Cours 3

Version textuelle

Introduction à l’échantillonnage

La version interactive de cette cour est disponible gratuitement à l’adresse suivante :

https://elearning.fao.org/?lang=fr

Dans cette cours

Ce cours explique les aspects généraux de l'échantillonnage dans les inventaires forestiers.

Ce cours aborde les aspects généraux de l'échantillonnage dans les inventaires forestiers, et cherche à introduire les concepts et les caractéristiques de base d'une étude par échantillonnage, et à fournir une vue d'ensemble des composants les plus importants d'un inventaire forestier national (IFN).

Attention : Ce cours ne prétend pas former convenablement des experts pour les statistiques sur échantillon nécessaires pour planifier, analyser, communiquer et interpréter correctement les estimations à partir d'échantillons d'un IFN.

À qui ce cours s'adresse-t-il?

Ce cours s'adresse principalement aux personnes impliquées dans les phases d'échantillonnage et d'analyse d'un IFN, mais peut être suivi par quiconque s'intéresse au sujet. Ce cours vise particulièrement:

1. Les techniciens forestiers responsables de la mise en œuvre des IFN de leur pays.
2. Les équipes du suivi national des forêts.
3. Les étudiants et les chercheurs, en tant que matériel programmatique dans les écoles forestières et les cours universitaires.
4. Les jeunes et les nouvelles générations d'agents forestiers.

Structure du cours

Ce cours comprend trois leçons.

Leçon 1: L'échantillonnage	Cette leçon introduit les concepts de base et les termes associés à l'échantillonnage statistique. Elle fournit une vue d'ensemble des caractéristiques importantes d'une étude par échantillonnage et explique les bases de l'échantillonnage pour un public non expert.
Leçon 2: Éléments de conception d'une étude par échantillonnage	Cette leçon présente les bases des éléments de conception des études par échantillonnage qui sont pertinentes pour les IFN, et les concepts à prendre en compte pour préparer une stratégie

	d'échantillonnage. Elle explique aussi comment calculer la taille de l'échantillon associée.
Leçon 3: Conception de l'estimation	Cette leçon s'intéresse aux méthodes et aux formules nécessaires pour dériver des estimations non biaisées des données relevées en suivant une certaine stratégie d'échantillonnage.

À propos de la série

Ce cours conclut une série de huit cours individualisés couvrant divers aspects d'un IFN. Voici un aperçu de la série complète.

Cours	Apprentissages
Cours 1: Pourquoi un inventaire forestier national (IFN)?	Objectifs et but d'un IFN, et comment les IFN informent la conception de politiques et la prise de décisions dans le secteur forestier.
Cours 2: Préparation d'un inventaire forestier national	La planification et le travail nécessaire pour mettre en place un IFN efficace ou un système national de suivi des forêts (SNSF).
☞ **Cours 3: Introduction à l'échantillonnage**	**(Vous suivez actuellement ce cours).**
Cours 4: Introduction au travail de terrain	Considérations pour le travail de terrain, les variables au niveau parcellaire et les mesures au niveau de l'arbre.
Cours 5: Gestion de données dans un inventaire forestier national	Collecte d'information et gestion de données pour les IFN.
Cours 6: Assurance qualité et contrôle qualité dans un inventaire forestier national	Procédures d'AQ et de CQ dans la collecte et la gestion de données d'un inventaire forestier.
Cours 7: Éléments de l'analyse de données	Approches/calculs typiques dans les analyses de données et questions connexes.

	Cours 8: Résultats de l'inventaire forestier national: notification et diffusion	Publication des résultats de l'IFN et importance de la notification dans le contexte des actions REDD+.

Leçon 1: L'échantillonnage

Introduction de la leçon

Cette leçon introduit les concepts de base et les termes associés à l'échantillonnage statistique.

Elle fournit aussi une vue d'ensemble des caractéristiques importantes d'une étude par échantillonnage et explique les bases de l'échantillonnage pour un public non expert.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire l'importance de l'échantillonnage dans les inventaires forestiers.
2. Définir la logique d'un échantillonnage statistique.
3. Expliquer les concepts de base et la terminologie associés à l'échantillonnage.
4. Expliquer l'importance de l'exactitude et de la précision pendant le processus d'estimation.

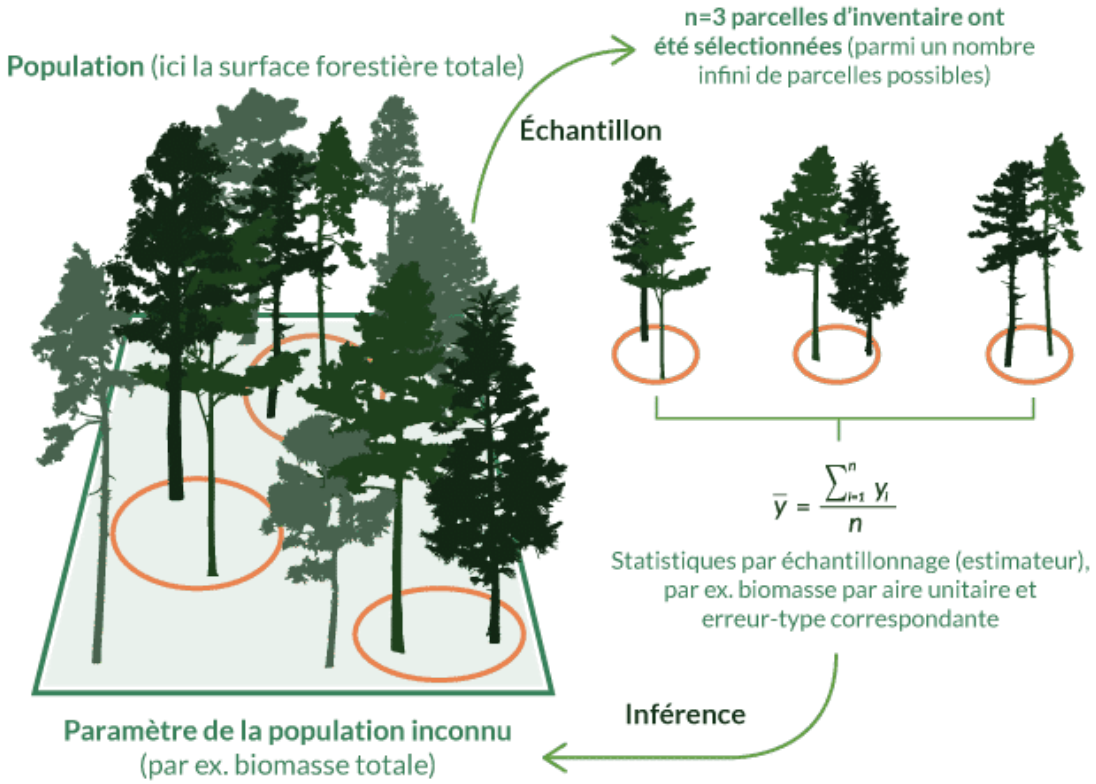
Pourquoi l'échantillonnage est-il nécessaire?

Avant de commencer à étudier les aspects importants de l'échantillonnage statistique, revenons en arrière pour nous intéresser brièvement à la logique fondamentale des études par échantillonnage en général.

Pourquoi l'échantillonnage est-il un concept si fondamental dans le contexte des inventaires forestiers et du suivi des forêts?

La réponse à cette question est très simple: Lorsque l'on se penche sur l'évaluation de terrain des variables centrales, il n'est **ni possible ni efficace d'observer tous les éléments** dans la surface forestière d'un pays. Les experts doivent plutôt établir des inférences concernant la situation actuelle et les changements des variables cibles en réalisant des observations sur des sous-ensembles ou «échantillons» relativement petits de la surface forestière totale.

On peut imaginer que l'échantillonnage s'apparente à ouvrir de petites fenêtres qui permettent de regarder des parties de la population afin d'obtenir une impression de l'ensemble de cette population.



Une observation plus fine de ces échantillons de l'aire totale montre que la crédibilité des résultats d'une étude par échantillonnage est influencée par **la manière dont on choisit les échantillons, les méthodes utilisées pour obtenir les observations singulières et les techniques et calculs d'estimation appliqués**. Ce sont les trois éléments de conception d'une étude par échantillonnage qui doivent être planifiés, avec les considérations statistiques – ce que nous aborderons plus en détail dans la prochaine leçon.

Qu'est-ce qu'un échantillonnage statistique?

Le processus de sélection garantit que les éléments de l'échantillonnage peuvent être considérés comme représentatifs de la population. Lorsque l'échantillonnage suit les règles des statistiques, on parle d'**échantillonnage statistique**. L'échantillonnage statistique est largement déterminé par la randomisation (et l'absence de considérations subjectives ou arbitraires), ce qui signifie qu'en appliquant une sélection aléatoire, on garantir que chaque élément de la population a une probabilité d'être sélectionné définie et connue.

D'autres critères de sélection, comme l'**impartialité** ou l'**objectivité** ne sont pas suffisants. Puisque les probabilités de sélection jouent un rôle central dans l'échantillonnage statistique, ces techniques sont aussi appelées **échantillonnage probabiliste**. Ainsi, la représentativité de l'échantillon est garantie, et des estimateurs non biaisés (soit des approches d'estimation statistiquement correctes) sont disponibles pour la plupart des conceptions courantes de l'échantillonnage et de l'observation.

Une sélection subjective des éléments de la population «les plus représentatifs» ne correspond pas à l'échantillonnage statistique et ne permet pas d'estimation statistique ni d'inférence.

Imaginez envoyer des experts avec pour tâche de trouver la parcelle «la plus représentative» dans une surface forestière (en termes de densité des arbres, mélange d'espèces, déclivité, conditions du sol, etc.). Il est assez évident qu'une estimation dérivée d'une telle parcelle serait exclusivement liée au choix de l'expert (alors qu'un autre expert pourrait très bien arriver à un choix différent).

Bien qu'une estimation logique experte puisse être correcte et proche de la valeur de la population cible, tout dépend de l'expert et aucune approche méthodologique objective n'est définie qui pourrait être répétée ailleurs. L'échantillonnage statistique, au contraire, est transparent à chaque étape méthodologique.

Le saviez-vous?

Un grand nombre de techniques d'échantillonnage statistique ont été inventées et présentées dans le contexte des inventaires forestiers. Bien que les parcelles d'échantillonnage étaient déjà largement utilisée dans la foresterie au 19e siècle, une technique plus formalisée d'échantillonnage statistique pour de grandes populations n'a été mise au point – et graduellement acceptée – en tant que méthodologie pour produire des résultats valides qu'autour de 1900: en 1895, le statisticien norvégien A.N. Kiaer a présenté une approche de l'échantillonnage alors appelée «la **méthode représentative**», où la «**représentativité**» jouait un rôle central.

Les statisticiens des inventaires forestiers ont significativement contribué à cette époque à l'analyse de l'échantillonnage systématique en ligne. Le premier IFN se fondant sur l'échantillonnage statistique a été mis en œuvre en 1919-1930 en Norvège. Cela a été suivi par d'autres pays d'Europe du Nord au début des années 1920: la Finlande en 1921-1924 et la Suède en 1923-1929..

La validité statistique est l'une des principales caractéristiques de l'échantillonnage statistique, tel qu'appliqué au suivi des forêts. C'est uniquement en adhérant aux principes de l'échantillonnage statistique que la conception de l'inventaire choisie peut être défendue de manière probante, lorsque, par exemple, des doutes sont émis concernant les résultats.

Dériver des inférences à partir d'un échantillon

Les statistiques descriptives s'intéressent à la caractérisation quantitative d'une population d'intérêt, ou au domaine pour lequel ces relevés descriptifs doivent être produits.

L'échantillonnage vise à dériver des inférences/conclusions concernant la population totale à partir d'un nombre limité d'éléments de l'échantillonnage sélectionnés. Dans les inventaires forestiers, ces éléments sont typiquement les parcelles d'échantillonnage, qui sont des sous- ensembles de la surface forestière totale.

À partir de l'analyse des observations relevées des variables cibles dans ces parcelles d'échantillonnage, on peut dériver une estimation statistique du vrai paramètre de la population inconnu. Par exemple: à partir de la biomasse par parcelle de *n* parcelles d'échantillonnage, on peut produire une estimation de la biomasse par hectare de la population entière. On comprend intuitivement que l'on ne pas attendre de cette estimation qu'elle soit égale à la vraie valeur – c'est une approximation, et elle variera dès que l'on prendra un autre échantillon en suivant la même conception de l'inventaire.

Note

Les vraies valeurs dans une population sont appelés **paramètres**, tandis que les estimations produites à partir des études par échantillonnage sont désignées par **statistiques**. La vraie Valeur moyenne d'une population, la moyenne paramétrique, est estimée à partir de la valeur moyenne dans l'échantillon. Il est important de faire clairement cette distinction : **les vrais paramètres ne seront jamais connus, mais estimés par l'analyse statistique de l'échantillon**. La vraie valeur est une constante, une valeur fixe. La valeur produite par l'analyse statistique de l'échantillon (= la valeur estimée) est une variable aléatoire qui peut prendre de nombreuses valeurs différentes – selon quel échantillon a été sélectionné – et suit une certaine distribution.

Voyons quelques exemples de l'application à un inventaire forestier pour la biomasse des définitions fournies.

1. La population, par exemple, des arbres, est déterminée par une aire, représentée par un nombre infini de points centraux adimensionnels où les parcelles d'échantillonnage peuvent être sélectionnées.
2. L'échantillon consiste en un certain nombre de parcelles (effectif de l'échantillon) qui ont été sélectionnées en fonction du plan d'échantillonnage.
3. La vraie valeur - ou le paramètre de la population - par exemple, de la biomasse, serait la biomasse moyenne estimée de toutes les positions infinies d'échantillonnage possibles dans l'aire. Puisque l'on dispose uniquement du plan parcellaire en cours, la vraie valeur reste

inconnue.

4. En utilisant un plan parcellaire et un plan d'estimation adaptés, on peut dériver une estimation non biaisée de l'échantillon à disposition.

Estimateur et estimation

Lorsque l'on parle d'un **estimateur** dans un échantillonnage statistique, il s'agit de l'algorithme ou la formule de calcul utilisée pour produire une estimation. Afin de produire des estimations statistiques, l'estimateur doit refléter: **Le processus de sélection sous-jacent** des éléments de l'échantillonnage; et **la manière dont les observations singulières de l'élément de l'échantillonnage ont été obtenues**

Qu'est-ce que le concept de population sous-jacente dans les inventaires forestiers?

Lorsque les parcelles d'échantillonnage sont les «éléments de l'échantillonnage» qui sont sélectionnés, la question qui se pose est «qu'est-ce que la population»? En termes généraux, la population est définie comme l'ensemble des éléments de l'échantillonnage qui peuvent théoriquement être sélectionnés. Dans un inventaire forestier, on utilise habituellement des parcelles d'échantillonnage dont la position est sélectionnée en choisissant un point échantillon. La population est alors définie comme tous les points échantillons possibles dans l'aire d'intérêt. Combien sont-ils?

Le nombre de points dans toute aire est infini. Les points échantillons sont sélectionnés dans un continuum, et on appelle cela une population infinie. La population est alors «le nombre total de parcelles d'échantillonnage possibles dans l'aire d'étude définie», où les parcelles d'échantillonnage sont installées autour des points échantillons.

Néanmoins, puisque de nombreuses variables d'intérêt sont des agrégats des mesures des arbres singuliers trouvés sur ces parcelles d'échantillonnage, elles ne varieront que lorsque la composition des arbres inclus changera.

Ainsi, pour ces variables, on peut affiner le concept de population est dire que «la population est composée de tous les clusters d'arbres mutuellement exclusifs qui ont une probabilité positive d'être inclus conjointement par le plan parcellaire défini». L'effectif de cette population n'est pas infini, il y a un nombre fini d'options pour les inclusions conjointes des arbres spatialement dispersé.

