statistique**	**Population cible**	**Caractéristique étudiée**
Ensemble des plantes de maïs d'un champ	Plante de maïs	Pontes de pyrale présentes dans le champ	Nombre d'oeufs parasités par ponte
Ensemble de parcelles obtenu par quadrillage régulier d'un champ de tournesol	Parcelle	Ensemble des plantes de Tournesol du champ	Nombre de plantes atteintes de crispation (due à l'action d'aphides)
Ensemble des arbres d'une forêt	Arbre	Ensemble des arbres de la forêt	Type de l'arbre (feuillu ou conifère)

TAB. 4.1 - *Exemples de paire (population statistique, population cible).*

Pour ce qui concerne la population biologique, on parle de *population fermée* quand les phénomènes de migration, de mortalité et de naissance d'individus sont négligeables pendant la période d'étude dans la région observée. Dans le cas contraire, on dit que la *population est ouverte* et les procédures d'analyse statistique sont en général plus complexes et délicates que pour une population fermée car les arrivées et disparitions d'individus doivent être prises en compte.

En général, l'étude de quelques échantillons *d'investissement* fournit une information a priori très utile pour la construction de protocoles de sondage *de routine* dans le but d'estimer ou prédire des niveaux de densité de population. Supposons par exemple qu'on arrive à modéliser correctement la façon dont se répartit une certaine espèce d'insectes sur les cultures de blé. Si ce modèle probabiliste nous donne les fréquences théoriques du nombre d'individus en divers points de la région étudiée à partir de très peu de paramètres, cette connaissance a priori nous permettra en général d'estimer la densité d'occupation de l'espèce en question, de façon plus fiable (Bliss et Fisher, 1953; Bliss et Owens, 1958).

Dans la suite de ce chapitre, on se propose donc de présenter brièvement quelques méthodes d'étude de répartitions spatiales d'individus, puis la façon dont fut abordée l'étude statistique des répartitions spatiales des pontes de pyrale à partir de données recueillies en 1980 et 1984 dans la région parisienne.

4.2 Méthodes d'étude de répartitions spatiales d'individus

En vue d'étudier des répartitions aussi bien spatiales que temporelles d'individus en écologie, les travaux concernant l'analyse statistique des processus ponctuels[1] fournissent une base d'étude intéressante (Bailey, 1964; Renshaw, 1972; Ogata, 1981; Jacobsen, 1982). En ce qui concerne les processus spatiaux, il s' avère que les techniques statistiques d'analyse des processus à support[2] continu sont largement développées (Diggle, 1975; entre autres) tandis que les études statistiques de processus à support discret sont moins nombreuses et s'avèrent pour la plupart non paramétriques (Cliff et Ord, 1981). Il existe cependant des travaux concernant les répartitions spatiales sur réseaux[3] réguliers (Besag, 1974; Guyon, 1985) mais certains phénomènes de répartitions d'individus, et c'est le cas des pontes de pyrale sur champ de maïs, ne peuvent être raisonnablement assimilés à un processus sur réseau régulier (les sites sont en effet les plantes de maïs dont la répartition spatiale est loin d'être régulière).

Comme on l'a déjà dit, l'étude de répartitions spatiales d'individus en écologie peut être abordée de différentes façons puisque les objectifs biologiques qu'on cherche à atteindre déterminent le mode d'échantillonnage, les variables à mesurer et les méthodes statistiques à employer (TAB. 4.2 et FIG. 4.1). Quand on a peu d'informations préalabres sinon aucune sur le phénomène à étudier, des échantillons d'investissement (en général exhaustifs) peuvent permettre le choix et la validation de modèles spatiaux ou de certaines hypothèses sur la répartition spatiale observée. Ces échantillons d'investissement sont souvent d'une certaine lourdeur. Les contraintes de temps et de moyen étant, dans la majorité des cas, importantes, on préfère échantillonner par distance (Diggle, 1975) ou par placettes. Les possibilités du statisticien deviennent alors plus modestes : elles se résument à l'estimation de densité ou du nombre moyen d'occurrences du phénomène par unité statistique, mais aussi à une description qualitative de la structure spatiale revenant souvent à tester l'hypothèse de répartition au hasard ou l'indépendance entre les répartitions de différents caractères. L'hypothèse de répartition aléatoire pure (processus de Poisson) constitue en effet une hypothèse de travail centrale; elle est une hypothèse de base et la construction de son test est, en général, facile. Pourtant, dans la pratique, l'analyse du caractère agrégatif ou répulsif des répartitions engendrées ont plus d'importance que le rejet de l'hypothèse de répartition au hasard. On s'intéresse alors aux échelles d'hétérogénéité ou de dispersion. Les tests basés sur les distances entre points d'échantillonnage et individus ou entre individus concernent essentiellement les petites échelles; ceux basés sur les comptages

1. répartitions aléatoires de points dans un espace

2. le support d'un processus est l'espace dans lequel les points du processus peuvent se trouver

3. un réseau est un ensemble de sites répartis dans un espace, e.g plantes, placettes, ...

par placettes examinent des échelles de l'ordre de la taille des placettes. Si l'on veut donc identifier les différentes échelles d'hétérogénéité et les diverses autocorrélations spatiales (à supposer qu'elles existent), il nous faut alors un échantillonnage plus conséquent voire quasi-exhaustif, qui n'est pas toujours réalisable. Un quadrillage de la région étudiée à partir d'un échantillonnage systématique peut cependant fournir une bonne information sur la répartition spatiale (Debouzie *et al.*, 1987), en tenant compte toutefois des réserves formulées au chapitre 3.

Pour revenir aux méthodologies statistiques elles-mêmes, on peut distinguer globalement, quatre catégories :

1) Celles basées sur la comparaison entre fréquences du nombre d'individus situées dans des placettes et des fréquences théoriques

En faisant certains types de suppositions quant au mécanisme présidant la répartition spatiale d'individus, on peut obtenir la loi théorique du nombre d'individus par placette. L'approche peut être conditionnelle (Besag, 1974a) c'est-à-dire qu'on tient compte du nombre d'individus en un site conditionnellement à ce qui se passe ailleurs (en général sur les sites voisins). On a ainsi une correspondance entre certains types de répartitions d'individus et certaines lois de probabilités discrètes. Les plus classiques sont :

1. Répartition au hasard et loi de Poisson
2. Contagion apparente (due à l'hétérogénéité du milieu naturel) et loi composée (mélange de lois)
3. Contagion vraie (affinité réelle entre individus) et loi généralisée (somme aléatoire de variables aléatoires équidistribuées)

Le problème est que l'ensemble des lois composées et celui des lois généralisées ne sont pas disjoints et que deux mécanismes de nature différente (par exemple, contagion apparente et contagion vraie) peuvent engendrer la même distribution de fréquences théoriques. Cependant, si les suppositions sous-jacentes au modèle probabiliste sont fiables, ces méthodes peuvent s'avérer tout de même intéressantes (Pielou, 1977 Chap. 2; Rogers, 1973 Chap. 3).

2) Celles basées sur des indices d'agrégation ou d'autocorrélation spatiale

La démarche n'est plus d'essayer d'expliquer la répartition spatiale observée mais de tenter de mesurer le degré d'agrégation entre individus avec un certain critère. Pour cela, comme pour la première catégorie de méthodes, des comptages sont effectués dans des placettes contigues ou pas. Par contre, pour les indices d'autocorrélation spatiale (qui peuvent d'ailleurs servir pour d'autres caractéristiques que le nombres d'individus par placettes), il est important d'avoir des placettes contigues ou, sinon, disposées de façon régulière. On peut

alors tester diverses formes d'autocorrélation pour la caractéristique observée dans ces placettes (Cliff and Ord, 1981).

