COURS/ Master 1 : Protection des écosystèmes

Module : Méthodes d'échantillonnage des peuplements

Suite les cours du mois d'avril

Concepts fondamentaux de l'échantillonnage

L'échantillonnage consiste essentiellement à tirer des informations d'une fraction d'un grand groupe ou d'une population, de façon à en tirer des conclusions au sujet de l'ensemble de la population. Son objet est donc de fournir un échantillon qui représentera la population et reproduira aussi fidèlement que possible les principales caractéristiques de la population étudiée.

Les principaux avantages de la technique d'échantillonnage par rapport à une énumération complète sont le moindre coût, la rapidité, la portée et la précision accrue. Tous ceux qui soutiennent que le seul moyen d'obtenir des informations exactes sur une population est de faire un recensement exhaustif oublient que les sources d'erreurs sont nombreuses dans un dénombrement complet et qu'un recensement à cent pour cent peut non seulement être faussé par un grand nombre d'erreurs, mais être pratiquement irréalisable. En effet, avec un échantillon on peut obtenir des résultats plus exacts car il est plus facile de contrôler les sources d'erreurs liées à la fiabilité et à la formation des agents de terrain, à la clarté des instructions, aux mesures et à l'enregistrement, au mauvais entretien des instruments de mesure, à l'identification des unités d'échantillonnage, au travail des enquêteurs et au traitement et à l'analyse des données. Plus l'échantillon est petit, plus la supervision est efficace. De plus, le degré de précision des estimations tirées de certains types d'échantillons, peut être estimé à partir de l'échantillon même. En fin de compte on obtient souvent avec une enquête par sondage une réponse plus exacte qu'avec un recensement complet, le tout en peu de temps, avec moins de personnel, moins de travail et moins d'argent.

La méthode d'échantillonnage la plus simple consiste à sélectionner un certain nombre d'unités d'échantillonnage considérées comme "représentatives" de l'ensemble de la population. Par exemple, pour estimer le volume global d'un peuplement forestier, l'enquêteur peut choisir un petit nombre d'arbres qui lui paraissent de dimensions moyennes et typiques de la zone considérée, et mesurer leur volume. Les méthodes simples, telles que

marcher dans la forêt, s'arrêter au hasard et lancer une pierre les yeux fermés, ou tout autre démarche excluant en apparence toute possibilité de choix délibéré des unités d'échantillonnage, sont très attirantes à cause de leur simplicité, mais elles ont évidemment des chances d'être faussées par le jugement de l'enquêteur, de sorte que les résultats seront biaisés et non fiables. Même si l'objectivité de l'enquêteur ne fait pas le moindre doute, d'importantes erreurs de jugement, conscientes ou inconscientes, peuvent se produire, et elles seront rarement identifiées. Or ces erreurs peuvent être bien supérieures à l'avantage de l'exactitude accrue qui est censée dériver de la sélection délibérée ou intentionnelle des unités d'échantillonnage. Sans compter qu'un échantillonnage subjectif ne permet pas d'évaluer la précision des estimations calculées à partir des échantillons. Un échantillonnage subjectif est statistiquement irrationnel et en tant que tel, il est à éviter.

Si l'échantillonnage est fait de façon à ce que chaque unité de la population ait quelque chance d'être incluse dans l'échantillon et si la probabilité de sélection de chaque unité est connue, on parle de méthode d'échantillonnage probabiliste. L'une de ces techniques est la sélection aléatoire, à ne pas confondre avec la sélection au hasard, qui implique un processus de sélection rigoureux de type tirage au sort. Dans ce manuel, le terme échantillonnage se réfère, sauf indication contraire, à une forme quelconque d'échantillonnage probabiliste. La probabilité qu'une unité d'échantillonnage quelconque soit incluse dans l'échantillon dépend de la procédure adoptée. Il faut toutefois savoir que la précision et la fiabilité des estimations obtenues à partir d'un échantillon peuvent être évaluées uniquement dans le cas d'un échantillon probabiliste, le contrôle des erreurs y étant relativement facile.