Limitations pour les conclusions

À partir d'un échantillon, on peut dériver des conclusions/inférences uniquement concernant une partie de la population qui a une probabilité positive de faire partie de l'échantillon. On appelle cette partie de la population le **cadre d'échantillonnage.** Dans le meilleur des cas, le cadre d'échantillonnage comprend la population d'intérêt complète, mais en réalité, il exclut en général des parties de la surface forestière pour diverses raisons, comme le manque d'accessibilité ou le risque d'entrée.

Toutes nos estimations se réfèrent uniquement à l'ensemble des éléments de l'échantillonnage qui sont dans le cadre d'échantillonnage et il faut s'assurer que le cadre d'échantillonnage couvre le maximum possible de la population complète.

Population versus cadre d'échantillonnage

Imaginons un pays qui n'utilise pas de définition biophysique des forêts – fondée sur des critères quantitatifs et qualitatifs –, mais une définition administrative ou juridique des «terres forestières». Si la sélection des éléments de l'échantillonnage est limitée à ce cadre d'échantillonnage, on ne peut pas dériver de conclusions concernant les arbres et les forêts biophysiques qui se trouvent en dehors des terres forestières définies. Toutes les conclusions se réfèreraient exclusivement à la surface forestière comprise dans les terres forestières définies.

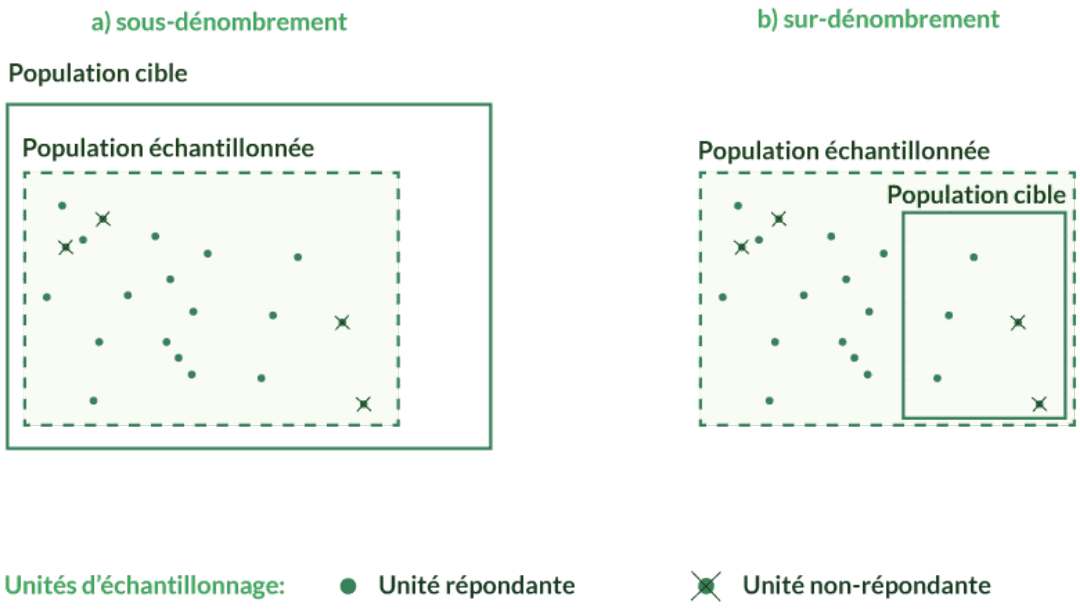
Ainsi, à la fois la population et le cadre d'échantillonnage doivent être clairement définis et mentionnés durant la communication et l'interprétation des résultats.

En outre, certains points dans le cadre d'échantillonnage peuvent ne pas être accessibles pour des questions de refus d'autorisation ou de sécurité. Ces observations manquantes sont appelées **non-réponses.** La différence entre **le cadre d'échantillonnage** et la **non-réponse** est ainsi définie: Le cadre d'échantillonnage définit les éléments de l'échantillonnage qui peuvent supposément être sélectionnés et mesurés. Mais il arrive que certains points échantillons s'avèrent inaccessibles, constituant alors des points de non-réponse, et différentes techniques sont disponibles pour traiter ce problème.

Généralement, les taux de non-réponse sont relativement faibles dans les IFN. Bien qu'il existe des techniques d'imputation pour réaliser des prédictions à partir de modèles des observations hypothétiques des parcelles de non-réponse, elles sont généralement ignorées dans les IFN et la taille de l'échantillon est réduite.

La figure ci-dessous montre qu'une population cible (aire dans la ligne pleine) peut souvent ne pas coïncider avec la population échantillonnée (aire dans la ligne pointillée). L'exemple A montre un sous-dénombrement, très typique dans les IFN où certaines aires de la population ont été classifiées au préalable comme non accessibles, par exemple. L'exemple B illustre un sur-dénombrement, plus rare dans le contexte des IFN, mais possible si la population cible d'intérêt est une sous-population particulière du pays, alors que l'échantillonnage a initialement été planifié pour le pays entier.

Dans les deux cas, des unités d'échantillonnage étaient accessibles (répondantes) et certaines inaccessibles (non-répondantes).



Concepts de base de l'échantillonnage

Avant d'approfondir des considérations pratiques concernant l'échantillonnage dans le contexte des IFN, familiarisons-nous avec des concepts importants et des termes centraux. Bien que les statistiques tendent à devenir complexes et difficiles à assimiler, beaucoup de choses exprimées dans des formules complexes sont en fait relativement faciles à comprendre avec des mathématiques de base – et c'est souvent assez intuitif.

Dans la section suivante, nous allons aborder plusieurs concepts statistiques et nous pencher uniquement sur ceux qui sont pertinents pour les IFN. Néanmoins, il y a beaucoup plus à apprendre sur les inventaires forestiers que juste quelques concepts.

Concepts et terminologie importants

Lorsque l'on prend un échantillon d'une population (ou d'un cadre d'échantillonnage), il n'y a pas de résultat unique: chaque sélection d'un nouvel échantillon alternatif donnera une estimation différente qui est aussi valide que toutes les autres.

Comme l'on ne peut pas déterminer la seule et unique **vraie valeur** (appelée **paramètre**) de la population à partir d'un échantillon, l'estimation que l'on dérive présente une incertitude. On ne pourra déterminer la marge de cette incertitude uniquement si la sélection des échantillons suit des critères statistiques, et si les formules appliquées ou les estimateurs sont corrects.

En fait, lorsque l'on détermine cette marge, il s'agit aussi d'une estimation. **L'intervalle** de confiance est une mesure d'incertitude typique, qui définit un intervalle autour de la valeur estimée, dans lequel la vraie valeur devrait se trouver avec une probabilité définie.

Qu'est-ce que la taille de l'échantillon?

La taille de l'échantillon se réfère au nombre d'observations (éléments de l'échantillonnage observés) **sélectionnés indépendamment** qui sont tirés du cadre d'échantillonnage. Ici, le terme indépendant signifie que la sélection d'un élément n'a aucun effet sur la sélection d'un autre. Ce processus de sélection a lieu si des éléments singuliers de l'échantillon sont sélectionnés aléatoirement.

Dans les inventaires forestiers, cependant, c'est rarement le cas, car les échantillons sont relevés à intervalles fixes. Il est important de remarquer que la «sélection indépendante» décrite ici ne doit pas être confondue avec l'«indépendance des variables», qui un concept complètement différent.

Plus d'information sur la manière de déterminer la taille de l'échantillon pour diverses conceptions d'inventaire forestier sera fournie dans la leçon 2.

Quelle est la différence entre l'intensité d'échantillonnage et la taille de l'échantillon?

L'**intensité d'échantillonnage** se réfère à la proportion du cadre d'échantillonnage qui est observée. La **taille de l'échantillon**, par ailleurs, concerne le nombre absolu d'éléments de l'échantillonnage

sélectionnés (indépendamment).

L'intensité d'échantillonnage est définie comme la fraction de la population des éléments de l'échantillonnage contenue dans l'échantillon. Cependant, un tel concept n'étant pas applicable à une population infinie, on définit l'intensité d'échantillonnage dans les inventaires forestiers par surface: c'est la fraction de l'aire échantillonnée (= la somme de toutes les surfaces des parcelles) par rapport à la surface totale qui définit la population.

Qu'est-ce que la variance de la population?

La variance de la population **quantifie la variabilité dans la population**. C'est une caractéristique de la population des éléments de l'échantillonnage. Autrement dit, pour chaque élément de la population, il existe une valeur pour une variable cible spécifique, comme la biomasse par hectare. La variance de la population paramétrique est la vraie variance de toutes ces valeurs. Et cette vraie variance (paramétrique) peut être estimée à partir d'un échantillon.

Quelle est la différence entre la variance de la population et la variance d'erreur?

Cette distinction est un élément clé pour comprendre une grande partie des statistiques à partir d'échantillon importantes dans les IFN. Alors que la variance de la population est une estimation de la variabilité parmi les éléments de la population (observations des parcelles inventoriées), la variance d'erreur est une propriété de l'échantillon. Cela signifie qu'elle quantifie la variation attendue parmi les estimations répétées de la même variable cible (par ex. la biomasse moyenne par hectare).

Supposons qu'un IFN est réalisé de manière répétée mille fois, chaque fois avec une nouvelle sélection de parcelles. Dans ce cas, la variation parmi toutes les moyennes singulières est une estimation de la variance d'erreur. Cette information est importante pour juger la qualité d'un échantillon, car elle répond à des questions importantes: qu'arriverait-il si l'on répétait un échantillon encore et encore? Obtiendrait-on toujours des résultats assez similaires, ou pourrait-on attendre des inventaires répétés qu'ils donnent des résultats largement variables?

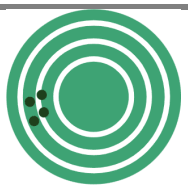
Dans le premier cas, on dirait que l'estimation est précise, tandis que dans le second cas, l'estimation est moins précise. Habituellement, on ne rapporte pas la variance d'erreur mais sa racine carrée: l'**erreur-type**. C'est l'une des plus importantes statistiques estimées à partir d'un échantillon, car elle quantifie la précision de l'estimation. La raison pour laquelle la racine carrée est rapportée et non pas la variance

d'erreur est très simple: l'erreur-type est exprimée dans la même unité que l'estimation elle-même, ce qui la rend beaucoup plus facile à comprendre.

Il apparaît intuitivement que la confiance et la crédibilité ou la certitude des résultats dépendent de cette variance d'erreur. Si aucune information n'est donnée sur la variance d'erreur, un utilisateur de l'information peut conclure qu'un inventaire unique n'est pas suffisant (car le prochain produira probablement une estimation différente).

Exactitude et précision

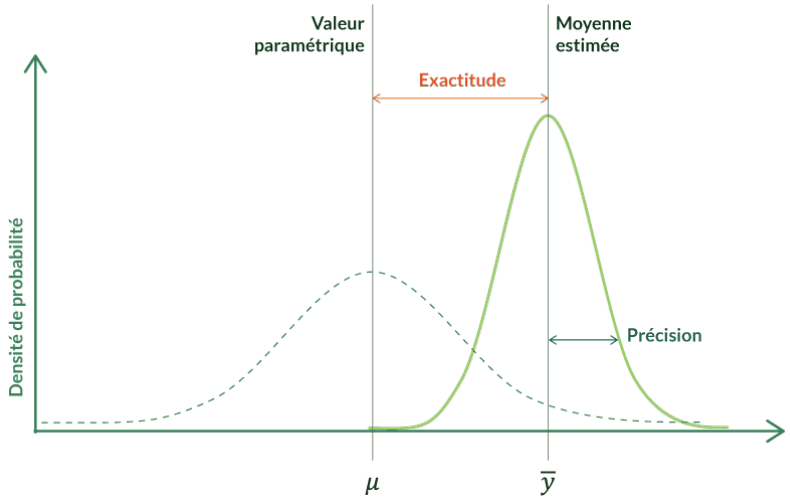
Nous avons abordé le concept de précision plus haut, mais penchons-nous maintenant sur l'importance et la signification de l'exactitude et de la précision telles qu'on les utilise dans l'échantillonnage des inventaires forestiers. Nous comprendrons mieux ces concepts avec l'exemple d'un jeu de fléchettes.

Faible précision, haute exactitude	**Faible exactitude, haute précision**
Imaginons que vous tirez quatre fléchettes. La distribution des tirs autour du centre de la cible est une expression de votre exactitude (la moyenne des tirs donnera une position proche du centre).	Poursuivons avec les fléchettes: la répartition des tirs individuels est une expression de votre précision (les tirs répétés sont proches).
	

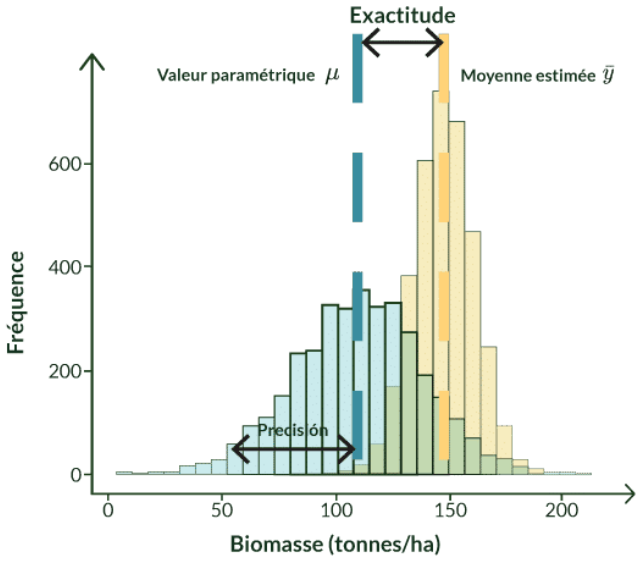
Comme on le voit dans le graphique ci-dessous, on peut représenter une distribution de toutes les observations relevées (qui sont ici des valeurs singulières sur l'axe x). Les valeurs de l'axe y indiquent la fréquence relative des observations pour les valeurs respectives.

Alors qu'une distribution solide donne une précision relativement haute (distribution étroite) de la moyenne estimée $\bar{y}$, comme dans le graphique, elle n'est pas très exacte (elle est biaisée) du fait de son écart du vrai paramètre μ. Au contraire, la distribution de la ligne pointillée donne une estimation très

exacte de la moyenne (ici, identique à la valeur paramétrique μ), mais avec une précision relativement faible.



Prenons maintenant deux exemples où la biomasse moyenne par hectare est mesurées dans 3 500 parcelles d'échantillonnage. L'histogramme jaune montre une distribution avec une précision relativement haute (distribution étroite) de la moyenne estimée, mais une exactitude faible (biaisée) du fait de son écart du vrai paramètre. Au contraire, la distribution de l'histogramme bleu reflète une estimation très exacte de la moyenne (ici, identique à la valeur paramétrique μ), mais avec une précision relativement faible, du fait d'une plus grande amplitude de la distribution.



Si l'on revient au graphique précédent, rappelons que la valeur produite par l'analyse statistique de l'échantillon calculée pour un échantillon n'est qu'une estimation de la vraie biomasse de la population, fondée sur un ensemble de parcelles sélectionnées. Imaginons que l'on répète l'estimation de la valeur produite par l'analyse statistique de l'échantillon avec une sélection différente mais le même plan d'échantillonnage: on produira des estimations différentes de la biomasse de la population. La distribution de ces moyennes estimées représente la fréquence relative de ces différentes estimations de la biomasse de la population.

L'amplitude de la distribution est une expression de la variabilité (ou dispersion) autour de la valeur moyenne estimée ($\bar{y}$). Si la dispersion de ces valeurs est faible et elles sont relativement proches, on peut conclure que des échantillons alternatifs répétés seront susceptibles de donner des estimations similaires. Ainsi, l'amplitude de cette distribution permet aussi de se prononcer sur la précision (voir le graphique ci-dessus).