3) Celles basées sur l'observation de distances

Quand l'échantillonnage par placettes pose certaines difficultés théoriques ou pratiques, on peut avoir recours, si l'on dispose d'un support d'observations continu et isotropique, à des méthodes reposant sur des mesures de distances (distance entre individus et point d'origine de l'échantillonnage ou entre individus entre eux). On peut alors estimer la densité de la population et tester certaines hypothèses, telles que la répartition au hasard des individus (Diggle, 1975).

4) Celles basées sur une analyse de variance

Des techniques inspirées de l'analyse de variance hiérarchisée du nombre d'individus sont utilisables quand l'information sur la répartition des individus dans la région étudiée est donnée sous la forme d'une grille de dénombrement (Greig-Smith, 1952; Mead, 1974; Byth, 1982). Elles sont utilisées pour la détection d'échelles d'hétérogénéité dans la répartition spatiale.

Pour étudier une répartition spatiale d'individus en écologie, à partir de données recueillies à cet effet, il existe diverses méthodes statistiques. Ces méthodes dépendent du type :

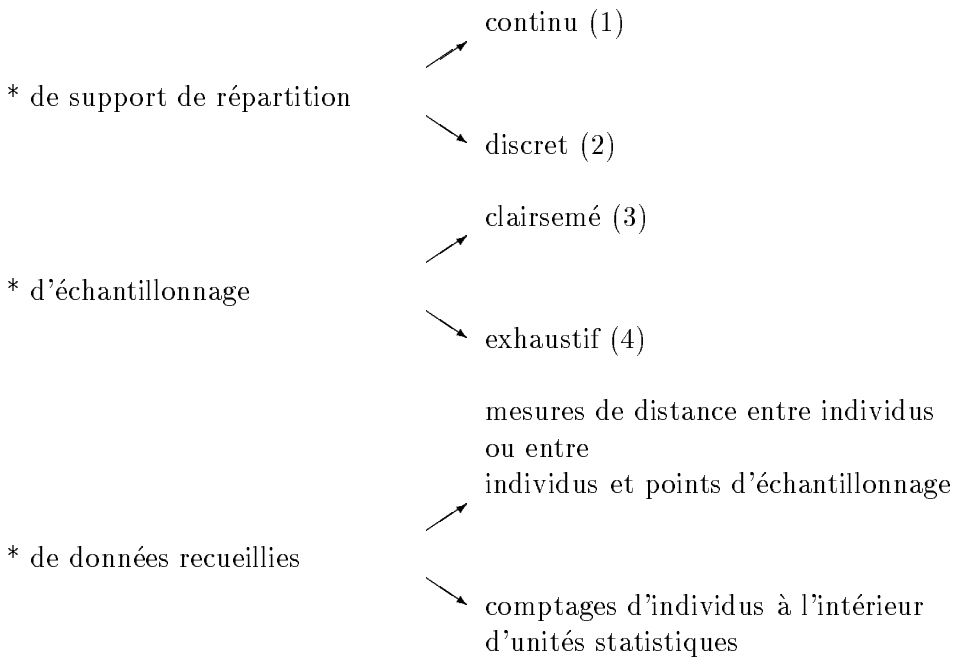


FIG. 4.1 - *Récapitulatif sur les facteurs limitant le choix de méthodes statistiques pour l'étude de répartition spatiale d'individus.*

D'autre part, les objectifs biologiques qu'on se donne, entrent bien sûr en

compte dans la détermination de la méthode statistique à employer.

(1) les individus peuvent être situés en n'importe quel point de la région observée. (2) les individus ne peuvent être qu'en des *sites* particuliers (par exemple, des pontes d'insecte sur des plantes. La disposition spatiale des plantes crée des contraintes sur la répartition des pontes) . (3) les unités statistiques (pièges, placettes ou individus ...) sont implantées ou choisies de façon aléatoire ou systématique. (4) toute la zone géographique à étudier est recensée.

Les procédures que nous avons vues tout au long de ce paragraphe servent à analyser le mode de répartition des individus par le biais de calculs d'indices appropriés et aussi de tests d'hypothèse (TAB. 4.2).

Observations effectuées	Plan d'échantillonnage	Méthodes statistiques envisageables
comptage total	exhaustif	1, 2, 3, 5, 6
	aléatoire systématique	1,2,3,5,6
	aléatoire non systématique	1, 2, 6
mesure de distance	exhaustif, aléatoire simple, systématique	4, 6
notation en présence-absence	exhaustif, systématique	3, 6
	aléatoire simple, stratifié	1, 6

TAB. 4.2 - *Incidence de certaines caractéristiques d'échantillonnage sur la façon d'étudier une répartition spatiale.*

Méthodes 1 : comparaisons de fréquences théoriques et observées. Tests d'ajustement.

Méthodes 2 : calculs d'indices d'agrégation. Tests associés.
Méthodes 3 : calculs d'indices d'autocorrélation. Tests associés.
Méthodes 4 : estimations et tests d'hypothèse basés sur les distances entre individus, ou entre individus et points arbitraires de la région étudiée.
Méthodes 5 : analyses de variance hiérarchisées du nombre d'individus pour blocs emboîtés.
Méthodes 6 : redistributions d'individus ou permutations d'effectifs d'unité d'échantillonnage par simulations numériques. tests de Monte-Carlo associés.

Mais certaines peuvent également être utilisées pour estimer la densité de population. C'est le cas des méthodes basées sur les distances (Diggle, 1977; Byth, 1982) et aussi des méthodes par comptages d'individus sur des unités naturelles (par exemple des plantes) ou artificielles (par exemple, des placettes constituées de regroupement de plantes voisines ou d'unités de surface). Pour ces dernières méthodes, tous les plans d'échantillonnage vus au chapitre 2 peuvent être pratiqués avec les avantages et inconvénients déjà soulignés. Il convient cependant de noter que pour les populations constituées d'animaux très mobiles, les méthodes par comptage ou par distance ne sont pas praticables. Dans de telles situations, les méthodes de capture-recapture (Southwood, 1978; Seber, 1973) sont parfois utilisées. Ces méthodes consistent à capturer dans une région donnée un nombre n_1 d'animaux puis de les marquer et, ensuite, les relâcher. Après un temps suffisamment long pour qu'il y ait eu brassage de la population, on procède à une nouvelle capture de n_2 animaux. Soit m le nombre d'animaux marqués parmi les n_2 recapturés, et soit N le nombre total d'animaux dans la région (nombre qu'on ne connait pas et qu'on cherche à estimer). Si l'on admet que les conditions suivantes sont vérifiées : brassage réel de la population entre la capture et la recapture, pas de disparition de marquage ou d'animaux, pas d'animaux *incapturables*, alors

$$\frac{m}{n_2} \simeq \frac{n_1}{N} \qquad \text{d'où l'estimation suivante pour } N :$$

$$\hat{N} = \frac{n_1 n_2}{m}.$$

La technique de capture-recapture présente l'avantage d'être adaptée à l'étude de certaines populations très mobiles (pratiquées, par exemple pour l'étude d'abondance d'espèces de poisson). Par contre, plusieurs problèmes peuvent se poser : les inconvénients dus au marquage des animaux sont mal connus (on peut imaginer des changements comportementaux comme l'agrégation ou la répulsion); le marquage peut disparaitre entre les deux captures; en cas de phénomène migratoire, les totaux d'individus marqués et non marqués sont modifiés entre les captures.

Une présentation plus détaillée des méthodes de capture-recapture (de nombreuses variantes existent) et des techniques de marquage peut être vue dans Seber (1973).

4.3 Exemple : Pontes de la pyrale du maïs en plein champ

A la suite de ces réflexions sur certains aspects de l'analyse statistique de répartition d'individus en écologie, nous allons maintenant voir brièvement comment furent abordés les problèmes concernant la répartition des pontes de pyrale sur culture de maïs dans le Bassin Parisien lors de la campagne d'observation de 1984 menée par la station de zoologie et le laboratoire de biométrie de Versailles (Vaillant, 1985).