Les principales étapes d'une enquête par sondage :

i) Définition des objectifs de l'enquête: Pour commencer, les objectifs de l'enquête doivent être examinés attentivement. Par exemple, pour une enquête forestière, on détermine la superficie qui sera couverte par l'enquête. Les caractéristiques sur lesquelles des informations seront collectées et le niveau de détail souhaité seront précisés. Si l'enquête porte sur des arbres, on déterminera les espèces d'arbres qui devront être recensés et l'on décidera s'il convient d'énumérer uniquement les arbres faisant partie de classes de diamètres déterminées ou si l'on estimera aussi le volume des arbres. C'est aussi durant la première étape que l'on détermine le degré de précision que devront avoir les estimations.

ii) Elaboration d'un diagramme des unités: Dans tout échantillon probabiliste, la première exigence est l'établissement d'une base de sondage. La structure d'une enquête par sondage est largement déterminée par cette base. La base de sondage est une liste des unités d'échantillonnage qui peuvent être clairement définies et identifiées dans la population. Ces unités peuvent être des compartiments, des sections topographiques, des bandes d'une certaine largeur ou des parcelles de forme et de taille définies.

L'élaboration d'une base de sondage adaptée aux objectifs d'une enquête demande de l'expérience et peut fort bien absorber une part importante des travaux de planification, en particulier dans les enquêtes forestières où il peut être nécessaire de dresser une liste artificielle des unités d'échantillonnage, faites de sections topographiques, de bandes ou de parcelles. Par exemple, dans une enquête forestière, une base de sondage peut se présenter sous la forme d'une carte appropriée de la superficie forestière. Le mode de sélection des unités d'échantillonnage doit permettre d'identifier sur le terrain une unité spécifique devant être incluse dans l'échantillon. Le choix est fonction de plusieurs facteurs: l'objet de l'enquête, les caractéristiques qui doivent être observées dans les unités sélectionnées, la variabilité entre des unités d'échantillonnage d'une taille donnée, le plan d'échantillonnage, le plan des travaux de terrain, et le coût total de l'enquête. Le choix est aussi déterminé par des considérations pratiques. Par exemple, dans des zones de collines, il n'est pas toujours possible de prendre des bandes comme unités d'échantillonnage, et les compartiments ou les sections topographiques peuvent être plus appropriés. En général, pour une intensité d'échantillonnage donnée (proportion de la surface recensée), plus les unités d'échantillonnage sont petites, plus l'échantillon est représentatif et plus les résultats ont de chances d'être précis.

Choix d'un plan d'échantillonnage: Si le plan d'échantillonnage doit être de nature à fournir une mesure statistiquement significative de la précision des estimations finales, l'échantillon doit être probabiliste, en ce sens que chaque unité de la population doit avoir une probabilité connue d'être incluse dans l'échantillon. Le choix des unités à inscrire sur la liste doit être basé sur une règle objective qui ne laisse aucune part à l'opinion de l'homme de terrain. La détermination du nombre d'unités à inclure dans l'échantillon et la méthode de sélection sont également fonction du coût admissible de l'enquête et de la précision des estimations finales.

Organisation des travaux sur le terrain: Une enquête par sondage n'est pleinement réussie que si les opérations de terrain sont fiables. Dans le domaine forestier, les travaux sur le

terrain doivent être organisés avec le plus grand soin autrement, même si le plan d'échantillonnage est excellent, les résultats de l'échantillon risqueraient d'être incomplets ou trompeurs. Le choix d'un personnel adéquat, une formation intensive, des instructions claires et une bonne supervision des opérations de terrain sont essentiels pour obtenir des résultats satisfaisants. Les équipes itinérantes doivent être capables de localiser correctement les unités sélectionnées et enregistrer les mesures nécessaires conformément aux instructions spécifiques reçues. Les superviseurs vérifient une partie de leur travail sur le terrain et s'assurent que l'enquête qui y effectuée correspond en tous points au plan.

Analyse des données : En fonction du plan d'échantillonnage utilisé et de l'information collectée, on utilisera les formules adéquates pour obtenir les estimations et calculer leur degré de précision. Une contre-vérification des calculs est souhaitable pour garantir l'exactitude de l'analyse.

Enquête préliminaire (essais pilotes): La conception d'un plan d'échantillonnage approprié à une enquête forestière demande une bonne connaissance de la théorie statistique et des données concernant la nature de la zone forestière, le mode de variabilité et le coût opérationnel. Dans le cas où l'on ne possède pas ces connaissances, il est parfois nécessaire d'effectuer une enquête pilote à petite échelle statistiquement planifiée, avant de se lancer dans une enquête à grande échelle sur toute la superficie de forêt. Ces enquêtes préparatoires, ou pilotes, fourniront les renseignements voulus sur la variabilité du matériel et offriront la possibilité d'essayer et d'améliorer les procédures en champ, de former des hommes de terrain, et d'étudier l'efficacité opérationnelle d'un plan. Une enquête pilote donnera aussi des renseignements pour estimer les différentes composantes du coût des opérations, par exemple le temps de trajet, le temps de localisation et de recensement des unités d'échantillonnage etc... Ces informations seront essentielles pour définir le type de plan et l'intensité d'échantillonnage appropriés aux objectifs de l'enquête.