Par ailleurs, **l'exactitude** – ou la correction – est l'écart de la valeur attendue à partir des échantillons répétés par rapport au vrai paramètre de la population. Cet écart est aussi appelé biais ou **biais de l'estimateur**. Puisque la vraie valeur reste inconnue, la taille de cet écart ne peut pas être quantifiée à partir de l'échantillon. Il s'agit plutôt d'une propriété de l'estimateur appliqué et une expression d'une erreur systématique qui ne peut pas être compensée en augmentant la taille de l'échantillon.

La seule manière de garantir des estimations «non biaisées» est une preuve mathématique du fait que le plan d'échantillonnage et les méthodes appliquées permettent des estimations correctes (non-biaisées par conception) ou des simulations empiriques (en cas où les estimations s'appuient sur l'application d'un modèle).

¡ Rappelez-vous: Dans les études par échantillonnage, on n'a aucune information sur la vraie valeur de la population (le centre de la cible), on ne dispose que de l'échantillon (les fléchettes). On navigue à l'aveugle quant à la position du centre, et seule l'utilisation d'estimateurs non biaisés peut garantir l'exactitude.

Quelles sont les raisons possibles d'estimations biaisées? Découvrons-le.

Biais de sélection	Une sélection non-statistique a été utilisée, et il n'y a pas de garantie de la représentativité de l'échantillon (par ex. une sélection subjective de parcelles proches de la route).
Biais de l'observateur	Les observations ou mesures sont systématiquement fausses (par ex. le dhp est toujours mesuré à une hauteur de 1 m au lieu de 1,3 m).
Biais de l'estimateur	Un calcul systématiquement faux (par ex. l'application constante d'un mauvais facteur d'extension des parcelles, de sorte que toutes les observations des parcelles sont trop élevées).
Biais du modèle	En cas de techniques d'échantillonnage fondées sur modèle ou assistées par modèle, mais aussi en cas d'observation modélisée (par ex. l'application des mauvais modèles de biomasse), un biais potentiel du modèle affectera directement le biais de l'estimation.

Note

Signification limitée de l'intensité d'échantillonnage concernant la précision de l'estimation

Dans les directives d'inventaire, ou même dans les réglementations gouvernementales, on trouve parfois des seuils d'intensité d'échantillonnage (proportion minimum des aires) qui doivent être échantillonnés (par ex. au moins 3 pour cent de la surface forestière). Cependant, cette intensité d'échantillonnage est très peu significative pour la précision des estimations produites. La précision dépend de la taille de l'échantillon. Regardez de plus près les estimateurs présentés à la fin de cette leçon, et vous verrez que l'«intensité d'échantillonnage» n'est présente dans aucune des formules.

Estimations ponctuelles et par intervalle

Généralement, la valeur estimée seule n'est pas une information suffisante pour une interprétation

correcte ou une publication et prise de décisions. Rappelons que l'on n'a pas observé l'ensemble de la population, mais dérivé une estimation à partir d'un échantillon. Si l'on rapporte une moyenne estimée (par ex. le volume moyen ou la biomasse par unité de surface), que l'on appelle une **estimation ponctuelle**, cette information seule ne permettra pas de juger de la qualité (ou la fiabilité, crédibilité ou certitude) de cette estimation.

Plus d'information sera nécessaire concernant la précision estimée de cette estimation ponctuelle afin d'informer sa qualité. Cette information est donnée en termes d'un intervalle autour de la moyenne estimée, dans lequel la vraie valeur devrait se trouver avec une certaine probabilité – et c'est ce que l'on appelle une **estimation par intervalle**.

Astuces rapides!

Communication des estimations

Lorsque l'on rapporte des estimations, la bonne pratique consiste à dire: «d'après notre étude par échantillonnage, nous estimons le matériel sur pied à 200 m^3/ha ± x», et non pas «d'après notre étude par échantillonnage, nous concluons que le matériel sur pied est égal à 200 m^3/ha.»

Le fait qu'il s'agisse exclusivement d'estimations est aussi évident du fait que les estimations des valeurs moyennes (estimations ponctuelles) sont accompagnées d'estimations de la précision de ces valeurs moyennes estimées (estimations par intervalle).

Les questions qui se posent alors immédiatement sont:

- À partir de combien d'observations indépendantes (parcelles) cette moyenne a-t-elle été estimée?
- Comment la variation entre ces observations singulières nous informe-t-elle quant à la variance de la population?
- Quelle est la variation attendue de cette moyenne si l'on répète (virtuellement) l'échantillon de

nombreuses fois avec le même plan (= variance d'erreur)?

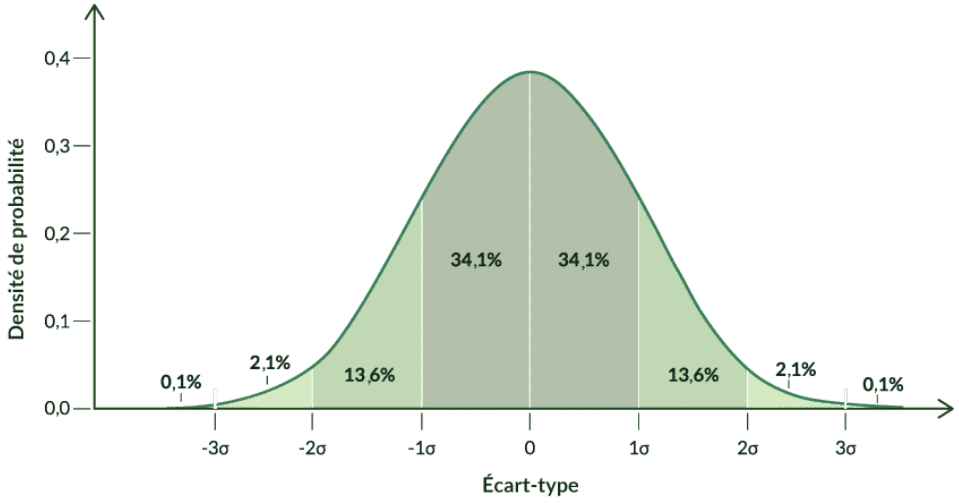
Toutes les questions ci-dessus influencent l'amplitude de ce que l'on appelle l'intervalle de confiance autour de la moyenne estimée. L'intervalle de confiance est une affirmation probabiliste à partir de laquelle on peut apprendre dans quel intervalle autour de la moyenne estimée on peut s'attendre à trouver le vrai paramètre de la population (inconnu) avec une probabilité définie.

Cependant, cela est uniquement possible si l'on adopte une certaine distribution des estimations, et c'est là où une propriété intéressante des échantillons statistiques entre en jeu. This, however, is only possible if we assume a certain distribution of estimates, and this is the point where an interesting property of statistical samples comes into play.

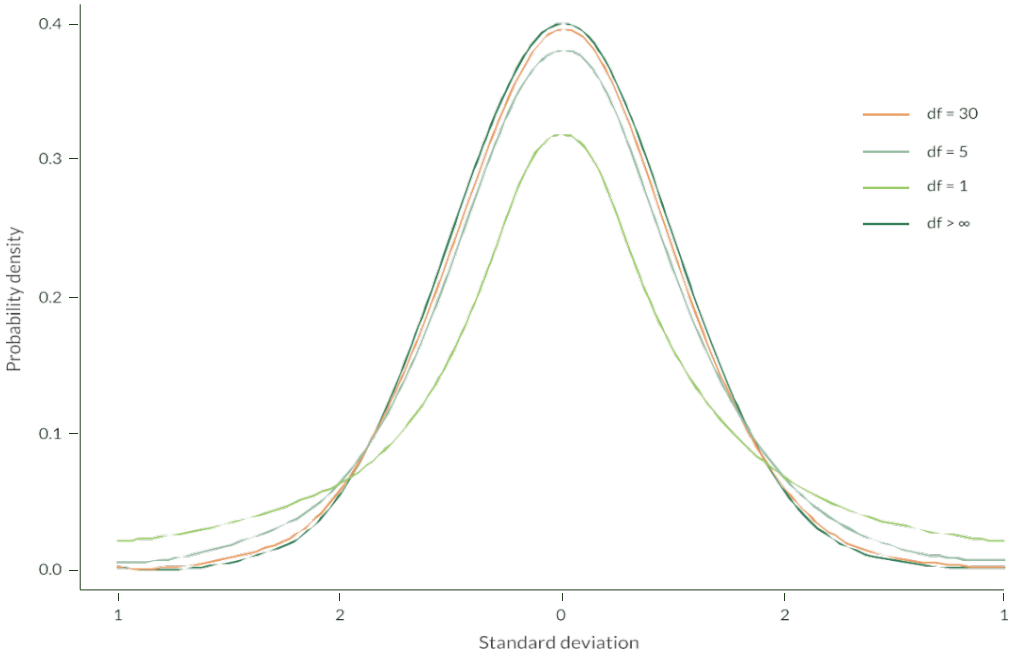
Distribution des échantillons

Une caractéristique très intéressante des échantillons permet la définition de cet intervalle: les estimations à partir d'un échantillonnage répété tendent à suivre une loi normale (ou distribution normale). Cela est vrai pour les grands échantillons avec un effectif de l'échantillon supérieur à 30, ce qui est considéré en statistiques par échantillonnage comme un seuil indicatif qui distingue les petits des grands échantillons; les valeurs de la moyenne estimée des échantillons plus petits suivent la loi t de Student. Pour les grands échantillons, on peut utiliser la loi normale pour déterminer les limites supérieure et inférieure de l'intervalle dans lequel on peut s'attendre à trouver la vraie valeur avec une probabilité définie (par ex. 95 pour cent).

Le graphique ci-dessous illustre une loi normale et une loi t de Student légèrement différente. Toutes deux permettent de dériver un intervalle dans lequel on peut s'attendre à trouver la vraie valeur paramétrique avec une probabilité définie.



Graphique illustrant la loi normale des échantillons. Diagramme tiré de Wikipédia(opens in a new tab), auteur M. W. Toews, sous licence Creative Commons License.



Intervalles de confiance

Dans le cadre du processus d'estimation, il convient d'évaluer le niveau de confiance que l'on peut avoir dans les estimations. Cela est reflété par la proximité de l'estimation avec le vrai paramètre, pour chaque échantillon relevé. Si pour tous les échantillons possibles, les estimations étaient très proches du

vrai paramètre de la population, on aurait une confiance élevée dans les estimations. Pour l'évaluer, on utilise souvent les intervalles de confiance.

Formellement, on peut affirmer que la probabilité P que le vrai paramètre μ se trouve entre une limite inférieure et une limite supérieure est de x %. Plus cette probabilité est élevée, plus la confiance dans l'estimation le sera aussi. Par exemple, dans le cas spécifique de l'estimation de la moyenne, les intervalles de confiance estimés (exprimés dans les mêmes unités que l'estimation moyenne) définiront les limites comme:

$$\bar{y} - C.I. \leq \bar{y} \leq \bar{y} + C.I.$$

où l'intervalle de confiance C.I. est défini par la valeur de la loi t de Student et l'erreur-type de l'estimation:

$$C.I. = t\, S_{\bar{y}}$$

Habituellement, des intervalles de confiance de 95 pour cent sont donnés. L'origine de ces intervalles de confiance de 95 pour cent remonte loin dans l'histoire des statistiques et il n'y a pas d'argument parfaitement probant justifiant une probabilité d'erreur de 5 pour cent. Un autre intervalle de confiance (par exemple, 90 pour cent) pourrait aussi bien être utilisé, tant que cela est clairement dit.

Erreur-type des estimations

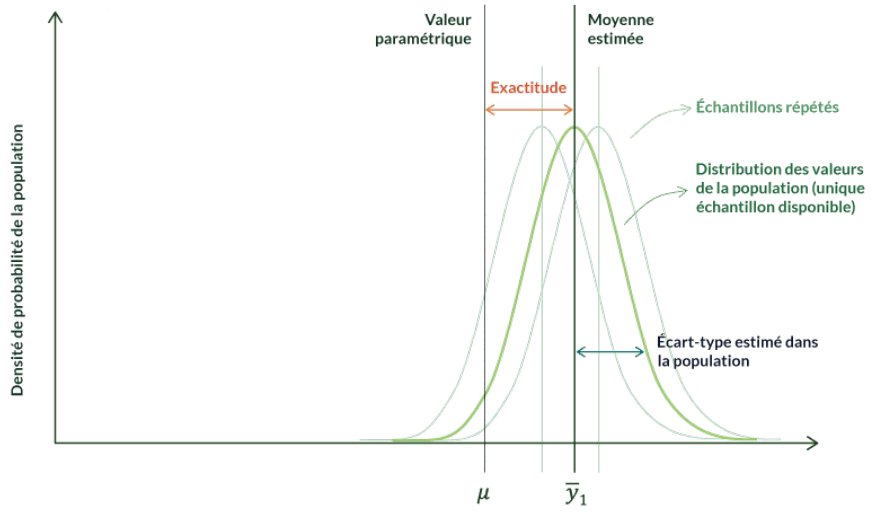
Si l'on regarde l'échantillon singulier à disposition (le seul inventaire que l'on a mené), comment peut-on dériver une attente de la variation de tous les autres échantillons possibles avec le même plan pour la même population? En pratique, bien sûr, on ne peut répéter l'IFN de nombreuses fois.

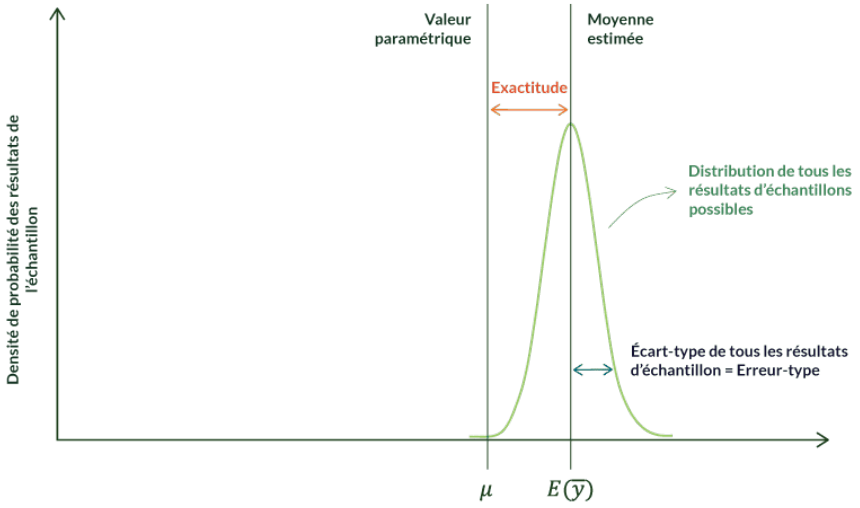
Mais l'on a appris qu'il est possible de tirer des conclusions concernant la variabilité (imaginée) d'échantillons répétés à partir de l'échantillon singulier dont on dispose. La mesure de cette variabilité est la **variance d'erreur.**

Et ce que l'on appelle **l'erreur-type** est la racine carrée de cette variance d'erreur. En d'autres termes, c'est l'écart-type estimé de tous les résultats d'échantillonnage possibles. L'erreur-type est la mesure de

la précision de l'estimation la plus fréquemment rapportée. Contrairement à la variance d'erreur, elle est exprimée dans la même unité que la statistique estimée. Elle est donc plus facile à interpréter que la variance d'erreur.

La figure suivante aide à distinguer ces deux perspectives différentes. Dans le graphique supérieur, on peut voir la distribution (variabilité) des éléments de la population (par ex. les valeurs des parcelles) à partir d'un échantillon singulier (ligne en gras). Cependant, l'échantillon singulier disponible n'est que l'un des nombreux échantillons possibles (vert clair) que l'on pourrait potentiellement concevoir. Dans le graphique inférieur, on voit la distribution de tous les résultats d'échantillonnage potentiels autour de la «valeur attendue» et l'erreur-type est l'écart-type de cette distribution.