Des parcelles rectangulaires (meilleure détection des probables échelles d'hétérogénéité par rapport à une parcelle carrée, (Cliff and Ord, 1981; Kershaw, 1957)), de près de 1200 plantes chacune, furent observées exhaustivement durant l'été 1984 à 2 dates différentes. Ces parcelles permirent d'acquérir certaines connaissances sur la façon dont les cultures de maïs sont infestées durant l'été : indice de dispersion du nombre de pontes par plante significativement supérieur à un (valeur de cet indice sous l'hypothèse de processus de Poisson) en cas de fortes infestations; disposition des séquences d'infestation au sein des rangées de maïs parfois proche de ce qu'on observerait sous l'hypothèse selon laquelle les plantes seraient infestées indépendamment les unes des autres; moyennes de pontes très différentes d'une rangée à une autre; autocorrélation inter-rangée parfois négative de façon significative pour le nombre de pontes situées sur des groupes de plantes contigues. Avant de présenter plus en détail quelques unes des procédures statistiques employées, précisons la façon dont furent recueillies les données sur lesquelles on a travaillé.

4.3.1 Données récoltées

Les données utilisées sont des observations de pontes effectuées en 1984, les motivations de cette récolte de données étant la mise au point de procédures d'échantillonnage plus fiables pour l'estimation des niveaux d'infestation en pontes de pyrale du maïs. Etant dans une phase essentiellement exploratoire, il convenait de mettre en place des échantillons *d'investissement* afin de tirer le maximum d'information sur les répartitions de pontes de pyrale, d'où le choix suivant :

Toutes dates confondues, 10 parcelles de 3 rangées contigues de près de 60m de long soit au total près de 14000 plantes observées

(Ce type de dispositif permettant, à coût fixé, d'explorer davantage les probables échelles d'hétérogénéité par rapport à une parcelle carrée).

Le tableau TAB. 4.3 nous montre, pour les 30 rangées observées, certaines caractéristiques

descriptives discutées plus loin.

Rg	N	p	μ	σ^2	I_D	N_0	N_1	N_2	N_3	N_4	$\bar{l}$	l_1	l_{max}
1	485	.157	.169	.166	.979	409	70	6	0	0	6.13	12	22
2	513	.150	.166	.174	1.048	436	70	6	1	0	6.54	12	29
3	503	.181	.203	.218	1.074	412	82	8	0	1	5.42	15	26
4	412	.114	.119	.120	1.006	365	46	0	1	0	8.89	7	52
5	634	.142	.151	.148	.975	544	84	6	0	0	6.70	17	42
6	585	.096	.104	.111	1.061	529	51	5	0	0	10.20	10	62
7	314	.248	.283	.293	1.034	236	69	8	0	1	4.01	22	14
8	495	.293	.329	.302	.918	350	129	14	2	0	3.38	50	16
9	449	.283	.334	.330	.988	322	105	21	1	0	3.40	42	15
10	533	.161	.176	.176	.996	447	78	8	0	0	6.19	16	29
11	525	.179	.196	.192	.980	431	85	9	0	0	5.59	17	22
12	548	.201	.219	.212	.966	438	101	8	1	0	4.93	25	17
13	586	.118	.125	.126	1.014	517	66	2	1	0	8.21	8	39
14	529	.102	.106	.102	.967	475	52	2	0	0	9.30	7	35
15	631	.100	.103	.099	.960	568	61	2	0	0	10.02	15	69
16	511	.239	.282	.305	1.081	389	104	14	4	0	4.19	28	13
17	442	.201	.235	.267	1.133	353	77	10	1	1	4.90	18	19
18	455	.213	.237	.230	.968	358	86	11	0	0	4.64	23	22
19	390	.223	.262	.276	1.055	303	73	13	1	0	4.12	25	25
20	390	.259	.321	.383	1.195	289	84	11	5	1	3.79	28	18
21	390	.200	.251	.307	1.221	312	61	14	3	0	4.53	14	21
22	390	.141	.154	.161	1.049	335	51	3	1	0	6.59	12	36
23	390	.215	.251	.266	1.058	306	71	12	1	0	4.66	17	19
24	390	.123	.136	.143	1.055	342	43	5	0	0	8.23	9	28
25	390	.244	.318	.377	1.185	295	68	25	2	0	3.88	34	20
26	390	.267	.333	.377	1.131	286	82	18	4	0	3.55	33	14
27	390	.282	.354	.399	1.127	280	86	21	2	1	3.23	42	23
28	390	.197	.228	.264	1.157	313	69	5	2	1	4.72	16	38
29	390	.213	.262	.307	1.173	307	67	13	3	0	4.55	19	25
30	390	.177	.197	.200	1.013	321	61	8	0	0	5.32	11	20

TAB. 4.3 - *Quelques paramètres descriptifs des rangées de maïs et de leur infestation par les pontes de pyrale.*

- Rg est le numéro de rangée de maïs observées.
- N est le nombre de plantes de la rangée
- p est la proportion de plantes infestées.
- μ, σ^2 et I_D sont respectivement la moyenne, la variance et le rapport variance-moyenne du nombre de pontes par plante.
- N_i est le nombre de plantes ayant i pontes.
- $\bar{l}$ est la longueur moyenne des séquences d'infestations, l_1 le nombre de séquences de longueur unité et l_{max} la longueur maximale des séquences observées.

4.3.2 Calcul de l'indice de dispersion

L'indice de dispersion I_D est tout simplement le rapport variance sur moyenne d'échantillon. On a donc :

$$I_D = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)\bar{x}} \quad \text{avec} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad x_i \text{ étant le nombre d'individus}$$

observés dans la i^e placette (ici, groupe de plantes contigues dans une rangée de maïs).

Cet indice donne une idée de la variabilité du nombre d'individus à l'échelle de la taille des placettes utilisées. Il permet aussi de vérifier l'hypothèse de distribution au hasard des individus observés à l'échelle de la taille des placettes. Le test qu'on applique alors est cependant conservatif quand le nombre moyen d'individus par placette est faible. Il est basé sur la comparaison de I_D avec 1 qui est la valeur théorique du rapport variance-moyenne sous l'hypothèse de distribution au hasard d'individus. Ainsi, six des rangées observées présentent une surdispersion significative.

4.3.3 Etude de la disposition des plantes infestées au sein des rangées de maïs

On s'intéresse ici aux nombres de plantes saines entre 2 plantes infestées (= séquence d'infestation) d'une même rangée de maïs. Sous l'hypothèse de répartition au hasard des pontes, on peut calculer les fréquences espérées des tailles des séquences d'infestations et les comparer aux fréquences observées à l'aide de notre échantillonnage. Un écart significatif entre ces deux types de fréquence nous instruit sur la tendance à l'agrégativité spatiale (ou à la régularité) des infestations en pontes. Il convient, en outre, de calculer l'indice de dispersion des pontes à l'échelle de la plante pour avoir une idée de la fluctuation du nombre de pontes d'une plante à une autre. Les tests de Monte-Carlo (voir par exemple Besag et Diggle, 1977) peuvent être très intéressants à pratiquer dans de telles situations à cause de leur simplicité : Vaillant et Badenhausser (1989) ont ainsi montré une agrégativité spatiale significative pour quelques unes des 30 rangées de maïs présentées ci-dessus.