Echantillonnage aléatoire simple

Dans un échantillonnage aléatoire simple toutes les combinaisons possibles d'unités d'échantillonnage tirées de la population ont les mêmes chances d'être sélectionnées. Théoriquement, l'échantillonnage aléatoire simple est la procédure la plus simple, dont s'inspirent de nombreuses autres techniques. Elle s'applique surtout au stade initial d'une enquête et aux études impliquant l'échantillonnage d'une petite surface où la taille de

l'échantillon est relativement petite. Si l'enquêteur connaît un peu la population sondée, il peut adopter d'autres méthodes plus pratiques et plus efficaces pour organiser l'enquête sur le terrain. Dans un échantillonnage aléatoire simple, la répartition irrégulière des unités d'échantillonnage sur la surface de forêt peut être un gros inconvénient dans les zones difficilement accessibles où les frais de déplacement et de localisation des parcelles sont considérablement plus élevés que les coûts de l'énumération des parcelles.

1. Sélection des unités d'échantillonnage

Dans la pratique, la sélection d'un échantillon aléatoire se fait unité par unité. Nous expliquerons dans cette section deux méthodes de sélection aléatoire pour un échantillonnage aléatoire simple sans remise.

i) Echantillonnage par tirage: Les unités de la population sont numérotées de 1 à *N*. Symboliquement, on peut assimiler ces unités à *N* boules identiques numérotées de 1 à *N*. Si on en sélectionne une au hasard après les avoir mélangées, toutes les boules ont la même possibilité d'être sélectionnées. Ce processus est répété *n* fois sans remettre en jeu les boules sélectionnées. Les unités correspondant aux numéros inscrits sur les boules sélectionnées forment un échantillon aléatoire simple de taille *n* tiré dans la population de *N* unités.

***ii) Echantillonnage au moyen de tables de nombres aléatoires* :** la procédure d'échantillonnage par tirage devient fastidieuse si *N* est élevé. Pour surmonter cette difficulté, on peut utiliser une table de nombres aléatoires, du type de celles publiées par Fisher et Yates (1963) (voir Annexe 6). Les tables de nombres aléatoires ont été conçues de manière à ce que les chiffres de 0 à 9 apparaissent indépendamment les uns des autres, à peu près le même nombre de fois dans la table. La méthode la plus simple pour choisir un échantillon aléatoire de la taille requise consiste à sélectionner un ensemble de *n* nombres aléatoires l'un après l'autre, de 1 à *N*, dans la table, puis de prendre les unités correspondant à ces numéros. Cette procédure peut comporter un certain nombre de rejets du fait que tous les nombres supérieurs à *N* qui apparaissent dans la table sont exclus d'office. Dans ces cas là, la procédure est modifiée comme suit. Si *N* est un nombre à *d* chiffres, on commence par déterminer le plus grand multiple de *N* à *d* chiffres, noté *N*'. Ensuite, on choisit un nombre aléatoire *r* de 1 à *N'* et l'unité portant le numéro égal au restant obtenu après avoir divisé *r* par *N*, est considérée comme sélectionnée. Si le reste est égal à zéro, la dernière unité est sélectionnée. Un exemple numérique est donné ci-après.

EXEMPLE1 :

Supposons que l'on doive choisir un échantillon aléatoire simple de 5 unités dans une liste de 40 unités numérotées en série, que l'on consulte l'Annexe 6 : Table de nombres aléatoires et que l'on choisisse dans la colonne 5) des nombres à deux chiffres les nombres suivants :

39, 27, 00, 74, 07

Pour donner les mêmes probabilités de sélection aux 100 unités, il faut rejeter tous les nombres supérieurs à 79 et considérer que (00) équivaut à 80. Ensuite, on divise les nombres ci-dessus par 40, et l'on prend les restes comme numéros des bandes sélectionnées pour l'échantillon, en rejetant les restes qui sont répétés. On obtient ainsi les 16 numéros de bande comme échantillon, soit : 39, 27, 40, 34, 7.