Estimation par échantillonnage aléatoire simple (EAS)

Nous arrivons au dernier segment de cette leçon. Dans cette section, nous aborderons des explications plus détaillées qui seront traitées dans la prochaine leçon, et considèrerons des estimateurs pour l'échantillonnage aléatoire simple (EAS).

L'échantillonnage aléatoire simple se réfère à une sélection aléatoire indépendante de chaque élément singulier de l'échantillonnage. Cela signifie que l'on adopte une sélection aléatoire non restreinte des positions d'échantillonnage dans une surface forestière. L'échantillonnage aléatoire non restreint

signifie que les éléments de l'échantillonnage ont la même probabilité de sélection. C'est le fondement de base des statistiques par échantillonnage, et un échantillonnage très adapté pour expliquer les estimateurs, car il est assez simple de déterminer les probabilités de sélection, qui sont ici égales pour tous les éléments.

Même s'il est rarement appliqué aux inventaires forestiers, ce plan d'échantillonnage (ou procédure de sélection) est fondamental pour toutes les statistiques, car les estimateurs existants sont assez simples, les caractéristiques de l'échantillonnage statistique peuvent facilement être expliquées, et il est intéressant à mentionner du point de vue de l'exhaustivité.

Dans le tableau ci-dessous, on trouve à gauche les formules de calcul pour la (vraie) valeur paramétrique de la population, qui reste inconnue, et à droit on voit la valeur estimée (à partir d'échantillons) de la population correspondante. Rappelons que le concept derrière la variance d'erreur dans le tableau a déjà été expliqué (Leçon 1, Concepts de base de l'échantillonnage, Concepts et terminologie importants, Quelle est la différence entre la variance de la population et la variance d'erreur)..

Statistique	Calcul paramétrique	Estimateur à partir d'un échantillon
Moyenne	$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$	$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}$	$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$
Écart-type	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}}$	$S_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$
Coefficient de variation (CV)	$CV = \frac{\sigma}{\mu}$	$CV = \frac{S}{\bar{y}}$
Erreur-type	$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$	$S_{\bar{y}} = \frac{S_y}{\sqrt{n}}$
Variance d'erreur	Erreur-type ^{2}	

Les estimateurs donnés ici sont ceux utilisés pour les EAS. Dans des prochaines leçons, nous apprendrons qu'ils deviennent légèrement plus complexes lorsque l'on aborde d'autres plans d'échantillonnage.

Ici, tout comme dans les leçons suivantes, on considère un échantillonnage sans remplacement et un échantillonnage d'une population infinie – et l'on ignorera par conséquent ce que l'on appelle la correction pour population finie (fpc). Pour plus de détail sur la fpc, consultez ***le wiki de l'Université de Göttingen*** (en anglais), ou tout manuel d'échantillonnage statistique.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Les experts doivent dériver des inférences et tirer des conclusions concernant la situation actuelle et les changements des variables cibles en réalisant des observations sur des sous-ensembles ou «échantillons» relativement petits de la surface forestière totale.
- Lorsque l'échantillonnage suit les règles des statistiques, on parle d'«échantillonnage statistique». La validité statistique est l'une des principales caractéristiques de l'échantillonnage statistique, tel qu'appliqué au suivi des forêts.
- Un «estimateur» en échantillonnage statistique se réfère à la formule de calcul utilisée pour produire une estimation.

Leçon 2: Éléments de conception d'une étude par échantillonnage

Introduction de la leçon

Cette leçon introduit les éléments de conception de base des études par échantillonnage qui sont pertinents pour les IFN, et les concepts à prendre en compte pour préparer une stratégie d'échantillonnage.

Elle montre également comment calculer la taille de l'échantillon associée.

Rappelons que suivre cette leçon ne fera pas de vous un expert des techniques décrites, mais améliorera votre appréhension des concepts généraux. Comme pour les autres leçons de ce cours, il s'agit uniquement d'une «amorce» pour les apprenants n'ayant pas de base solide en statistiques, indispensable pour une bonne compréhension de l'échantillonnage statistique.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire les trois éléments de conception technique d'une étude par échantillonnage.
2. Décrire un plan d'échantillonnage.
3. Identifier les types de plans d'échantillonnage.
4. Expliquer la logique et l'approche de la stratification.
5. Décrire un plan parcellaire/plan d'observation.
6. Résumer le concept de correction de pente.

Trois éléments de conception d'une étude par échantillonnage

La planification de tout étude par échantillonnage peut être divisée en trois éléments de conception technique de base qui fournissent un cadre aux projets d'échantillonnage. Rappelons que pour préparer une étude par échantillonnage, ces trois éléments de conception doivent être pris en compte de manière exhaustive.

Plan d'échantillonnage	Le plan d'échantillonnage répond à la question: «Comment les éléments de l'échantillonnage sont-ils sélectionnés?». En suivi des forêts, les points échantillons sélectionnés dans une aire d'inventaire sont théoriquement infiniment petits, et donc considérés adimensionnels. Ces points définissent la position des parcelles d'échantillonnage. 
Plan d'observation	Le plan d'observation, aussi appelé plan parcellaire ou plan de réponse, aborde la question: «Comment les observations sont-elles réalisées pour chaque élément de l'échantillonnage?». Le plan d'observation est défini par les règles qui guident la manière dont les arbres échantillons sont inclus dans la parcelle d'échantillonnage, en référence au point échantillon adimensionnel. 
Plan d'estimation	Le plan d'estimation répond à la question: «Comment les estimations sont-elles calculées, et quels estimateurs statistiques sont-ils utilisés?». C'est l'ensemble des estimateurs, ou formules, à utiliser pour le plan d'échantillonnage et le plan parcellaire donnés. Dans le plan d'échantillonnage et le plan parcellaire, on peut librement choisir les plans «optimaux» ou qui remplissent le mieux les objectifs.  Cependant, on ne peut pas choisir librement les estimateurs. En effet, ceux-ci doivent correspondre au plan d'échantillonnage et au plan

parcellaire sélectionnés. Généralement, il existe peu d'estimateurs de la sorte.

Rappelons que dans cette leçon, nous nous concentrerons que les plans d'échantillonnage et les plans parcellaires typiques dans les IFN. Les plans d'estimation seront abordés dans la prochaine, et dernière, leçon de ce cours.

Détermination de la taille de l'échantillon

L'un des aspects définis dans le plan d'échantillonnage est le nombre d'éléments de l'échantillonnage (parcelles) qui doivent être observés. On parle aussi d'effectif de l'échantillon. Dans une perspective purement statistique, il y a deux principaux critères qui déterminent la taille de l'échantillon nécessaire pour une précision cible définie dans une situation d'inventaire donnée:

① La variabilité dans la population, autrement dit la variance de la population. Elle peut être estimée à partir d'une étude pilote, ou prise dans des inventaires préalables ou des inventaires dans des aires comparables. On se réfère ici à la population des parcelles d'échantillonnage et les variances de population seront différentes pour différents plans parcellaires pour la même surface forestière.

② La précision cible désirée, qui est une question de définition. Généralement, la précision est définie comme la moitié de l'amplitude de l'intervalle de confiance cible.

Que se passe-t-il s'il n'existe pas de connaissance préalable ou d'information d'inventaires antérieurs?

En l'absence de données d'inventaires préalables ou d'estimations de variabilité de la variable cible, une étude pilote peut aider à obtenir l'information pertinente. Puisque la variance estimée se réfère toujours au plan parcellaire spécifique utilisé, un nombre relativement petit de parcelles pourrait être distribué entre les différents types de forêt présents dans un pays. Cette étude pilote peut alors fournir des estimations de la variance de la population (même si elles seront probablement peu précises).

Il se peut aussi que des types de forêt similaires existent dans les pays voisins, pour lesquels des estimations de la variance de la population sont disponibles pour informer notre plan d'échantillonnage.

Lorsqu'aucune information n'est disponible, les statisticiens forestiers doivent alors s'appuyer sur de l'information alternative, souvent non fondée sur des plans probabilistes, comme une opinion experte ou un examen de la littérature.

Pour un échantillon aléatoire, la taille de l'échantillon est comme suit:

$$n = \frac{t^2 * S^2}{A^2} = \frac{t^2 * (CV\%)^2}{(e\%)^2}$$

où A désigne l'intervalle de confiance, en valeur absolue, que l'on cherche à atteindre dans les estimations (en pourcentage e% si exprimé relativement à la moyenne), t est la valeur correspondante de la loi t de Student, et S^2 (généralement pré-estimée à partir d'études pilotes ou d'information préalable) est la variance d'échantillon de la variable d'intérêt, comme le volume par ha. CV pour cent est le coefficient de variation dans cette information préalable, exprimé en pourcentage relativement à la moyenne. L'exercice suivant montre un exemple pratique pour calculer la taille de l'échantillon.

Exercice pratique

On cherche à calculer combien de parcelles seront nécessaires pour estimer le stock de carbone des forêts avec une précision de 10 % (se référant à un intervalle de confiance de 95 %). Plusieurs études ont montré des valeurs AGB autour de 100 t/ha avec un écart-type de 70 t/ha (CV %=70). Combien de parcelles doivent être mesurées si l'on adopte un échantillonnage aléatoire simple?

Afin de calculer cela, on a besoin de la valeur correspondante de la loi t pour probabilité d'erreur de 5 % (ou 0,05, bilatéral). Cependant, pour déterminer la valeur t, on doit connaître la taille de l'échantillon – soit ce qui est recherché ici. Par conséquent, on doit d'abord adopter un effectif de l'échantillon et ensuite réaliser un calcul itératif. On peut commencer avec une valeur t de 2 dans la première itération – ce qui correspond à un effectif de l'échantillon de plus de 30 et obtenir: $2^2*70^2/10^2 = 196$.

En référençant le ***tableau T*** pour cet effectif de l'échantillon de n=196 (à partir de la première itération), on arrive à une valeur t de ~1,97 – et l'estimation ci-dessus peut de nouveau être calculée par $1{,}97^2*70^2/10^2 \sim 190$.

Remarquez que cette estimation de la taille de l'échantillon nécessaire est uniquement valide pour un EAS..

En réalité, cependant, les ressources sont limitées et seul un certain nombre de parcelles de terrain peuvent être observées. Dans ce cas, on cherche à atteindre le résultat le plus précis avec le budget donné. Comme nous l'avons déjà appris, augmenter la taille de l'échantillon améliorera la précision – on

voudra alors planifier le plus de parcelles possibles avec le plan parcellaire et les restrictions pratiques données.

Quelle est ma variable cible?

Un inventaire forestier peut uniquement être optimisé en fonction d'une variable cible singulière (pour laquelle la précision devra être maximisée avec les ressources données). La surface terrière des peuplements, qui est fortement corrélée au volume et à la biomasse, est fréquemment utilisée comme variable cible. Néanmoins, la prise en compte d'objectifs exige des compromis dans les plans d'échantillonnage et les plans parcellaires, et il est possible que la taille de l'échantillon qui optimise la précision de l'estimation de la surface terrière ne soit pas optimale pour d'autres variables.

Plan d'échantillonnage

Jusqu'ici, nous avons vu les concepts de base de l'échantillonnage et considéré les trois éléments du plan d'échantillonnage. Intéressons-nous maintenant à certaines options du plan d'échantillonnage.

Le plan d'échantillonnage définit la procédure de sélection des éléments de l'échantillonnage, autrement dit, comment ces éléments de l'échantillonnage sont sélectionnés, et combien (effectif de l'échantillon). Le résultat de cette procédure de sélection est une liste de toutes les coordonnées des positions d'échantillonnage.

Nous nous limiterons ici à aborder exclusivement certains plans d'échantillonnage typiques utilisés dans le contexte des IFN. Rappelons que nous avons déjà abordé l'EAS (voir Leçon 1, Estimation par échantillonnage aléatoire simple) comme un plan principalement théorique peu utilisé en pratique dans les IFN, mais utile pour établir une référence simple avec laquelle comparer les options suivantes.

Échantillonnage systématique – le plan d'échantillonnage le plus commun dans les IFN

Utiliser une grille systématique des positions d'échantillonnage est le plan d'échantillonnage normalisé dans les IFN. Un tel échantillon systématique a l'avantage que la surface forestière est uniformément couverte par les positions d'échantillonnage, et il assure une distance minimum entre toutes les positions. Il mène à une «allocation proportionnelle» des positions d'échantillonnage entre les types de forêt trouvés.

Et, comme il couvre uniformément l'aire d'intérêt entière, on peut s'attendre à ce qu'une telle grille

d'échantillonnage donne un échantillon «représentatif» de la population.

Des considérations théoriques et de nombreuses études de simulation ont montré que l'échantillonnage systématique apporte virtuellement toujours une plus grande précision que l'EAS, avec le même nombre de points d'observation.

Cela peut s'expliquer par le fait que l'échantillonnage systématique couvre uniformément l'ensemble de la population et toutes les conditions sont ainsi couvertes; une autre raison est que dans l'échantillonnage systématique, les points échantillons sont toujours à une distance définie et ne peuvent pas être très proches: dans les forêts, et dans beaucoup d'autres populations naturelles, les parcelles qui sont trop proches sont généralement plus auto-corrélées que les parcelles distantes, et cela est inefficace.

Note

La **taille de l'échantillon** se réfère au **nombre d'éléments de l'échantillonnage sélectionnés indépendamment**, où indépendamment signifie sélectionnés par randomisation. Puisque toutes les positions d'échantillonnage dans une grille systématique sont fixes une fois que le point de départ et l'orientation de la grille ont été sélectionnés, il n'existe qu'un échantillon systématique fondé sur une randomisation (effectif de l'échantillon = 1). Cependant, à partir d'une seule observation indépendante, on ne peut pas dériver d'estimation de la variance, et par conséquent, pas non plus d'estimation de la précision!

Si l'on reprend l'estimateur de variance pour l'EAS, si la taille de l'échantillon est n=1, alors le dénominateur n-1 sera égal à zéro et, par conséquent, la variance de la variable d'intérêt S2 y n'est pas définie:

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

L'estimateur de l'EAS est fréquemment utilisé pour calculer la variance d'erreur d'un échantillon systématique. On sait que cette variance d'erreur va surestimer la vraie variance d'erreur, et l'on dit alors que l'estimateur de l'EAS est ici un **estimateur conservateur.**

Cela signifie que la vraie précision est supérieure à celle que l'on estime avec l'estimateur de l'EAS, mais on ne peut pas dire dans quelle mesure elle est plus précise. Cette sous-estimation de la précision affectera aussi d'autres estimations, comme la taille de l'échantillon nécessaire.

Stratification

Nous avons déjà appris dans les leçons précédentes que notre but est de restreindre la distribution des observations le plus possible car cela augmentera la précision des estimations.

Que peut-on faire d'autres pour rendre la population «plus homogène» afin d'améliorer la précision?

La stratification vise à sous-diviser la population en sous-populations plus homogènes. On appelle ces sous-populations des strates. Un échantillon indépendant est prélevé dans chaque strate. Lorsque l'on utilise l'échantillonnage aléatoire simple dans chaque strate, on parle de plan d'échantillonnage aléatoire stratifié. Autrement dit: on n'introduit pas ici de nouveau plan d'échantillonnage, mais l'on applique l'EAS indépendamment dans chaque strate; ce qui est nouveau a trait à la combinaison finale des estimations par strate pour arriver à un total composé de toutes les strates.

Pour avoir plus de précision dans ce plan, la stratification doit être «homogène au sein des strates et hétérogène entre les strates».