4.3.4 Autocorrélation spatiale du nombre de pontes déposées sur des groupes de plantes voisines

Ici, on considère la variable aléatoire nombre de pontes déposées sur un groupe de plantes voisines d'une même rangée (= placette) afin de voir s'il n'y a pas une certaine similitude dans les observations effectuées sur les différentes pla-

cettes de notre échantillonnage, en fonction de la proximité entre ces placettes. En effet, si l'on trouve qu'il n'y a pas équiprobabilité des permutations spatiales des réalisations de pontes dans ces placettes, cela veut dire qu'en échantillonnant en plaçant d'une façon non aléatoire des placettes dans un champ, on pourrait sous-estimer ou sur-estimer le niveau d'infestation en pontes. Pour tester cette autocorrélation spatiale, on peut calculer l'indice de Moran I_M. Cet indice permet alors de tenir compte du type de voisinage qu'on désire étudier entre nos placettes. Il s'écrit :

$$I_M = \frac{n \sum_{i,j} z_i z_j w_{ij}}{(\sum_{i,j} w_{ij}) \sum_{i=1}^{n} z_i^2}$$

où $z_i = x_i - \bar{x}$ et où $w = (w_{ij})$ est la matrice des coefficients positifs définissant la nature de l'autocorrélation à étudier (par exemple, w_{ij} non nul si et seulement si $|i - j| = 1$ pour l'autocorrélation d'ordre 1).

Les tests de Monte-Carlo utilisés avec I_M ont montré la présence parfois d'autocorrélation positive d'ordre un (Vaillant et Badenhausser, 1989).

4.3.5 Etude d'une zone préférentielle de ponte au niveau de la plante de maïs.

Durant l'été 1980, des observations ont été effectuées par le laboratoire de Zoologie de Versailles sur 1512 plantes de maïs de la variété LIZA et 2016 plantes de la variété INRA 252. La parcelle expérimentale située dans la Beauce a été observée 2 fois par semaine durant toute la période de ponte de la pyrale, soit au total 15 dates d'observation. Entre autres caractéristiques, le numéro de la feuille porteuse de la ponte fut notée. Le tableau TAB. 4.4 nous montre la distribution des pontes sur les différents niveaux foliaires (les feuilles sont numérotées par ordre croissant à partir de la feuille la plus proche du sol). F_i est le numéro de la feuille porteuse de la ponte. Les 15 dates d'observation couvrent toute la période de ponte de la pyrale, 2 dates consécutives correspondant à une semaine.

On constate que quelles que soient la date de ponte et la variété considérées, les feuilles 4 à 9 reçoivent la grande majorité des pontes. Sur toute la période de ponte, le pourcentage de pontes situées sur cette strate foliaire est 85% pour LIZA et 89% pour INRA 252. En ajoutant la feuille 10 à cette strate préférentielle de ponte, ces pourcentages deviennent 91% pour LIZA et 93% pour INRA 252. Ces constatations furent mises à profit pour l'élaboration d'une procédure de détection des pontes au niveau de la plante et dans le

cadre de la lutte biologique contre la pyrale du maïs à l'aide du Trichogramme (Hawlitzky *et al.*, 1994). Ceci fera, en partie, l'objet du paragraphe 4.3.6 de ce chapitre. Notons déjà que les pontes sont déposées sur la face inférieure des feuilles, ce qui complique un peu plus leur détection puisqu'il faut pour cela retourner les feuilles. Il arrive aussi qu'on puisse trouver une ponte sur un entre-noeud de la tige mais ceci reste extrèmement rare.

Date	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	Tot.
1	0	0	0	1	1	4	0	1	0	0	0	0	0	7
2	0	0	0	0	1	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	1	0	0	0	1	0	0	2
4	0	0	0	1	0	3	1	3	3	1	2	0	0	14
5	0	0	1	3	3	7	5	7	1	2	0	0	0	29
6	0	0	7	3	3	11	9	9	4	1	2	0	0	49
7	1	5	9	23	31	41	41	28	14	4	2	1	0	200
8	0	0	7	9	24	28	27	24	18	8	1	2	0	148
9	0	3	15	23	57	70	62	67	50	20	13	7	1	388
10	0	1	5	8	38	34	39	36	19	20	3	4	1	208
11	0	0	3	27	44	55	55	68	47	28	12	3	0	342
12	0	0	4	6	19	10	24	17	10	5	7	2	0	104
13	0	0	0	1	0	4	5	7	2	4	1	2	1	27
14	0	0	0	0	0	4	7	2	2	1	2	2	0	20
15	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Tot.	1	9	51	105	221	271	277	269	170	94	46	23	3	1540

Variété LIZA (1512 plantes observées)

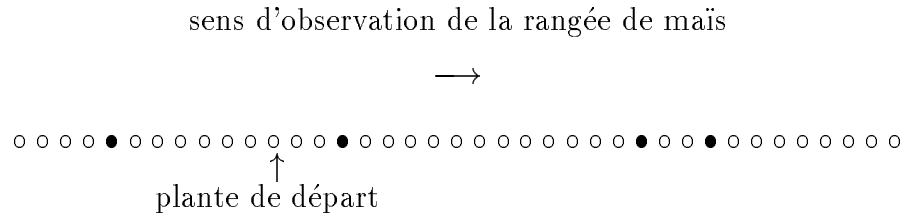
Date	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	Tot.
1	0	0	0	0	1	0	0	1	0	0	0	0	0	2
2	0	0	1	0	0	0	0	0	0	0	0	0	0	1
3	0	0	1	1	3	1	3	1	1	0	0	0	0	11
4	0	0	2	1	1	3	1	1	1	0	0	0	0	10
5	0	0	4	1	0	3	7	6	1	0	0	0	0	22
6	1	0	4	7	9	9	1	5	2	0	0	0	0	38
7	0	4	4	16	25	22	17	10	7	3	0	0	0	108
8	0	1	6	16	24	17	23	16	11	8	1	1	0	124
9	0	3	13	43	64	80	66	54	39	7	4	2	1	376
10	0	4	6	24	37	31	38	26	13	9	3	0	0	191
11	0	1	8	15	28	41	33	35	23	15	1	0	0	200
12	0	0	0	3	8	2	10	7	3	2	1	0	0	36
13	0	0	0	0	1	3	3	4	0	1	0	0	0	12
14	0	0	0	0	0	1	0	7	3	1	0	0	0	12
15	0	0	0	0	0	0	0	0	1	1	0	0	0	2
Tot.	1	13	49	127	201	213	202	173	105	47	10	3	1	1145

Variété INRA 252 (2016 plantes observées)

TAB. 4.4 - *Effectifs de pontes de pyrale suivant la date d'observation et le numéro de strate foliaire.*

4.3.6 Etude de stratégies d'échantillonnage pour l'estimation des niveaux d'infestations en pyrale

Il existe dans la théorie des sondages des résultats classiques sur l'admissibilité ou l'optimalité de certains estimateurs de la moyenne d'une caractéristique donnée, aussi bien dans le cadre de population fixe que dans celui de superpopulation (Cassel *et al.*, 1977). Il en va de même pour les stratégies d'échantillonnage qui sont, en termes de théorie des sondages, des couples définis par la donnée d'un plan de sondage (loi de probabilité sur l'univers des échantillons possibles) et d'un estimateur. Ces résultats sont basés sur le risque défini à partir de l'erreur quadratique moyenne d'estimation. Or, l'une des préoccupations essentielles de l'échantillonneur en écologie est la facilité de mise en oeuvre du plan de sondage et le coût qui lui est associé. L'échantillonnage informatif (CHAP. 5), et en particulier le séquentiel, peut être une solution intéressante (Denechere *et al.*, 1982) mais peut aussi s'avérer impraticable dans certains cas (Badenhausser et Vaillant, 1987). Pour ce qui concerne le problème d'estimation de la moyenne des pontes de pyrale dans une région donnée, des procédures d'observation des plantes de maïs permettant de réduire le coût moyen d'échantillonnage ont été étudiées (Vaillant, 1985; Vaillant et Derridj, 1989). En effet, sous certaines hypothèses sur le type de répartition des pontes dans les rangées de maïs, la loi des tailles de séquence d'infestation est connue et la vraisemblance de l'échantillon observé peut être explicitée. Ceci permet alors de développer les méthodes statistiques liées au maximum de vraisemblance et, en particulier, d'obtenir des estimateurs de la densité de population (ainsi que certaines de leurs propriétés) sous les différentes hypothèses concernées. Ces stratégies basées sur les séquences d'infestation ou, en d'autres termes sur l'observation partielle des séquences présence-absence d'infestation, ne nécessitent pas le comptage des pontes sur les plantes observées mais seulement d'examiner l'état d'infestation de la plante (FIG. 4.2).