2. Estimation de paramètres

Soient $y_1, y_2, \ldots, y_n$ les mesures d'une caractéristique spécifique, effectuées sur n unités sélectionnées d'un échantillon d'une population de N unités d'échantillonnage. On constate dans le cas d'un échantillonnage aléatoire simple sans remise que la moyenne de l'échantillon

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (5.11)$$

est un estimateur non biaisé de la moyenne $\bar{Y}$ de la population. Une estimation non biaisée de la variance d'échantillonnage de $\bar{y}$ est donnée par

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{Nn} s_y^2 \quad (5.12)$$

où $$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \quad (5.13)$$

Si l'estimation $\bar{y}$ suit une loi normale, il est possible d'établir un intervalle de confiance sur la moyenne de la population $\bar{Y}$, les limites de confiance inférieure et supérieure étant définies par,

Limite inférieure $$\hat{\bar{Y}}_L = \bar{y} - z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \quad (5.14)$$

Limite supérieure $$\hat{\bar{Y}}_U = \bar{y} + z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \quad (5.15)$$

où *z* est la valeur de la table qui dépend du nombre d'observations incluses dans l'échantillon. Si leur nombre est égal ou supérieure à 30, on peut extraire ces valeurs de la table de la distribution normale (Annexe 1). Si le nombre d'observations est inférieur à 30, la valeur tabulaire sera extraite de la table de distribution *t* (Annexe 2), avec *n* - 1 degrés de liberté.

Nous allons illustrer ces calculs par un exemple. Supposons qu'une forêt ait été divisée en 1000 parcelles de 0,1 hectare chacune et qu'un échantillon aléatoire simple de 25 parcelles ait été selectionné. Pour chacune de ces parcelles d'échantillon, les volumes de bois, en m3, ont été enregistrés. Ces volumes étaient les suivants:

7 10 7 4 7

8 8 8 7 5

2 6 9 7 8

6 7 11 8 8

7 3 8 7 7

Si le volume de bois de la *i*-ème unit d'échantillonnage est noté y_i, un estimateur non biaisé de la moyenne $\bar{Y}$ de la population, s'obtient à l'aide de l'Equation (5.11), soit :

$$\hat{\bar{Y}} = \bar{y} = \frac{7+8+2+\ldots+7}{25} = \frac{175}{25}$$

$= 7\ m^3$

qui est le volume moyen de bois par parcelle de 0.1 ha , dans la superficie de forêt.

Une estimation (s_y^2) de la variance des valeurs individuelles de *y* s'obtient à l'aide de l'équation (5.13).

$$s_y^2 = \frac{(7-7)^2 + (8-7)^2 + \ldots + (7-7)^2}{25-1}$$

$$= \frac{82}{24} = 3.833$$

L'estimation non biaisée de la variance d'échantillonnage de $\bar{y}$ est donc

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1000-25}{(1000)(25)}\right)3.833$$

$$= 0.1495 \text{ (m}^3\text{)}^2$$

$$SE(\hat{\bar{Y}}) = \sqrt{0.1495} = 0.3867 \text{ m}^3$$

L'erreur-type relative, $\frac{SE(\hat{\bar{Y}})}{\hat{\bar{Y}}}(100)$ est une expression plus commune. Ainsi,

$$RSE(\hat{\bar{Y}}) = \frac{\sqrt{0.1495}}{7}(100) = 5.52\ \%$$

Les limites de confiance attachées à la moyenne de la population $\bar{Y}$ s'obtiennent par les équations (5.14) et (5.15).

Limite inférieure $\hat{\bar{Y}}_L = 7 - (2.064)\sqrt{0.1495}$

= 6.20 cordes

Limite supérieure $\hat{\bar{Y}}_U = 7 + (2.064)\sqrt{0.1495}$

= 7.80 cordes

L'intervalle de confiance de 95% associé à la moyenne de la population est de (6.20, 7.80) m^3. Cela signifie que l'on peut estimer qu'il y a 95 chances sur cent que l'intervalle de confiance de (6.20, 7.80) m^3 inclura la moyenne de la population.

On obtiendra facilement une estimation du volume total de bois dans la surface de forêt échantillonnée en multipliant l'estimation de la moyenne par le nombre total de parcelles comprises dans la population. Ainsi

$$\hat{Y} = 7(1000) = 7000 \text{ m}^3$$

avec une intervalle de confiance de (6200, 7800) obtenu en multipliant les limites de confiance associés à la moyenne par $N = 1000$. L'erreur-type relative RSE de $\hat{Y}$, n'est cependant pas modifiée par cette opération.