Observons le diagramme: la surface forestière totale est sous-divisée en deux strates différentes (claire et foncée), et l'on suppose que chacune est plus homogène que la surface totale et qu'elles diffèrent clairement dans leurs valeurs moyennes. Dan le plan d'échantillonnage, elles sont traitées comme des sous-populations indépendantes et des grilles systématiques différentes sont utilisées.

Il existe de nombreuses manières de sous-diviser une population en sous-populations: les critères de stratification sont, par exemple, les types de forêt, ou les régions de croissance avec des conditions physiques homogènes. Parfois, des limites administratives sont aussi utilisées, bien que cela ne donne pas nécessairement des sous-populations plus homogène ou une meilleure précision.

Cela peut cependant faciliter la mise en œuvre de l'inventaire ou assurer que des estimations plus précises par unité administrative peuvent être fournies..

Calcul de la taille de l'échantillon et allocation des échantillons aux strates

Pour déterminer la taille de l'échantillon dans un échantillonnage stratifié, deux questions se posent:

- combien d'échantillons utiliser au total; et
- comment distribuer/allouer les échantillons aux strates.

La taille de l'échantillon nécessaire ne dépend pas toujours de l'erreur permise avec une probabilité d'erreur donnée ou de la variabilité au sein d'une population; dans la stratification, on traite un certain nombre de sous-populations, et il faut considérer que les variances de sous-population diffèrent entre les strates.

Comme les strates ont généralement des tailles différentes, ces variances de sous-population peuvent être pondérées en calculant la taille de l'échantillon totale. S'il y a un nombre de strates L annoté avec l'indice h, et chaque strate a la taille (par exemple en termes de surface) Nh, le poids de chaque strate est donné par Nh/N.

Mais la conception d'un inventaire peut aussi impliquer de gérer des contraintes en termes de coûts d'inventaire. Donc si l'on veut allouer des parcelles d'échantillonnage pour minimiser la variabilité, on peut aussi vouloir penser au coût total engagé dans l'inventaire, où Ch est le coût par unité d'échantillonnage dans la strate h. Alors, la taille de l'échantillon totale peut être calculée par:

$$n = \frac{t^2 \sum \frac{N_h^2 S_h^2}{C_h}}{N^2 A^2}$$

où A est l'erreur permise, exprimée comme la moitié de l'amplitude de l'intervalle de confiance cible. L'erreur permise est une question de définition. Comme dans l'estimation de la taille de l'échantillon dans un EAS, vue plus haut, S et A peuvent être substitués par des expressions relatives: CV(%) et e(%).

Après avoir calculé la taille de l'échantillon totale, ces échantillons peuvent être alloués aux différentes strates. Pour cela; il faut considérer trois caractéristiques des strates, individuellement ou dans leur ensemble:

1. **La taille de la strate**: plus une strate sera grande, plus on lui allouera d'échantillons.

2. **La variabilité dans la strate**: plus une strate sera variable, plus on lui allouera d'échantillons.
3. **Le coût par unité d'échantillonnage**: plus le coût sera élevé, moins on allouera d'échantillons.

Allocation proportionnelle	Allocation de Neyman	Allocation optimale avec minimisation des coûts
Allocation d'échantillons en fonction de la seule taille des strates	Prise en compte de la taille des strates et de la variabilité au sein des strates pour l'allocation	Dans cette option, les implications de coût (c) sont aussi incluses, en plus de la taille des strates et de la variabilité au sein des strates
$n_h = n \frac{N_h}{N}$	$n_h = n \frac{N_h S_h^2}{\sum_{h=1}^{L} N_h S_h^2}$	$n_h = n \frac{\frac{N_h S_h^2}{\sqrt{C_h}}}{\sum_{h=1}^{L} \frac{N_h S_h^2}{\sqrt{C_h}}}$

Note

Rappelons que t**oute technique d'échantillonnage peut être appliquée par strate.** Il peut aussi y avoir différentes techniques d'échantillonnage utilisées dans les différentes strates. Il est important que, pour chaque strate, les estimations ponctuelles et par intervalle des variables cibles puissent être produites, dans le meilleur des cas, de manière non biaisée.

En fait, **la principale caractéristique de l'échantillonnage stratifié est qu'il consiste en plusieurs études par échantillonnage mises en œuvre indépendamment**. La seule chose nouvelle est qu'il faut trouver comment combiner finalement les estimations qui proviennent de L strates différentes de manières à générer des estimations pour la population entière.

Post-stratification

On peut aussi stratifier l'inventaire après la mise en œuvre d'un échantillonnage non stratifié (par ex. en dérivant des estimations distinctes pour les différents types de forêt). On appelle cela la post-stratification, et elle peut être considérée comme une sorte de groupement de données à des fins d'analyse.

Néanmoins, cette analyse post-stratifiée doit être menée avec prudence, car l'estimation ne suit plus strictement le plan d'échantillonnage. Par exemple: le groupement de données pour l'analyse ne doit pas être fait avec la variable cible, en formant par exemple trois groupes d'amplitude égale (post-strates) de valeurs faibles, moyennes et hautes; cette approche serait entièrement fausse, même si elle mènerait à des valeurs de précision élevées (mais fausses)!

Avant de réaliser une analyse post-stratifiée avec des estimations de la précision de l'estimation, il faut consulter un expert en échantillonnage pour éviter des erreurs inutiles et des inférences et conclusions trompeuses: tous les estimateurs qui sont recommandés dans les manuels pour les analyses post-stratifiées sont associés à des hypothèses qui doivent être observées.

Échantillonnage à deux phases ou double

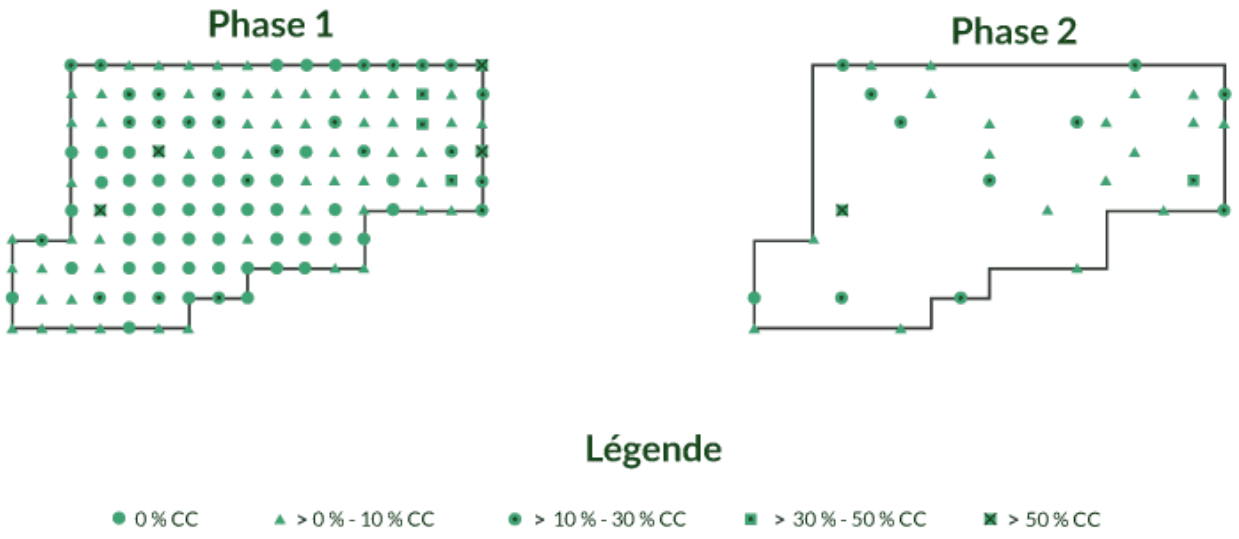
Dans l'échantillonnage double, une nouvelle caractéristique est introduite: l'utilisation de **variables auxiliaires**, également appelées **variables subordonnées ou co-variables**. Afin d'améliorer la précision de l'estimation de la variable cible, il est essentiel de comprendre la corrélation entre la cible et les variables auxiliaires. L'idée est de prélever un échantillon relativement grand – mais à bas coût – dans une première phase, pour obtenir de l'information sur ces variables auxiliaires, comme par exemple avec la télédétection.

Puis, dans une seconde phase, un échantillon plus petit est sélectionné où la variable cible et la variable auxiliaire sont observées à la fois. Ceci implique généralement un coût beaucoup plus élevé par parcelle – voyons un exemple pour mieux comprendre.

Lorsque l'on estime la biomasse de la forêt, il est possible d'utiliser une première phase d'estimation de la variable auxiliaire grâce à l'imagerie de télédétection et de déterminer un indice de végétation autour de nombreux points échantillons. Ceci est rapide et peu cher et peut aussi être réalisé automatiquement.

Puis, dans la seconde phase, un effectif d'échantillon de parcelles d'échantillonnage beaucoup plus petit est installé sur le terrain où des mesures sont prises et la biomasse des parcelles est estimée. Ceci est beaucoup plus cher qu'un échantillon dans la première phase. Pour toutes les parcelles de terrain, non seulement la variable cible (la biomasse, dans cet exemple) est déterminée, mais la variable auxiliaire est également observée (l'indice de végétation, dans cet exemple) à partir de données de télédétection.

Les paires de données de la seconde phase, avec les variables cible et auxiliaire, sont alors utilisées pour établir un modèle entre la variable cible et la variable auxiliaire, à partir duquel les estimations peuvent être produites. Les modèles les plus communs utilisés ici sont le ratio simple entre les deux variables ou un modèle de régression. Ceci mènerait alors à un échantillonnage double avec l'estimateur par ratio, et à un échantillonnage double avec l'estimateur par régression, respectivement.



Il peut apparaître désormais clairement que plus la corrélation positive entre les deux variables sera élevée, plus l'estimateur sera efficace en termes de précision. Autrement dit, pour une variable auxiliaire efficace, il faut rechercher une variable très fortement positivement corrélée à la variable cible. Finalement, il s'agit bien entendu aussi d'une considération des coûts, car l'introduction d'une première phase augmente le coût de l'inventaire.

Dans la prochaine leçon sur le plan d'estimation, nous verrons comment cela peut améliorer la précision de l'estimation de la variable cible.

L'échantillonnage double est une manière très efficace de capter une variable auxiliaire (peu chère à observer) pour améliorer la précision de l'estimation d'une variable cible (plus chère à observer).

Échantillonnage double pour la stratification

L'échantillonnage double est aussi pertinent dans le contexte de la stratification. Il existe des cas d'inventaires où l'on sait ou l'on suppose que la stratification pourra améliorer la précision, mais parfois, il n'est pas possible de dessiner clairement les limites des strates (par ex. à travers l'imagerie de télédétection), car elles sont «floues» ou constituent plus des transitions continues que des lignes strictes. En outre, une telle délimitation prend du temps et nécessite une connaissance préalable solide!

Jusqu'ici, pour l'échantillonnage stratifié, on a supposé que les strates étaient définies avant l'échantillonnage, où l'on peut donc parler de **pré-stratification**. Ce faisant, on suppose que cette définition préalable des strates est libre d'erreur; autrement dit: la taille des strates et leur poids dans les estimateurs ne sont pas considérés comme une source d'erreur.

Dans **l'échantillonnage double pour la stratification (EDS) ou échantillonnage à deux phases pour la stratification**, les strates n'ont pas besoin d'être définies avant l'échantillonnage, mais sont définies pendant le processus d'échantillonnage où la taille des strates est estimée.

Les deux phases dans l'EDS sont comme suit: un **échantillon relativement grand est sélectionné dans la première phase** (fréquemment à partir d'imagerie de télédétection, car cela est peu cher), et il est déterminé à quelle strate appartient chaque point échantillon. Autrement dit: la variable auxiliaire qui est observée dans la première phase est la strate; dans les IFN il peut s'agir du type de forêt.

Dans la seconde phase, un sous-échantillon stratifié des parcelles de la première phase est sélectionné, et ces parcelles sont visitées pour observer la variable cible – relativement chère –, ce qui est fréquemment mené sur le terrain. L'allocation de la taille de l'échantillon totale aux strates peut être effectuée selon la même stratégie qu'avec la pré-stratification: uniforme, proportionnelle à la taille, proportionnelle à la taille et à la variabilité, ou encore proportionnelle à la taille, à la variabilité et au coût. La décision de cette allocation devra être menée à partir de l'information disponible sur la variabilité attendue et le coût par parcelle dans les strates qui ont été distinguées.

Avec les mêmes tailles d'échantillon dans l'échantillonnage en seconde phase dans la pré-stratification

normale, l'EDS sera moins précis que la pré-stratification. La raison pour cela est que dans l'EDS, la taille des strates est estimée à partir de l'échantillon de la première phase, et une telle estimation de la taille entraîne une erreur d'échantillonnage qui se propage à l'erreur totale. On peut aussi le voir avec les estimateurs pour l'EDS, qui ne sont pas présentés ici mais peuvent être trouvés dans les manuels d'échantillonnage..

Le saviez-vous?

Peut-on aussi utiliser une classification par télédétection pour séparer les strates?

Oui, c'est possible et c'est souvent le cas. Imaginons, par exemple, une classification à partir de la télédétection de différents types de forêt, pour laquelle on attend des différences de la biomasse forestière. Néanmoins, comme pour l'interprétation visuelle mentionnée plus haut, toute classification comportera des erreurs. Puisque que les estimations des aires des strates contiennent des erreurs, il faut tenir compte de cette source d'incertitude supplémentaire dans l'estimateur!

Astuces rapides!

N'«inventez» jamais un nouveau plan d'échantillonnage ou plan parcellaire en ignorant la question de la dérivation d'un estimateur statistique non biaisé! L'exactitude d'un estimateur dépend d'une réflexion prudente de la procédure de sélection et d'inclusion. On peut facilement tomber dans des pièges statistiques insolubles en faisant juste de petits changements dans le plan d'échantillonnage ou le plan parcellaire. Par exemple, une règle simple comme «étendre la même parcelle si une certaine condition est remplie» peut donner lieu à des problèmes statistiques inattendus (les probabilités d'inclusion des arbres en résultant ne peuvent pas être calculées facilement)! D'autres règles, comme l'inclusion complète de parcelles qui chevauchent les limites de la forêts, sont une violation de la définition de la population. Elles sont simplement fausses et peuvent donner lieu à des estimations biaisées.

Plan parcellaire ou plan d'observation

Nous allons maintenant nous intéresser au plan d'observation.

Le plan d'échantillonnage détermine comment les points échantillons sont sélectionnés, tant que le plan parcellaire aborde la manière dont les arbres à échantillonner sont choisis autour du point sélectionné. La question est: quels objets (ex. Arbres) doivent être inclus dans chaque position d'échantillonnage autour du point échantillon?

Comme dans pratiquement toutes les étapes de conception/planification pour un IFN, si l'on cherche à optimiser ou adapter le plan parcellaire aux conditions spécifiques de la forêt, on doit penser avec soin comment allouer les ressources limitées (temps, budget et personnel) de la manière la plus efficace. L'efficacité peut être comprise comme la relation entre les coûts et la précision des estimations en résultant. Si les ressources ne sont pas suffisamment prises en compte, cela peut compromettre la durabilité d'un IFN permanent.

Saisir la variabilité comme un objectif majeur de la conception du plan parcellaire

D'un point de vue purement statistique dans l'optimisation d'un plan parcellaire, on cherche à saisir **un maximum de variabilité à l'intérieur de chaque parcelle.** La logique derrière cela est que l'on réduit ainsi la variabilité entre les parcelles. Et cela se traduit par une distribution étroite des observations des parcelles autour de la moyenne estimée (*voir la leçon 1 de ce cours*), ce qui signifie une plus haute précision de l'estimation.

L'auto-corrélation spatiale est un concept pertinent dans ce contexte, que l'on observe aussi dans les populations des forêts: cela signifie que les objets – ici, les parcelles – qui sont proches tendent à donner des observations plus corrélées.