○ correspond à une plante non infestée.
● correspond à une plante infestée.

FIG. 4.2 - *Illustration d'une procédure d'échantillonnage basée sur les séquences présence-absence de ponte.*

Description de la méthode : On choisit d'abord l'emplacement des n *plantes dites de départ*, de préférence de façon systématique afin de couvrir tout le champ et d'éviter des redondances d'information. Le nombre n est choisi à partir de considérations de coût et de précision. On part, ensuite d'une des plantes de départ comme illustré ci-dessus et on observe les plantes suivantes dans un sens déterminé de la rangée. Une fois une ponte détectée, on va à la prochaine plante de départ et on opère comme précédemment, ainsi de suite jusqu'à la $nième$ plante de départ. On a alors n observations (en occurrence, les n nombres de plantes à observer avant détection d'une ponte) qui permettent d'estimer la moyenne de pontes par plante par un procédé statistique approprié. Pour limiter le coût d'échantillonnage, on peut limiter l'observation d'une séquence d'infestation à n_0 plantes (d'après les valeurs l_{max} de TAB. 4.3, $n_0 = 20$ est une valeur correcte). Cela signifie que si, en partant d'une plante de départ, on n'a, à la $n_0 ième$ plante, encore détecté aucune ponte, on va à une autre plante de départ (ou on s'arrête si l'on était déjà à la $nième$ plante de départ). Il est à noter qu'ainsi, on limite aussi, mais de façon raisonnable, la précision d'estimation.

Remarque simple mais cruciale : une plante de maïs infestée n'est pas distinguable d'une plante non infestée immédiatement. Les pontes de pyrale sont minuscules et déposées sur la face inférieure des feuilles. Il faut donc ici passer du temps à examiner les plantes, ce qui n'est pas le cas pour d'autres types de répartition d'individus.

L'une de ces procédures, paraissant la plus intéressante d'un point de vue pratique, fut comparée à l'échantillonnage aléatoire en grappes. Il s'avère qu'en cas de faibles infestations, la procédure basée sur les séquences d'infestations est la meilleure : coût moyen d'observation d'une plante diminuée et bonne précision d'estimation même en cas d'observations tronquées des séquences d'infestation. Une présentation plus conséquente de ces méthodes peut être examinée dans l'article de Vaillant et Derridj (1989).

Protocole d'observation au niveau de la plante

Vu les résultats fournis au paragraphe 4.3.5 de ce chapitre et les spécificités de la plante de maïs, tous stades phénologiques confondus, voici ce que nous préconisons pour une détection plus rapide de la première ponte dans les stratégies présentées ci-dessus : examiner d'abord les feuilles situées dans la strate médiane de la plante (feuilles 4 à 10 pour un stade phénologique avancé), examiner ensuite les feuilles les plus proches du sol, et enfin terminer, ci besoin est, par les feuilles les plus hautes. Dans la pratique, on examinera donc la feuille 10 puis la feuille 9, *etc* jusqu'à la feuille 1, ensuite la feuille 11, la feuille 12, *etc* en s'arrêtant bien-sûr dès la découverte d'une ponte.

4.4 Bibliographie

BADENHAUSSER, I. ET VAILLANT, J. (1987): Echantillonnage séquentiel de populations distribuées dans l'espace. C.R. Acad. Agric. Fr., 73, 83-92.

BAILEY, N. T. J. (1964) : Elements of stochastic processes. Ed. Wiley Newyork 249pp.

BESAG, J. E. (1974a): Spatial interaction and the statistical analysis of lattice systems. JRSS 36, 192-236.

BESAG, J. E. (1974b) : On spatial-temporal model and Markov fields. Trans. European Meeting of Statisticians, Prague.

BESAG, J. E. AND DIGGLE, P. J. (1977): Simple Monte-Carlo tests for spatial patterns. App. Statist. 26, 3, 327-333.

BLISS, C. I. AND FISHER, R. A. (1953) : Fitting the Negative Binomial distribution to biological data. Biometrics, June 1953 176-200.

BLISS, C. I. AND OWENS, A. R. G. (1958) : Negative Binomial distributions with a common k. Biometrika, 45, 37-58.

BYTH, K. (1982): On robust distance-based intensity estimators. Biometrics 38, 127-135.

CASSEL, C. M., SÄRNDAL, C. E. AND WRETMAN, J. H. (1977): Foundation of inference in survey sampling. Wiley. 192 pages.

CLIFF, A. D. AND ORD, K. (1981). Spatial processes, models and applications. Pion Limited London. 266 pages.

COCHRAN, W.G. (1977): Sampling techniques. Wiley. 3*ième* édition. 428 pages.

DEBOUZIE, D., DENIS, J-B., ET ROSPARS, J-P. (1987): Echantillonnage et répartition spatiale. C.R Acad. Agric. Fr., 73, 73-82.

DENECHERE, M., DERRIDJ, S. ET DUBY, C. (1982): Etude d'une méthode d'échantillonnage séquentiel appliquée à l'estimation du nombre de pontes de la pyrale du maïs. Agronomie, 2, 341-346.

DIGGLE, P. J. (1975): Robust density estimation using distance methods. Biometrika 62, 1, 39-48.

DIGGLE, P. J. (1977): A note on Robust density estimation for spatial point patterns. Biometrika 64, 91-95.

DIGGLE, P. J. AND MILNE, R. K. (1983) : Negative Binomial Quadrat Counts and Point Processes. Scand. J. Statist., 10, 257-267.

GREIG-SMITH, P. (1952): The use of random and contiguous quadrats in the study of the structure of plant communities. Ann. Bot. 16, 293-316.

GUYON, X. (1985): Estimation d'un champ par pseudo-vraisemblance conditionnelle: Etude asymptotique et application au cas markovien. Proc. 6th. Franco-belgian Meet. FUSL, 15-62.

HAWLITZKY, N., DORVILLE, F. M. ET VAILLANT, J. (1994): Statistical Study of *Trichogramma Brassicae* efficiency in relation with characteristics of the European Corn Borer egg mass. Res. Popul. Ecol. 36, sous presse.

JACOBSEN, M. (1982): Statistical Analysis of counting Processes. Lecture Notes In Statistics, Vol 12, 226pp.

KERSHAW, K. A. (1957): The use of cover and frequency in the detection of pattern in plant communities. Ecology, 38, 2, 291-299.

KONIJN, H. S. (1973): Statistical theory of sample survey design and analysis. North-Holland Amsterdam. 429 pages.

MEAD, R. (1974): A test for spatial pattern at several scales using data from a grid of contiguous quadrats. Biometrics 30, 295-307.

OGATA, Y. (1981): On Lewis' Simulation Method for Point Processes. IEEE Trans. on inform. Theor., 27, 1, January 1981, 23-31.

PIELOU, E. C. (1977): Mathematical ecology. Wiley. 384 pages.

RENSHAW, E. (1972): Birth, death and migration processes, Biometrika, 59, 49-60.

ROGERS, A. (1973): Statistical analysis of spatial dispersion. Pion Limited. 164 pages.

SEBER, G. A. F. (1973): The estimation of animal abundance and related parameters. Griffin London. 506 pages.

SOUTHWOOD, T. R. E. (1978): Ecological methods. Chapman and Hall, London. 524 pages.

TAYLOR, L. R. (1979): The Negative Binomial as a dynamic ecological model for aggregation. Journ. of anim. ecol., 48, 289-304.

VAILLANT, J. (1985): Etudes statistiques des répartitions spatiales et temporelles des pontes de pyrale dans le Bassin Parisien. Problèmes d'échantillonnage. Thèse de 3ieme cycle, Univ. Paul Sabatier de Toulouse. 284 pages.

VAILLANT, J. (1989): Spatial-temporal models for counting processes. Cahiers du CERO, Operat. Res. Statist. and Appl. Math., 31, 277-296.

VAILLANT, J. ET BADENHAUSSER, I. (1989). Etude de répartition d'individus sur transect ou grille régulière par tests de Monte-Carlo. Biom. Praxim. 29, 153-172.

VAILLANT, J. ET DERRIDJ, S. (1989). Estimation of European Corn Borer egg masses density by sampling of runs. Res. Popul. Ecol. 31, 2, 1-16.

Chapitre 5

Echantillonnage informatif et échantillonnage à taille aléatoire

Plan :

5.1 Introduction

Au chapitre 2, nous avons présenté des procédures d'échantillonnage qui sont à taille fixe ou sont conçues comme tel. En fait, le point commun entre ces procédures sont qu'elles sont non informatives c'est-à-dire qu'en cours d'exécution du plan, on ne retire pas de l'information sur la population échantillonnée, conduisant à modifier le déroulement de ce plan. Il n'y a donc pas d'interférence entre l'exécution du plan et l'observation des variables d'intérêt. Tout se passe, en fait, comme si les observations étaient recueillies de façon instantanée alors qu'en pratique, il y a en général un ordre chronologique. On comprend cependant qu'il puisse exister des situations où le praticien s'accommode mal d'une telle abnégation : S'il avait projeté de choisir 500 plantes au hasard dans un champ et d'y dénombrer des individus d'une espèce d'insecte afin d'estimer le niveau moyen d'infestation dans ce champ, si au bout de 300 plantes observées, il n'a dénombré aucun insecte, il aura envie d'arrêter ses observations et de déclarer qu'il n'y a pas d'insecte dans le champ. C'est un peu l'idée des procédures informatives qui intègrent d'une certaine façon l'information concernant la population étudiée, pendant le déroulement du plan d'échantillonnage, afin de modifier en cours de route, et de façon appropriée,

le plan en question. Les deux procédures informatives les plus connues sont l'échantillonnage séquentiel (Wald, 1947; Millier, 1967; Kuno, 1972) et certains types de double échantillonnage (Cochran, 1977; Jessen, 1978; Grosbras, 1987). Nous les verrons plus en détail au § 5.3. Auparavant, on s'intéressera aux méthodes d'échantillonnage à taille non fixe.

5.2 Echantillonnage à taille aléatoire

5.2.1 Aléa dû à la structure de la population statistique

Nous avons dit au paragraphe précédent que certains plans d'échantillonnage présentés au chapitre 2 n'étaient pas forcément à taille fixe. Il s'agit, principalement, de l'échantillonnage en grappe quand ces grappes sont de tailles inégales. Dans ce type d'échantillonnage, seul le nombre de grappes est fixé au préalabre. Un exemple simple est le suivant : on a une population constituée de 3 grappes de taille 3, 10, 18. En tirant au hasard deux grappes (sans remise), la taille de notre échantillon peut prendre les valeurs 13, 21 ou 28. Deux exemples plus concrets sont fournis par ceux du § 2.4.2.3. L'exemple 1 concerne un ensemble de clochers abritant des jeunes chouettes et la variable d'intérêt est le poids de ces animaux. Un nombre fixe de clochers est tiré au hasard et tous les jeunes sont pesés. Les unités primaires sont donc les clochers et les unités secondaires les jeunes chouettes. Puisque les effectifs de ces derniers dans les clochers ne peuvent être connus d'avance, le nombre d'animaux échantillonnés sera variable d'un échantillon à un autre même si le nombre de clochers visités reste constant. On a donc un échantillonnage à taille aléatoire puisque le nombre total de jeunes chouettes ainsi observées l'est également. Le second exemple est l'étude des bovins de moins d'un an d'une région d'élevage : un nombre fixe de cheptels est examiné dans cette région, et une caractéristique donnée est analysée pour chaque jeune bovin. Là encore, le nombre total de bovins (qui sont nos unités statistiques de base) est aléatoire, et par définition, la taille de notre échantillon.
On voit donc que ce type d'aléa dans la taille d'échantillon peut arriver pour tout sondage à plusieurs niveaux (§ 2.4.3) dès lors que les tailles des unités à certains niveaux sont inégales et qu'un recensement est effectué à l'un des niveaux concernés. Par contre, si à tous les niveaux, les tailles d'unités statistiques sont égales et de valeurs connues, et le taux d'échantillonnage fixe, alors la taille de l'échantillon global sera constant et déterminé simplement par les taux d'échantillonnage à chaque niveau. Voici un exemple trivial : en effectuant un échantillonnage en grappes par prélèvement de n grappes toutes de taille g, alors tout échantillon ainsi réalisé sera de taille ng.

5.2.2 Aléa dû à la nature du plan

Les plans discutés au § 5.2.1 sont non informatifs et la taille de l'échantillon finalement observé n'est aléatoire qu'à cause du fait qu'on échantillonne à plusieurs niveaux dans des sous-populations de taille variée. Les plans informatifs, quant à eux, peuvent être aussi bien à taille fixe qu'à taille aléatoire. Dans ce dernier cas, le fait que le nombre d'unités statistiques à observer ne soit pas connu d'avance est souvent dû aux interactions entre observations effectuées et renseignements relatifs au phénomène étudié, obtenus de façon progressive. Ainsi, si dans le dénombrement d'oeufs d'insecte dans une parcelle de maïs, on veut atteindre une certaine précision d'estimation pour le nombre moyen d'oeufs par plante, on peut appliquer la procédure suivante : à chaque plante observée (choisie au hasard), on calcule une précision estimée et on n'arrête les observations que lorsque cette précision estimée est supérieure à la précision désirée au départ (Denechere *et al.*, 1982; Badenhausser et Vaillant, 1987). Telle est l'idée de l'échantillonnage séquentiel à précision choisie. L'application de tels plans conduit, en particulier, à un coût d'échantillonnage non connu à l'avance. Leur mérite est cependant de permettre la maîtrise de la précision des estimations effectuées, ou des risques d'erreurs en cas d'échantillonnage à but décisionnel (Iwao, 1975; Kuno, 1976; Kuno, 1991; Badenhausser, 1989). En outre, dans certaines situations, le coût moyen d'échantillonnage est diminué (Fowler, 1978; Denechere et al., 1982), ce qui peut rendre cette procédure attrayante malgré une mise en oeuvre parfois lourde. Nous nous étendrons davantage sur ce type de plan au paragraphe § 5.3.

5.3 Procédures informatives

Nous tenons tout d'abord à souligner une fois de plus que les procédures d'échantillonnage informatives ne sont ni synonymes d'échantillonnage à taille aléatoire (on vient de le voir au § 5.2.1), ni d'échantillonnage séquentiel, comme certains abus de langage pourrait le laisser supposer. Ainsi le double échantillonnage (Cochran, 1977; Gourieroux, 1981) est généralement informatif et à taille fixe, et c'est également le cas des procédures en plusieurs phases (discutée au § 5.3.2). Le double échantillonnage, dans sa version la plus classique, consiste à acquérir dans un premier temps de l'information sur une variable auxiliaire x à l'aide d'un échantillon s_1 puis d'utiliser cette information pour construire un "bon" échantillon s_2 pour étudier la ou les variable(s) d'intérêt. Quand la taille de s_1 et et celle de s_2 sont déterminées d'avance, on a bien-sûr un échantillonnage à taille fixe. Il arrive cependant que la taille de s_2 soit établie à partir de l'observation de s_1. C'est tout particulièrement le cas quand s_1 n'est qu'un pré-échantillon servant à estimer la variabilité au sein de la population afin de déterminer le taux optimal d'échantillonnage (par exemple, l'allocation de l'effort d'échantillonnage pour un plan stratifié vu au § 2.4.1.2). Dans la suite de ce paragraphe, nous allons d'abord parler de l'échantillonnage

séquentiel et des caractéristiques de son application en écologie des populations. Ensuite, nous discuterons des plans en plusieurs phases.