Une corrélation élevée signifie qu'en connaissant la valeur du premier objet, on peut prédire assez bien la valeur du second objet à une distance spatiale donnée. Si c'est le cas, la mesure de la seconde observation est peu efficace car elle apporte peu d'information supplémentaire; elle peut même constituer une perte d'argent.

Tenir compte de l'importance de l'auto-corrélation spatiale dans la conception d'un plan d'inventaire permet certaines conclusions concernant le plan parcellaire:

- Il est bon d'avoir une certaine distance entre les parcelles d'échantillonnage. Des parcelles d'échantillonnage qui sont spatialement proches ne sont pas efficaces.
- Il est bon d'avoir un plan parcellaire qui couvre une grande aire de sorte que les observations à l'intérieur des parcelles montre une moindre auto-corrélation:
    a) ainsi, avec la même aire donnée, les parcelles en bande allongées sont statistiquement plus efficaces que les parcelles rondes ou carrées; et
    b) une autre option pour augmenter l'efficacité d'une aire de parcelles donnée consiste à sous-diviser les parcelles en sous-parcelles spatialement séparées à une certaine distance les unes des autres: c'est ce que l'on appelle les «parcelles en cluster».

Avec ces deux options, rappelons que non seulement les considérations statistiques, mais aussi de coût, sont importantes. Le coût par parcelle sera supérieur pour des parcelles en bande allongées ou des parcelles en cluster que celui de parcelles compactes dans la même aire: ainsi, en pratique, ces considérations d'optimisation doivent toujours être pondérées avec les critères statistiques et de coût.

Typiquement, cette corrélation spatiale diminue après 50-200 m (selon le type de forêt et la gestion)..

Aire fixe et parcelles d'échantillonnage imbriquées

Les parcelles à aire fixe constituent le plan parcellaire le plus simple. La forme et la taille de ces parcelles à aire fixe peuvent différer selon l'objectif spécifique de l'inventaire et les conditions des forêts. En général, les parcelles circulaires sont plus communes que les rectangulaires dans le suivi des forêts, alors que les formes carrées sont plus fréquentes dans les enquêtes écologiques – et le terme «**quadrat**» est parfois utilisé dans les études écologiques au lieu de parcelle ou lot.

D'un point de vue théorique, toute forme de parcelle est admissible; cependant, il est crucial de considérer avec soin l'objectif de l'inventaire et les conditions des forêts pour sélectionner le plan parcellaire approprié, et pondérer les considérations de coût et pratiques avec le besoin de relever des données exactes.

Note

Le facteur d'expansion des parcelles (ou des arbres)

De nombreuses variables d'intérêt sont liées à l'aire, comme le «nombre d'arbres par hectare». Ceci signifie par exemple que si l'on double l'aire d'une parcelle, on s'attend aussi à ce que le nombre d'arbres trouvés double en moyenne.

Afin d'étendre ou de mettre à l'échelle l'observation pour l'unité de rapport typique d'un hectare, ces observations par parcelle liées à l'aire doivent être multipliées par un facteur d'expansion résultant de la relation 1 ha/aire de parcelle.

Les variables liées à l'aire sont généralement celle associées à des mesures quantitatives directes, comme le volume, la biomasse, le nombre d'arbres ou la densité de régénération.

Les arbres de différentes classes de diamètre apparaissent normalement avec des densités différentes (dans les forêts naturelles, par exemple, il y a beaucoup plus de petits arbres que de grands arbres). Les grands arbres sont porteurs d'une grande part de la biomasse forestière, mais ils sont moins nombreux. Si l'on utilise alors une parcelle relativement grande pour s'assurer d'avoir en moyenne de grands arbres dans l'aire de la parcelle, on devra mesure un grand nombre de petits arbres. Autrement dit: une aire de parcelle unique est ainsi généralement peu efficace.

Une solution courante est d'utiliser ici ce que l'on appelle un plan parcellaire imbriquées, où des sous-parcelles de différentes tailles sont imbriquées, de manière à observer les arbres de différentes classes de taille dans différentes aires des sous-parcelles. Il est important ici de s'en tenir à une terminologie stricte pour éviter les confusions: l'ensemble (la combinaison de toutes les sous-parcelles) est la parcelle, tandis que les diverses formes imbriquées de différentes tailles constituent les sous-parcelles.

Note

Prise en compte des probabilités inégales

L'inclusion de la probabilité dans un inventaire forestier se réfère à la probabilité qu'un arbre soit inclus dans un échantillon. Puisqu'effectivement, les unités d'échantillonnage sont des parcelles, fondées sur des aires, cette probabilité est en effet l'inverse du facteur d'expansion.

Ainsi, un plan avec des sous-parcelles donnera lieu à des probabilités d'inclusion inégales, et cela doit être reflété dans le plan d'estimation: le facteur d'expansion de la parcelle sera plus élevé pour les parcelles plus petites, où les arbres plus petits sont observés!

Puisque les sous-parcelles ont des tailles et aires différentes, les arbres auront des facteurs d'expansion différents selon la sous-parcelle où ils ont été enregistrés. On doit alors calculer le facteur d'expansion correct pour chaque arbre individuellement, en fonction de son dhp ou de son affiliation à une sous-parcelle et son aire correspondante. Les facteurs d'expansion normalisent alors toutes les aires en une base singulière par hectare.

La vidéo suivante explique comment établir et évaluer des parcelles d'échantillonnage imbriquées à aire fixe sur le terrain. La vidéo est en langue originale (anglais) '**How to assess (nested) fixed area forest inventory plots https://www.youtube.com/watch?v=IA-PflXW9_k&t=2s**

Correction de pente

L'aire à laquelle se réfèrent toutes les observations et les estimations est l'aire cartographiée, soit la projection horizontale du terrain sur le plan cartographié.

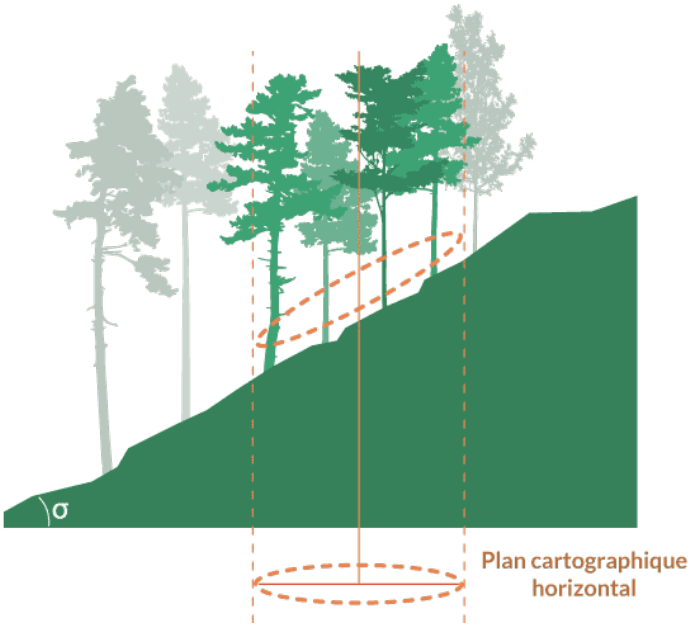
Lorsqu'il n'est pas possible de mesurer directement des distances horizontales pendant la mesure de la parcelle (en utilisant des instruments électroniques modernes), et les distances sont mesurées en suivant la pente, l'aire de projection horizontale de la parcelle est plus petite que l'aire de la parcelle attendue et les distances mesurées entre le centre de la parcelle et les arbres sont plus grandes que les distances horizontales projetées (excepté si l'on mesure exactement les lignes de contour).

Pour garantir une aire de parcelle égale sur le plan cartographique horizontal, ce qui est une exigence pour dériver des estimations non biaisées liées à l'aire, l'aire de parcelle oblique qui constitue la parcelle sur le terrain doit être agrandie en fonction de l'angle de la pente.

Si l'on utilise des parcelles à aire fixe circulaires, elles deviennent des ellipses une fois projetées sur la pente. Pour établir ces parcelles avec une pente, il y a essentiellement deux options::

1. Soit un télémètre électronique est utilisé, qui mesure directement la distance horizontale: alors, les arbres corrects dans la distance horizontale définie (rayon) sont automatiquement inclus. Une parcelle elliptique est établie sans qu'il soit nécessaire de la déterminer spécifiquement.
2. Soit – l'approche traditionnelle – on calcule l'aire (la plus grande) de l'ellipse projetée sur la pente et on établit un cercle sur la pente avec exactement cette aire. Pour cela, on doit corriger en fonction de la pente le rayon nominal de la parcelle sur le plan horizontal avec le facteur ci-dessous pour obtenir le rayon du cercle supérieur qui sera déterminé sur la pente.1

$$\sqrt{\cos \alpha}$$



Ressources vidéo

Comment évaluer les parcelles (imbriquées) a aire fixe d'un inventaire forestier

La vidéo est en langue originale (anglais).

Si cette correction de pente a été omise pendant l'établissement de la parcelle, les observations obtenues de cette parcelle peuvent être corrigés en aval (puisque les parcelles ont des tailles inégales dans la projection horizontale selon la pente). Puisque l'aire horizontale réelle est plus petite que celle attendue, le résultat devra être multiplié par le facteur de correction 1/cos α. Cependant, l'angle de la pente doit avoir été mesuré; sinon, aucune correction ne sera possible

Dans la plupart des IFN, la correction de pente est généralement considérée pour les angles de pente > 10 pour cent, soit l'une des conventions avec lesquelles les inventaires forestiers fonctionnent en pratique. En outre, en présence de pentes douces < 10 pour cent, les mesures de distance peuvent souvent être prises horizontalement par nivellement manuel.

Note

La correction de pente s'applique à tout plan parcellaire et doit toujours être prise en compte à l'avance, les corrections étant assez simples pour les parcelles à aire fixe circulaire. Les mêmes principes de correction de pente s'appliquent bien sûr aussi aux parcelles carrées et rectangulaires.

Mais pour ces deux formes de parcelle, la correction de pente est plus laborieuse: pour les parcelles carrées, les coins d'une aire de parcelle réelle plus grande doivent être marqués sur la pente afin que l'aire de la parcelle projetée corresponde à l'aire nominale. Pour les parcelles rectangulaires allongées, on marche généralement le long de la ligne centrale et on mesure les arbres à gauche et à droite à une distance définie : ici, les deux directions de la parcelle doivent faire l'objet d'une correction de pente, soit la longue ligne sur laquelle on marche, et les mesures à gauche et à droite.

Échantillonnage avec des parcelles en cluster

Dans les IFN, situer et se rendre sur les positions d'échantillonnage est un facteur de coût majeur, particulièrement lorsque le réseau routier est mauvais. La grille d'échantillonnage est généralement éparpillée et les distances entre les parcelles sont grandes. C'est pourquoi il convient d'évaluer le plus d'information possible une fois l'équipe déployée sur une parcelle. Ceci nécessite de grandes parcelles.

Cependant, nous avons appris – du fait de l'auto-corrélation spatiale – qu'il est bon de réaliser des observations à une certaine distance spatiale les unes des autres, de sorte qu'au lieu d'établir une grande parcelle par point échantillon, les IFN adopte souvent l'établissement de ce que l'on appelle des **parcelles en cluster**: les grandes parcelles individuelles sont sous-divisées en sous-parcelles, chacune déterminée à une certaine distance spatiale.

Le résultat est un ensemble de sous-parcelles ordonnées selon un certain modèle géométrique (par exemple les coins d'un carré, ou une forme de L). L'ensemble des sous-parcelles forme la parcelle et il est important de ne pas confondre parcelles et sous-parcelles. Les parcelles sont les éléments de l'échantillonnage centraux et le nombre de parcelles correspond au site des échantillons, pas au nombre de sous-parcelles.

La distribution et la distribution spatiales entre les sous-parcelles est déterminée de sorte que la parcelle en cluster puisse «saisir» plus de variabilité qu'une parcelle compacte singulière de la même surface. Pour planifier un plan parcellaire en cluster, il faut décider de certaines caractéristiques:

1. Le nombre de sous-parcelles par cluster.
2. La distance entre les sous-parcelles.
3. La taille et la forme des sous-parcelles.
4. L'ordonnancement spatial des sous-parcelles.

Considérations concernant la forme et la taille des (sous-)parcelles

Nous avons déjà conclu que chaque parcelle d'échantillonnage singulière doit saisir le plus de variabilité possible afin d'augmenter la précision générale de l'estimation. Si l'on disposait d'un nombre illimité de ressources, le même nombre de grandes parcelles serait toujours meilleur que le même nombre de

parcelles plus petites.

Mais, avec des ressources limitées, on doit décider si l'on utilise un plus grand nombre de petites parcelles ou un nombre moindre de grandes parcelles. Augmenter la taille des parcelles implique d'accroître les coûts marginaux de la précision pour chaque arbre supplémentaire mesuré, tandis qu'augmenter le nombre de parcelles implique des gains en précision de plus en plus faibles en échange d'un temps supplémentaire de marche entre les parcelles. Cependant, à partir d'une certaine taille de parcelles, l'effet sur la précision d'une taille de l'échantillon plus grande sera plus important que celui de l'augmentation de la taille des parcelles!

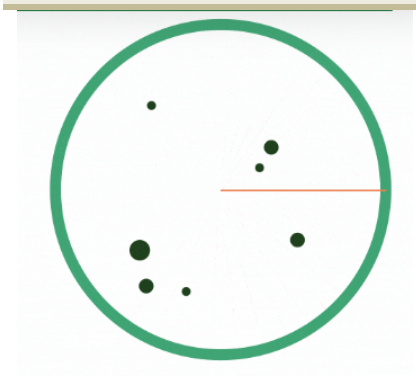
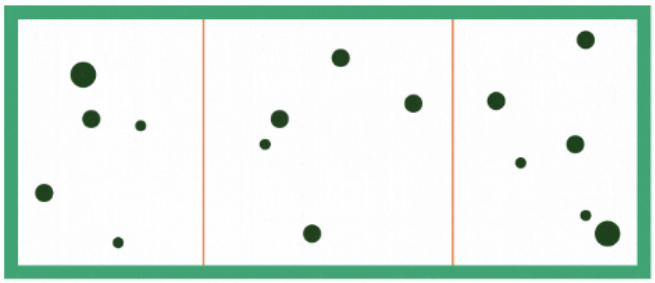
Le saviez-vous?

Taille de parcelle et efficacité statistique

Le gain d'information marginal que l'on peut attendre de la mesure d'un arbre supplémentaire par parcelle diminue à chaque nouvel arbre. Imaginons que l'on a déjà mesuré 99 arbres sur une parcelle d'échantillonnage, peut-on s'attendre à apprendre quelque chose de la mesure du 100e arbre? Probablement pas, parce qu'il apportera uniquement une information redondante.

Par ailleurs, les coûts d'évaluation de la parcelle augmenteront de manière linéaire avec chaque unité supplémentaire de l'aire observée (ou chaque hausse du nombre d'arbres). **Mais où doit-on placer la limite?** L'expérience et les études empiriques suggèrent que l'évaluation de plus de 15-20 arbres par (sous-)parcelle cesse d'être efficace. Il est donc préférable d'investir plutôt les ressources dans l'augmentation de la taille de l'échantillon (mieux vaut plus de petites parcelles que moins de grandes parcelles)!