5.3.1 Méthodes séquentielles

Le principe des méthodes séquentielles est de tenir compte de l'information accumulée en cours de prélèvement de l'échantillon pour maîtriser des critères statistiques d'intérêt, quant aux objectifs poursuivis par l'échantillonneur. Les deux principaux types de préoccupation auxquels ces méthodes peuvent répondre sont :

1. Test d'une hypothèse avec maîtrise des risques d'erreur (échantillonnage à but décisionnel, voir par exemple Millier (1967), Bechinski *et al.* (1983)).

2. Estimation de paramètres à précision fixée (voir, par exemple, Grambsch (1983)).

Un plan séquentiel est caractérisé, d'une part, par un choix au hasard et successif des unités à observer, et d'autre part, par une règle d'arrêt permettant de décider de l'arrêt des observations ou de leur poursuite (FIG. 5.1).
Pour l'étude de populations écologiques distribuées spatialement (estimation ou classement de niveau de densité), il nécessite généralement un modèle de description de la répartition spatiale des individus (Badenhausser et Vaillant, 1987; Badenhausser, 1989). Les principaux modèles utilisés dans la littérature sont soit probabilistes (loi de Poisson, binomiale Négative, ...), soit statistiques (loi empirique de Taylor, relation d'Iwao, ...). On pourra se rapporter à Taylor (1984) pour l'interprétation de différents modèles et aux auteurs suivants pour des exemples d'utilisation des méthodes séquentielles en écologie : Oger (1982), Harcourt (1983), Kuno (1986), Badenhausser et Vaillant (1987), Badenhausser (1989). L'ouvrage de base pour les méthodes séquentielles est de Wald (1947), vulgarisé par Millier (1967).

Exemple d'échantillonnage séquentiel à précision fixée

Supposons qu'on ait à analyser une répartition d'individus écologiques dans une parcelle expérimentale et qu'on veuille plus précisément estimer la moyenne du nombre d'individus X par unité statistique, ceci en cessant les observations dès qu'une précision P_0 fixée à l'avance est atteinte. La procédure d'arrêt peut être basée sur une relation f entre l'espérance m et la variance σ^2 de X :

$$\sigma^2 = f(m).$$

La précision relative pour n observations effectuées au hasard parmi un nombre N (supposé grand) d'unités échantillonnables (population statistique) est définie de la façon suivante :

$$P_0(n) = \frac{\sigma}{m\sqrt{n}} = \sqrt{\frac{f(m)}{nm^2}}$$

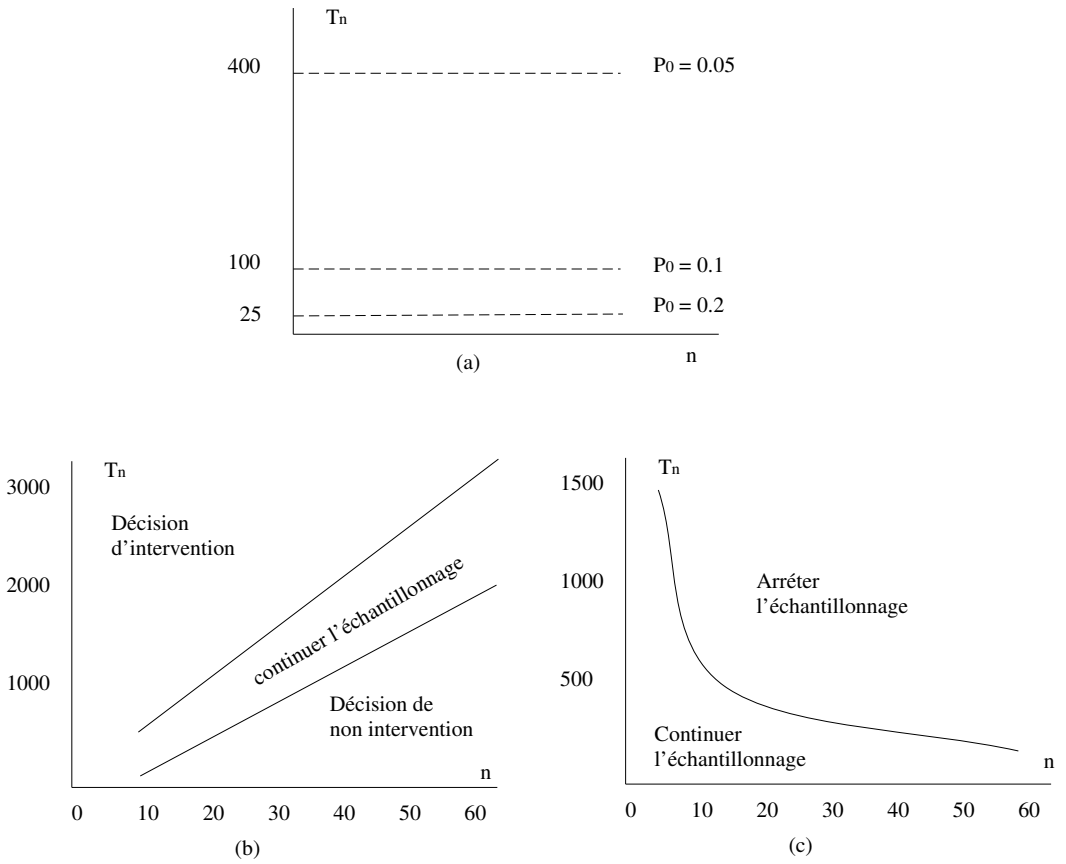


FIG. 5.1 - *Quelques courbes d'arrêt d'échantillonnage séquentiel pour l'étude de population.*
(a) Estimation de densité à précision relative fixée P_0, pour une loi du nombre d'individus par unité statistique dont l'espérance m et la variance σ^2 vérifient la relation $\sigma^2 = m$.
(b) Objectif décisionnel pour un seuil d'intervention fixé à 30 individus par unité, un risque de 5%, et une population vérifiant la loi de Taylor $s^2 = 4{,}05\bar{x}^{1{,}47}$, ($\bar{x}$ et s^2 étant respectivement la moyenne et la variance d'échantillonnage).
(c) Estimation de densité pour une précision relative P_0=0,20 et la relation moyenne-variance de Taylor $s^2 = 4{,}05\bar{x}^{1{,}47}$. (**(b)** et **(c)** tirés de Badenhausser (1989)).
n=nombre d'unités statistiques observées; T_n=nombre cumulé d'individus.

Dans le cas où N est faible, on applique à la variance d'échantillonnage σ^2/n le facteur correctif de population finie $1 - n/N$ d'où :

$$P_0(n) = \sqrt{\frac{(1 - n/N)f(m)}{nm^2}}$$

Au *nième* prélèvement, m peut être estimé par T_n/n où T_n est le nombre cumulé d'individus observés dans les n unités échantillonnées. La précision

relative $P_0(n)$ est alors elle même estimée par :

$$P_0^*(n) = \sqrt{\frac{nf(T_n/n)}{T_n^2}}.$$

L'échantillonnage s'arrête lorsque la précision désirée P_0 est atteinte par la précision estimée $P_0^*(n)$. Connaissant la fonction f, on peut définir la courbe d'arrêt de l'échantillonnage pour atteindre P_0 à partir des valeurs de T_n (FIG. 5.1). Kuno (1972) souligne cependant l'existence d'un biais dans l'estimation de la moyenne par T_n/n due au fait que n est aléatoire, contrairement au plan à taille fixe.
Un exemple de fonction f est donné par la relation d'Iwao (1968) :

$$f(m) = (\alpha + 1)m + (\beta - 1)m^2,$$

où α et β sont des paramètres caractérisant le mode de répartition de la population écologique.
On a alors une règle d'arrêt basée sur :

$$P_0^*(n) = \sqrt{\frac{(\alpha + 1)}{T_n} + \frac{\beta - 1}{n}} \quad .$$

Dans la pratique, la fonction f est supposée connue, mais ce n'est parfois pas le cas. Par exemple, Kuno (1972) discute de la robustesse de cette méthode d'estimation par rapport à l'hypothèse de validité du modèle défini par la fonction f de la relation d'Iwao.

5.3.2 Echantillonnage en plusieurs phases

Nous allons maintenant parler d'échantillonnage en plusieurs phases, à ne pas confondre avec échantillonnage à plusieurs niveaux (ou degrés) vu au § 2.4.3. Certains auteurs l'appellent aussi échantillonnage à différentes occasions (Cochran, 1977, § 11 et § 12; Scherrer, 1983, §8). Ce type d'échantillonnage regroupe des plans pour lesquels l'acquisition de l'information se fait en un nombre fixé d'étapes échelonnées dans le temps mais pas forcément de façon informative. En effet, acquisition par étape de l'information ne signifie pas que la procédure est informative car cette information n'est pas forcément utilisée pour modifier le déroulement du plan : Elle peut être simplement utilisée pour analyser, en fin d'exécution de ce plan, les interactions temps-variables d'intérêt. Un exemple immédiat est l'échantillonnage permanent pour lequel l'échantillon tiré pour la première étape, reste inchangé pour les étapes suivantes. Cette dernière procédure est souvent utilisée pour les relevés météorologiques en plusieurs sites.
Par contre, il peut s'avérer fort intéressant qu'à chacune des étapes, les renseignements obtenus au préalable puissent déterminer le choix des échantillons

des phases ultérieures. Par exemple, avec un double échantillonnage, on peut effectuer une stratification a posteriori du milieu étudié à l'aide du premier échantillon tiré. Cela permet alors, selon la situation, d'améliorer les qualités des estimations de paramètres de ce milieu.
En écologie, on peut distinguer généralement trois types de plans d'échantillonnage en plusieurs phases :

- l'échantillonnage permanent,
- l'échantillonnage partiellement renouvelé,
- l'échantillonnage renouvelé.

Ces plans sont surtout intéressants quand on désire analyser l'évolution, les modifications des caractéristiques du phénomène étudié, au cours du temps, ou entre chaque phase du plan.
Pour ce qui concerne l'échantillonnage partiellement renouvelé, le calcul de certains estimateurs peut s'avérer difficile à cause de contraintes concernant les unités observées à plusieurs reprises par rapport à celles qui ne seraient observées que durant une seule phase (Cochran, 1977).

Exemple d'échantillon permanent

Sur une parcelle de maïs de 40m sur 40m où des lâchers de trichogramme (*Trichogramma maidis*) ont été effectués, on désire estimer le nombre moyen d'oeufs par ooplaque de pyrale, parasités par cet oophage, et ceci tout au long de la période de ponte de la pyrale. Le trichogramme est utilisé dans la lutte biologique contre la pyrale du maïs et il est nécessaire de mieux apprécier son aptitude à attaquer les pontes de pyrale sous diverses conditions. L'étude de l'évolution dans le temps et dans l'espace du parasitisme par cet oophage est donc cruciale pour les chercheurs concernés par ce moyen de lutte. Un premier échantillon fut tiré en choisissant de façon quasi-systématique 504 plantes dans la parcelle. Puis, les mêmes plantes furent examinées à raison de deux fois par semaine durant toute la période de ponte de la pyrale. Un tel suivi des 504 plantes permit d'observer l'évolution temporelle du parasitisme sur les pontes présentes. D'autre part, la disposition quasi-systématique des plantes figurant dans l'échantillon faisait que la parcelle était suffisamment quadrillée pour fournir une idée appréciable de la progression spatiale du parasitisme (Hawlitzky *et al.*, 1994).

5.4 Bibliographie

BADENHAUSSER, I. (1989). Echantillonnage séquentiel et répartition spatiale des insectes : fondements méthodologiques et application au cas du puceron du pois. Acta Œcologica, Œcol. Applic., 10, 81-97.

BADENHAUSSER, I. ET VAILLANT, J. (1987). Echantillonnage séquentiel de populations distribuées dans l'espace. C. R. Acad. Agri. Fr., 73, 83-92.

BECHINSKI, E. J., BUNTIN, G. D., PEDIGO, L. P. AND THORVILSON, H. G. (1983) Sequential count and decision plans for sampling Green Cloverworm (Lepidoptera: Noctuidae) larvae in Soybean. J. Econom. Entom., 76, 806-812.

COCHRAN, W. G. (1977) Double sampling. In "Sampling techniques", chap. 12. Wiley. 3*ième* édition.

DENECHERE, M., DERRIDJ, S. ET DUBY, C. (1982). Etude d'une méthode d'échantillonnage séquentiel appliquée à l'estimation du nombre de pontes de la pyrale du maïs. Agronomie, 2, 341-346.

FOWLER, G. W. (1978) Errors in sampling plans based on Wald's sequential probability ratio test. US Dep. Agric. For. Ser. Gen. Tech. Rep., 46, 1-13.

GROSBRAS, J. (1987) Sondages en plusieurs phases. In "Méthodes statistiques des sondages", chap. 9. Ed. Economica, ESA.

GRAMBSCH, P. (1983) Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. Ann. of Statistics, 11, 68-77.

GOURIEROUX, C. (1981) Tirages successifs. In Théorie des sondages, chap. 6. Ed. Economica, Paris.

HARCOURT, D. G. (1983) A sequential decision plan for management of the Alfalfa Blotch leafminer, Agromyza Frontella (Diptera: Agromizidae). Can. Ent., 115, 1513-1518.

HAWLITZKY, N., DORVILLE, F. M. ET VAILLANT, J. (1994): Statistical Study of *Trichogramma Brassicae* efficiency in relation with characteristics of the European Corn Borer egg mass. Res. Popul. Ecol. 36, sous presse.

IWAO, S. (1968). A new regression method for analysing the aggregation pattern of animal populations. Res. Popul. Ecol., 10, 1-20.

IWAO , S. (1975). A new method of sequential sampling to classify populations relative to a critical density. Res. Popul. Ecol., 16, 281-288.

JESSEN, R. J. (1978) Double sampling (multiphase). In "Statistical Survey techniques", Chap. 10. Wiley.

KUNO, E. (1972). Some notes on population estimation by sequential sampling. Res. Popul. Ecol., 14, 58-73.

KUNO, E. (1976). Multi-stage sampling for population estimation. Res. Popul. Ecol., 18, 39-56.

KUNO, E. (1986). Evaluation of statistical precision and design of efficient sampling for the population estimation based on frequency of occurrence. Res. Popul. Ecol., 28, 305-319.

KUNO, E. (1991). Verifying zero-infestation in pest control : a simple sequential test based on the succession of zero-samples. Res. Popul. Ecol., 33, 29-32.

MILLIER, C. (1967). Une méthode statistique : l'analyse progressive. Ann. Sci. forest., 24, 327-343.

OGER, R. (1982). Une procédure séquentielle pour le contrôle des maladies et des populations d'organismes nuisibles des plantes cultivées. Biométrie-Prax., 22, 149-162.

SCHERRER, B. (1983). Techniques de sondage en écologie. Collection d'écologie, 17, Masson, PUL, 63-162.

VAILLANT, J. AND DERRIDJ, S. (1989) Estimation of European Corn Borer egg masses density by sampling of runs. Res. Popul. Ecol., 31, 289-304.

WALD, A. (1947). Sequential analysis. Wiley, New-York, 212 pages.

Index des auteurs

M

O

P

Q

R

S

T

V

W

Z

Index des sujets

R

S

T

U

V