Divers arguments pratiques et statistiques déterminent la forme des parcelles (ou des sous-parcelles), et les traditions et les normes courantes dans les différentes régions du monde doivent aussi être prises en compte. L'orientation générale qui suit s'applique aux mêmes aires de parcelles/sous-parcelle de différentes formes:

Les parcelles d'échantillonnage circulaires sont:	**Parcelles rectangulaires longues:**
	
• faciles à mettre en œuvre – pratiquement tout peut être mesuré à partir du centre; • relativement faciles concernant la correction de pente; mais • très compactes et susceptibles de saisir moins de variabilité.	• exigent plus de travail pour le marquage des parcelles (par ex. marquer le transect central avec une bande pour le suivre); • ont, en moyenne, plus d'arbres de bordure à vérifier; • intersectent, en moyenne, plus souvent avec les limites entre les types de forêt et nécessitent une plus grande prise en compte des corrections de bordure; • demandent plus de temps consacré à la correction de pente; • seront susceptibles de saisir une plus grande variabilité; et • sont préférables quand la visibilité est faible (sous-étage trop dense), car seules des courtes distances à gauche et à droite de la ligne centrale sont observées.

Concernant le nombre de sous-parcelles par cluster

Le regroupement de sous-parcelles en une observation conjointe sera toujours moins efficace que la sélection du même nombre de sous-parcelles comme parcelles sélectionnées indépendamment dans l'ensemble de la région inventoriée. L'échantillonnage avec des parcelles en cluster est un compromis utilisé pour réduire les coûts de transport et observer des aires de parcelles plus grandes à chaque position d'échantillonnage singulière, tout en réduisant la redondance causée par l'auto-corrélation spatiale en distribuant spatialement les sous-parcelles.

Par conséquent, les mêmes arguments que pour planification de la conception de parcelles singulières sont valables: augmenter l'aire observée signifie augmenter les coûts, mais l'erreur-type sera réduite jusqu'à une certaine limite, en-deçà de laquelle il n'y a proprement plus de réduction possible.

Ainsi, investir plus de temps et d'effort dans une parcelle singulière n'a, à partir d'un certain point, plus d'effet significatif sur la précision. Généralement, il n'y pas d'effet important sur la précision après un nombre de 3-5 sous-parcelles (selon la variabilité spatiale), et la mesure d'autres parcelles par cluster devient inefficace.

Rappel à la réalité

La viabilité comme ligne directrice

On peut dériver un grand nombre de considérations statistiques de la taille de l'échantillon et de la taille des parcelles, mais finalement, l'argument qui prévaut est la viabilité. Dans la plupart des cas, on est obligé de considérer les ressources disponibles et d'en tirer le meilleur parti.

À des fins de planification, il est souhaitable que, en moyenne, une parcelle en cluster entière puisse être mesurée par une seule équipe de terrain en une journée. Ceci affectera le nombre de sous-parcelles viables et leur effectif, en considérant des sous-parcelles relativement petites (dans de nombreux inventaires forestiers, les considérations statistiques donnent des tailles de parcelle d'environ 15-20 arbres en moyenne).

Il arrive fréquemment dans les IFN que plus de temps soit nécessaire pour atteindre le point échantillon et marcher de sous-parcelle en sous-parcelle. On peut considérer ces temps de marche

comme inefficaces en termes de mesures de nos variables cibles: souvent, la plus grande partie du temps sur le terrain est passée à marcher inefficacement. On peut alors facilement imaginer que plus de 4-5 sous-parcelles deviendront dans de nombreux cas un défi en termes de temps.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- La planification de toute étude par échantillonnage doit être divisée en trois éléments de conception technique de base – le plan d'échantillonnage, le plan d'observation/parcellaire, et le plan d'estimation.
- L'un des aspects définis dans le plan d'échantillonnage est le nombre d'éléments de l'échantillonnage (parcelles) qui doivent être observés.
- Un inventaire forestier doit généralement être optimisé en fonction d'une variable cible singulière (pour laquelle la précision devra être maximisée avec les ressources données). La surface terrière des peuplements, qui est fortement corrélée au volume et à la biomasse, est fréquemment utilisée comme variable cible.
- Dans les inventaires forestiers, le plan d'échantillonnage définit comment les échantillons sont sélectionnées parmi la population, et quelle sera la taille de l'échantillon.
- La stratification se rapporte à la «sous-division» de la population totale (surface forestière) en sous-populations plus homogènes que l'on appelle «strates».
- Le plan parcellaire définit ce qui est fait à chaque point échantillon; il définit aussi les règles d'inclusion des arbres échantillonnés qui seront observés..

Leçon 3: Conception de l'estimation

Introduction de la leçon

Dans cette leçon, nous allons nous pencher sur le plan d'estimation, qui consiste en méthodes et formules appliquées pour dériver des estimations non biaisées à partir de données relevées dans un plan d'échantillonnage et un plan parcellaire.

Objectifs

A la fin de leçon, vous serez en mesure de:

1. Décrire les estimateurs de base pour les approches d'échantillonnage communes.
2. Expliquer l'importance de l'application de l'estimateur correct.

Plan d'estimation

Commençons cette leçon en regardant des plans d'estimation typiques. Certaines de ces alternatives sont spécifiques au plan d'échantillonnage utilisé, et d'autres peuvent être appliquées à plusieurs plans d'échantillonnage différents.

Dans certains cas, on aura aussi la liberté d'appliquer différents estimateurs aux données relevées selon un plan d'échantillonnage donné. Par exemple, on peut inclure des données auxiliaires avec un estimateur par ratio (abordé dans les sections finales de cette leçon). On peut aussi dériver une estimation sans considérer de variable auxiliaire si cela n'aide pas à produire une estimation plus précise.

C'est à l'analyste de données de décider quel estimateur utiliser. Si des estimations alternatives multiples peuvent être produites, le choix est généralement le plan d'estimation qui donne la plus grande précision (ce qui équivaut à «l'erreur-type moindre des estimations»).

Inférence fondée sur un calcul, assistée par modèle et fondée sur un modèle

Dans la leçon 1, nous avons vu le terme «inférence». Parfois, les termes inférence et estimation sont utilisés de manière interchangeable, car chaque estimation signifie que l'on établit des inférences concernant les vraies valeurs de la population. Certains experts en inventaire préfèrent donc parler en général d'inférence quand ils se réfèrent à l'estimation, car l'inférence implique plus qu'une simple estimation: elle exprime aussi l'objectif de l'estimation.

Voyons maintenant trois paradigmes inférentiels: l'inférence fondée sur un calcul, fondée sur un modèle et assistée par modèle.

Inférence fondée sur un calcul	Inférence fondée sur un modèle	Inférence assistée par modèle	
On n'émet aucune hypothèse sur la structure (spatiale) de la population. On considère cette structure comme inconnue, et on cherche à estimer des caractéristiques de cette population fixe. L'absence de biais est garantie exclusivement par le plan d'échantillonnage et le plan parcellaire, soit par la randomisation.	La population est vue comme une réalisation d'un processus stochastique, et les hypothèses sur les processus sous-jacents ou le modèle peuvent être considérées pendant l'estimation. L'hypothèse est que l'on regarde uniquement l'une des nombreuses populations possibles (qui font une superpopulation). Puisqu'aucun modèle ne peut décrire parfaitement cette population, l'incertitude persistera même après un recensement complet, et provient de la «qualité» du modèle utilisé – et non pas du plans d'échantillonnage	Un modèle est utilisé pour appuyer une estimation fondée sur le plan, quelque part entre l'inférence fondée sur un calcul et l'inférence fondée sur un modèle. Cela signifie que même si le modèle n'était pas bien spécifié, cela n'introduira pas de biais, mais affectera la précision de l'estimation. Les estimateurs par ratio et par régression qui utilisent des modèles simples pendant l'estimation en établissant une relation entre une variable auxiliaire et la variable cible en sont des exemples.	**Hypothèses sur la population**
	La validité de l'estimation dépendra entièrement de la validité du modèle.	Les observations de terrain de la variable cible ainsi que des variables auxiliaires pour les parcelles sont prises en compte.	**Validité des estimations**
La validité des estimations (non	Les observations de terrain sont utilisées pour établir	La validité des estimations dépend du plan	

biaisées) dépend exclusivement du plan d'échantillonnage (sélection des parcelles d'échantillonnage, randomisation).	une relation (modèle) avec les variables auxiliaires qui sont généralement des indices issus de la télédétection. Le modèle est alors utilisé pour prédire la variable cible à partir d'une couverture exhaustive de ces indices.	d'échantillonnage – mais la précision de l'estimation peut être augmentée en intégrant l'information supplémentaire qui provient de la variable auxiliaire.
La télédétection ou des données auxiliaires ne sont pas intégrées dans la phase d'estimation, mais peuvent l'être dans la phase de planification, par exemple pour la stratification. Les estimations sont produites à partir des seules variables cibles des observations des parcelles.	**Exemple:** Pour chaque pixel d'une image par satellite, un modèle prédit la biomasse/ha, les statistiques sont ensuite dérivées comme agrégats des valeurs des pixels.	**Exemple:** Au lieu d'estimer directement la biomasse, un ratio entre la biomasse et, par exemple, l'IDNV est estimé, où les valeurs de l'IDNV servent de variable auxiliaire et sont disponibles pour l'ensemble de la surface forestière (population).

Estimation avec des parcelles en cluster

Contrairement à de nombreux manuels d'échantillonnage, nous ne considérons pas l'échantillonnage en cluster comme un plan d'échantillonnage en soi, mais comme un échantillonnage avec des parcelles en cluster, autrement dit, nous le considérons comme un plan parcellaire, car cela est plus cohérent avec la terminologie utilisée pour les plans parcellaires.

Néanmoins, les deux sont le même sens: un élément de l'échantillonnage singulier consiste en plusieurs sous-éléments, qui sont sélectionnés conjointement en une seule étape de randomisation. Puisque les sous-parcelles dans une parcelle en cluster ne sont pas sélectionnées indépendamment les unes des autres, la taille de l'échantillon se réfère au nombre de clusters sélectionnés et non pas au nombre de sous-parcelles. La parcelle en cluster peut être considérée comme une parcelle singulière avec une «forme étrange», où ces formes proviennent de l'ordonnancement spatialement disjoint de la parcelle.

Pour un échantillonnage aléatoire simple de parcelles en cluster: lorsque les observations des sous-parcelles sont agrégées au niveau du cluster = au niveau de la parcelle (une seule valeur, une moyenne

ou un total par cluster), l'estimation subséquente peut suivre les mêmes estimation introduits dans la leçon 1 (échantillonnage aléatoire simple: EAS). Cependant, il arrive souvent qu'on doive considérer des clusters avec différentes tailles (= différents nombres de sous-parcelles), car toutes les sous-parcelles ne sont pas toujours dans la population cible. L'estimateur par ratio sera alors un choix, utilisant la taille du cluster (le nombre de sous-parcelles) comme variable auxiliaire.

Il peut cependant être également intéressant de mener une analyse par sous-parcelles à l'intérieur des clusters; cela ne changera rien en termes de résultats des estimations ponctuelles et par intervalle, mais cela permettra des analyses supplémentaires de la structure spatiale des forêts et de l'efficacité du plan parcellaire en cluster.

Voyons les maths

Estimation avec des parcelles en cluster

Les clusters peuvent avoir un nombre de sous-parcelles (m) égal ou inégal pour tous les clusters. Dans cette section, on présente uniquement l'estimateur pour le cas de parcelles en cluster avec un nombre égal à partir d'un échantillonnage aléatoire. Pour les clusters de tailles inégales, une repondération des clusters sera nécessaire.

La moyenne estimée par sous-parcelle peut alors être calculée à partir de la moyenne estimée par cluster et du nombre moyen de sous-parcelles par cluster comme suit::

$$y = \frac{\bar{y}_{cl}}{\bar{m}}$$

et la variance d'erreur estimée de la moyenne par sous-parcelle est:

$$\hat{var}_{cl}(\bar{y}) = \frac{1}{\bar{m}^2}\frac{S_{y_i}^2}{n}$$

où $\mathbf{y}_i$ sont les observations par sous-parcelle. La variance estimée par cluster peut être dérivée en

calculant la variance des observations par cluster avec l'estimateur connu pour l'EAS:

$$S^2_{y_i} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{cl})^2}{n-1}$$

Les résultats totaux estimés, comme c'est habituel, proviennent de la multiplication de la moyenne par le total. Dans l'échantillonnage en cluster, on peut prendre la moyenne des clusters et le nombre de clusters (***N***), ou la moyenne par sous-parcelle et le nombre de sous-parcelles (***M***):

$$\tau = N * \bar{y}_{cl} = M * \bar{y}$$

Et la variance d'erreur respective pour le total peut être dérivée par::

$$var(\hat{\tau}) = N^2 \hat{var}(\bar{y}_{cl}) = M^2 \hat{var}(\bar{y})$$

Efficacité de l'échantillonnage avec des parcelles en cluster – corrélation intra-cluster

La similarité des observations au sein d'un cluster peut être quantifiée par les moyennes **du coefficient de corrélation intra-cluster (CIC**), parfois appelé coefficient de corrélation intraclasse. Plus cette corrélation est élevée, plus les observations issues de différentes sous-parcelles sont redondantes et le gain d'information deviendra moindre.

Cette analyse est très instructive pour comprendre et analyser la performance d'un échantillonnage en cluster pour des populations avec différentes structures d'auto-corrélation spatiale, car cela est directement reflété par le coefficient de corrélation intra-cluster. Lorsque le CIC est élevé, on peut considérer d'étendre les distances entre les sous-parcelles (ce qui augmente temps de marche et les coûts, bien sûr) ou de réduire le nombre de sous-parcelles.

Néanmoins, on doit prendre en compte que les CIC peuvent être différents pour différentes variables, et un plan en cluster pour une variable n'est pas nécessairement aussi optimal pour une autre. Prendre la surface terrière comme variable guide dans ces optimisations est une pratique commune, car elle présente une bonne corrélation avec plusieurs autres variables des arbres (par ex. la biomasse).

Pour les parcelles en cluster consistant en plusieurs sous-parcelles, il s'avère que si:

- CIC = 0 (observations non corrélées), il n'y a pas de différence de performance entre

l'échantillonnage en cluster de n clusters et l'EAS avec n*m sous-parcelles. On cherche à maintenir un CIC faible. Mais obtenir un CIC proche de zéro est impossible en pratique, car la distance entre les sous-parcelles serait trop importante.

- CIC < 0 (corrélation négative entre les sous-parcelles), l'échantillonnage avec n parcelles en cluster sera plus efficace qu'avec n*m sous-parcelles sélectionnées indépendamment. Cette situation est très peu probable dans les inventaires forestiers (du fait de l'auto-corrélation spatiale).
- CIC > 0 (redondance au sein des clusters) est le cas le plus typique. Échantillonner avec des parcelles en cluster est moins efficace qu'une sélection indépendante de parcelles singulières.

Néanmoins, si les coûts d'inventaire sont inclus, l'efficacité générale sera probablement supérieure grâce à la réduction des déplacements..

Échantillonnage stratifié

Nous avons appris que la stratification vise à sous-diviser la population totale en sous-populations plus homogènes, dans lesquelles des études par échantillonnage indépendantes sont mises en œuvre. Lorsque l'on combine les estimations singulières des différentes strates, il faut se rappeler que ces strates ont différentes tailles. Ainsi, il faut pondérer toutes les estimations par strate avec les tailles relatives des strates respectives. Dans le suivi des forêts, la taille d'une strate est généralement donnée en termes d'aire, et la somme tous les poids des strates sera 1 (soit égale à l'aire totale).

L'échantillonnage stratifié n'introduit pas de nouveau plan d'échantillonnage, c'est le cadre utilisé pour intégrer les estimations des différentes strates en une estimation pour l'aire totale qui est nouveau. L'échantillonnage stratifié introduit par conséquent une variation du plan d'estimation: la combinaison des estimations indépendantes de L strates en une seule estimation pour la population totale.

Voyons les maths

Estimation avec un échantillonnage stratifié

On utilise ici la notation h en tant qu'indice pour une strate et L pour le nombre total de strates.

Ainsi une moyenne non biaisée des strates multiples peut être estimée sous forme de somme pondérée. Les pondérations sont ici les proportions des aires des différentes strate ***N_h/N***:

$$y = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{y}_h$$

L'estimateur pour la variance d'erreur est:

$$\hat{var}(\bar{y}) = \sum_{h=1}^{L} \left\{ \left(\frac{N_h}{N} \right)^2 \hat{var}(\bar{y}_h) \right\} = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}$$

La racine carrée de cette variance d'erreur est l'erreur-type. Le total est dérivé par:

$$\hat{\tau} = N\bar{y} = \sum_{h=1}^{L} \frac{N_h}{N} \hat{\tau}_h = \sum_{h=1}^{L} N_h \bar{y}_h$$

Et la variance d'erreur du total estimé est:

$$\hat{var}(\hat{\tau}) = \hat{var}(N\bar{y}) = N^2 \hat{var}(\bar{y})$$

Efficacité de l'échantillonnage stratifié

Des considérations statistiques montrent que plus les moyennes des strates sont différentes, plus la stratification est efficace pour augmenter la précision de l'estimation de la moyenne. Ces effets positifs (soit une meilleure précision d'ensemble) tendent à se réduire avec un nombre croissant de strates.

D'un point de vue statistique, la formation de plus de six strates n'a généralement pas d'effet significatif sur l'amélioration de la précision de l'estimation. Cependant, on peut prendre en compte d'autres arguments, pas uniquement statistiques, pour former les strates. La question qui se pose aussi est de savoir si une post-stratification est plus indiquée dans ce cas.

Échantillonnage double pour la stratification

L'échantillonnage double pour la stratification a déjà été abordé dans la leçon 2. C'est un plan d'échantillonnage à deux phases pour estimer la taille des strates (qui ne peuvent pas être délimitées ou prédéfinies facilement). Puisque les aires (et les pondérations) des strates sont estimées à partir de l'échantillon de la première phase, l'erreur d'échantillonnage de l'estimation de ces aires doit être prise en compte lorsque l'on estime la moyenne et la variance pour l'ensemble de la population.

Voyons les maths

Estimation avec un échantillonnage double pour la stratification

Considérant que les pondérations des strates sont estimées à partir de l'échantillon de la

première phase (annotées avec l'apostrophe) comme

$$w'_h = \frac{n'_h}{n'}$$

Et une moyenne non biaisée peut être estimée avec:

$$\bar{y} = \sum_{h=1}^{L} w'_h \bar{y}_h$$

Ignorant la correction pour population finie et considérant que n' est grand, la variance d'erreur respective sera estimée par:

$$\hat{var}(\bar{y}) = \sum_{h=1}^{L} \left({w'_h}^2 * \frac{{s_h}^2}{n_h} + w'_h * \frac{(\bar{y}_h - \bar{y}')^2}{n'} \right)$$

Cet estimateur de variance est très similaire à l'estimateur dans un échantillonnage aléatoire stratifié – excepté pour le dernier terme entre parenthèses: un composant d'erreur est ajouté qui provient du fait que les tailles des strates sont uniquement estimées et pas connues.

L'outil Collect Earth (en anglais), qui fait partie du système Open Foris de la FAO, est utile dans ce contexte et a été appliqué de nombreuses fois. Il a été conçu pour utiliser l'imagerie par satellite et les images aériennes géoréférencées disponibles à partir de Google Earth, Bing, et d'autres sources, pour une interprétation visuelle des positions d'échantillonnage ou parcelles. À l'aide de cet outil, un plus grand nombre de points peuvent être visités et classifiés visuellement dans différentes strates. Ensuite, la taille des strates peut être estimée comme proportion des points échantillons par strate. Une variance de cette estimation peut être dérivée et incorporée à l'estimation présentée ci-dessus.

L'estimateur par ratio – exploiter l'information auxiliaire quantitative

Il existe des situations dans l'échantillonnage d'un inventaire forestier où l'on sait (ou l'on suspecte) que la valeur de la variable cible est fortement corrélée à une autre variable (appelée co-variable, variable auxiliaire ou variable subordonnée). Si un tel auxiliaire peut être observé dans la parcelle sans trop d'effort ni de coût (par ex. une analyse de télédétection), il sera efficace de l'observer également, et d'exploiter la corrélation avec la variable cible pour finalement améliorer la précision de l'estimation de la variable cible. C'est là que l'estimateur par ratio est appliqué.

Note

Imaginons qu'une classification d'imagerie par satellite a été menée pour produire une prédiction continue du **pourcentage de couvert arboré** pour une surface forestière avec une densité de couvert variable. Considérant une forte corrélation avec le volume ou la biomasse des parcelles, il s'agirait d'un cas où l'estimateur par ratio sera appliqué. Dans les surfaces forestières fermées, avec un couvert forestier complet partout, cela n'aurait aucun sens car le pourcentage de couvert arboré ne varierait pas mais serait constamment de 100 pour cent, de sorte que la corrélation entre la variable auxiliaire **pourcentage de couvert arboré** et la variable cible **biomasse** serait proche de zéro.

Au lieu d'estimer la biomasse du peuplement par unité de surface directement à partir des parcelles de terrain, l'estimateur par ratio prend un détour: on estime un ratio, r, des deux moyennes, ce qui donne **biomasse/pourcentage de couvert arboré**, et l'on utilise ensuite le couvert forestier connu pour dériver une estimation de la biomasse. La biomasse moyenne peut alors être estimée par **r*pourcentage de couvert arboré moyen**.

Un autre cas typique d'utilisation de l'estimateur par ratio se produit si une certaine proportion de grandes parcelles (ou parcelles en cluster) dépassent les limites de la région inventoriée et ne sont que partiellement dans la population cible. Dans ce cas, l'aire parcellaire dans la forêt n'est pas identique pour toutes les parcelles, et l'estimateur par ratio peut être appliqué, avec l'aire parcellaire comme variable auxiliaire. On peut en fait supposer que l'aire parcellaire sera fortement corrélée avec les variables de

stock (y compris la surface terrière, le volume, la biomasse, le carbone et le nombre d'arbres) enregistrées dans la parcelle. Pour une estimation de précision, on devra connaître la valeur paramétrique (moyenne) de la variable auxiliaire (ici, le pourcentage de couvert arboré moyen de la surface forestière totale, ou la surface totale de la forêt à inventorier dans l'exemple de l'aire parcellaire)

Voyons les maths

Estimation avec l'estimateur par ratio

Le ratio paramétrique entre la variable cible y et la variable auxiliaire *x*

$$R = \frac{\mu_y}{\mu_x}$$

est estimée à partir de l'échantillon avec:

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

La variance estimée de ce ratio estimé est:

$$var(r) = \frac{1}{n}\frac{1}{{\mu_x}^2}\frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$$

Le total estimé est dérivé de:

$$\tau_y = r\tau_x$$

avec une variance d'erreur associée de:

$$var(\hat{\tau}_y) = {\tau_x}^2 \hat{var}(r)$$

Étant donné le ratio estimé, r , la moyenne de la variable cible peut être estimée selon:

$$y_r = r\mu_x$$

Cette moyenne estimée comporte une variance estimée de :

$$var(\bar{y}_r) = \mu_x{}^2 v\hat{a}r(r) = \frac{1}{n}\left\{s_y{}^2 + r^2 s_x{}^2 - 2r\hat{\rho}s_x s_y\right\}$$

Voyons les maths

Plan d'estimation avec l'estimateur par régression

Alors que l'estimateur par ratio modélise la relation entre la cible et la variable auxiliaire, l'estimateur par régression utilise un modèle de régression avec à la fois le coefficient d'ordonnée à l'origine et le coefficient de pente. Rappelons qu'avec l'estimateur par ratio, on suppose que l'ordonnée à l'origine est zéro. La moyenne est estimée grâce à l'estimateur par régression comme suit:

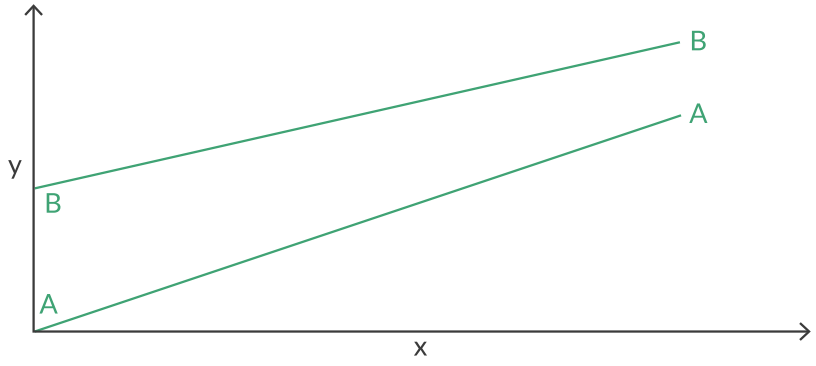
$$y_L = \bar{y} + b\left(\mu_x - \bar{x}\right)$$

La variance estimée de cette moyenne estimée est donnée par:

$$var(\bar{y}_L) = \frac{1}{n}\frac{1}{n-2}\left\{\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

Estimateur par ratio versus estimateur par régression

L'estimateur par ratio utilise un ratio fixe simple, ce qui signifie que si la variable cible, y, sera égale à zéro si la variable auxiliaire, x, est zéro. Néanmoins, il existe des situations ou cela n'est pas correct. Imaginons que nous pouvons trouver de petits arbres sur ces parcelles, pour lesquels aucun couvert arboré n'a été détecté dans les images de télédétection (par ex. du fait d'une faible résolution spatiale). Dans ce cas, une droite de régression avec un coefficient d'ordonnée à l'origine qui n'est pas forcément zéro (comme avec l'estimateur par ratio) serait plus adaptée; si, par exemple, le pourcentage de couvert arboré est zéro, il peut tout de même y avoir une biomasse considérable au sol. Ici, l'estimateur par régression utilise un modèle linéaire simple. Dans les deux cas, qu'il s'agisse des estimateurs par ratio ou par régression, l'efficacité d'ensemble dépend de la corrélation entre les variables cible et auxiliaire, qui devrait être fortement positive. Il s'avère parfois que cette corrélation est assez faible et les attentes étaient trop élevées, après, par exemple, une imagerie de télédétection très chère a été achetée.



Échantillonnage double (échantillonnage à deux phases)

Pour l'estimateur par ratio et l'estimateur par régression, la moyenne paramétrique ou le total paramétrique de la variable auxiliaire doivent être connus. Si cette information n'est pas disponible, on peut estimer ces valeurs à partir d'un échantillon.

C'est exactement ce que fait l'échantillonnage double, également appelé échantillonnage à deux phases: dans la première phase, la variable auxiliaire est estimée, généralement avec un grand échantillon d'une variable qui peut être observée facilement et à bas coût, et dont l'on sait qu'elle est fortement corrélée

positivement avec la variable cible.

Ensuite, dans l'échantillon de la seconde phase, un échantillon plus petit de la variable cible est prélevé, qui est fréquemment une variable beaucoup plus chère ou difficile à observer. Une relation entre la variable cible et la variable auxiliaire peut alors être établie (avec un ratio simple ou une régression, ce qui supposerait un échantillonnage double avec l'estimateur par ratio ou l'estimateur par régression, respectivement).

Ici, plus la corrélation positive est forte avec la variable auxiliaire, plus l'on pourra réduire la taille de l'échantillon nécessaire dans la seconde phase, où la variable cible plus complexe/chère/difficile est observée.

On considère ici des phases dépendantes, où l'échantillon de la seconde phase est un sous-ensemble de la première phase (et non pas un échantillon sélectionné indépendamment). Les estimateurs présentés sont valables exclusivement pour l'EAS.

Voyons les maths

Estimation avec un échantillonnage double

Pour un échantillonnage double avec l'estimateur par ratio, la moyenne de y peut être estimée par:

$$\bar{y}_{md.r} = \frac{\bar{y}}{\bar{x}}\bar{x}' = r\bar{x}'$$

Avec une variance estimée de la moyenne estimée de:

$$\hat{var}(\bar{y}_{md.r}) = \frac{S_y^2 + r^2 S_x'^2 - 2rS_{xy}}{n} + \frac{2rS_{xy} - r^2 S_x'^2}{n'} - \frac{S_{\cdot}}{N}$$

Et avec l'estimateur par régression, la moyenne est estimée par:

$$\bar{y}_{md.reg} = \bar{y} + b(\bar{x}' - \bar{x})$$

Avec une variance estimée de la moyenne estimée de:

$$\hat{var}(\bar{y}_{md.reg}) = \frac{S_y^2}{n}\left\{ 1 - \frac{n' - n}{n'}\hat{\rho}^2 \right\}$$

Où ρ est le coefficient de corrélation estimé entre x et y.

Dans les deux cas, la variance d'erreur du total est calculée, comme généralement, par:

$$\hat{var}(\hat{\tau}) = N^2 \hat{var}(\bar{y}$$

L'efficacité générale de l'échantillonnage double dépend de la relation de coût entre l'observation des échantillons de la phase 1 et de la phase 2, et de la corrélation entre les deux variables. En fait, on cherche à exploiter la variable auxiliaire le plus possible, pour pouvoir réduire le nombre d'échantillons (coûteux) de la seconde phase. Plus la corrélation est élevée et plus les observations de la seconde phase sont chères, plus l'on réduira l'échantillon de la phase 2.

Le saviez-vous?

Faire un choix parmi des estimateurs alternatifs

Selon le plan d'échantillonnage appliqué, plusieurs estimateurs alternatifs pourraient être appliqués. Par exemple, une estimation peut être produite à partir des échantillons de terrain uniquement, ou considérer des variables auxiliaires supplémentaires. Ou une post-stratification peut être appliquée

ou non aux données. Dans ces situations, la production d'estimations valides différentes, avec des estimateurs alternatifs, devrait donner la même moyenne, mais avec des estimations de précision différentes. Dans le cas où de multiples estimateurs non biaisés sont disponibles, on préfèrera celui qui produit la moindre erreur-type des estimations.

Résumé

Avant de conclure, voici les principaux points d'apprentissage de cette leçon:

- Pour l'inférence fondée sur un calcul, on n'émet aucune hypothèse concernant la structure (spatiale) de la population. On considère la structure inconnue, et l'on cherche à estimer les caractéristiques de cette population fixe.
- Avec l'inférence fondée sur un modèle, les observations de terrain sont utilisées pour établir une relation (modèle) avec des variables auxiliaires qui sont généralement des indices issus de la télédétection. Le modèle est alors utilisé pour prédire la variable cible à partir d'une couverture exhaustive de ces indices.
- Dans l'inférence assistée par modèle, un modèle est utilisé en appui à l'estimation fondée sur le plan, quelque part entre l'inférence fondée sur un calcul et l'inférence fondée sur un modèle.
- Lorsque l'on utilise des parcelles en cluster, un élément de l'échantillonnage singulier (parcelle) consiste en plusieurs sous-éléments (sous-parcelles), qui sont sélectionnés conjointement. Puisque les sous-parcelles dans une parcelle en cluster ne sont pas sélectionnées indépendamment les unes des autres, la taille de l'échantillon se réfère au nombre de clusters sélectionnés, et non pas au nombre de sous-parcelles.
- La similarité des observations à l'intérieur d'un cluster peut être quantifiée grâce au coefficient de corrélation intra-cluster (CIC), également connu comme le coefficient de corrélation intraclasse.
- L'échantillonnage stratifié ne constitue pas un nouveau plan d'échantillonnage, mais un cadre d'intégration d'estimations à partir d'échantillons générés indépendamment pour différentes strates, en une estimation pour la population totale, autrement dit: c'est plutôt une variation du

plan d'estimation.

- Dans l'échantillonnage double pour la stratification, la taille des strates n'est pas déterminée avant l'échantillonnage, mais est estimée dans la première phase d'échantillonnage.