Consortium de Recherche FOR@C
Pavillon Adrien-Pouliot
Université Laval
Québec, G1K 7P4
www.forac.ulaval.ca

2424, chemin Sainte-Foy
Sainte-Foy (Québec)
G1V 1T2
www.cerfo.qc.ca

STATISTIQUE APPLIQUÉE À LA FORESTERIE

Auteurs

CERFO - Développement du contenu
Donald Blouin, ing.f., M.Sc., Coordonnateur-adjoint, Aménagement forestier
Sylvie Côté, ing.f., M.Sc.
François Guillemette, ing.f., M.Sc.
Hugues Lapierre, ing.f., M.Sc.
Claire Roy, secrétaire

FORAC - Réalisation de la version en ligne
Amélie Tremblay, graphiste
Philippe Marier, ing. , M.B.A.

TABLE DES MATIÈRES

INTRODUCTION

Lors de la réalisation des interventions forestières sur les forêts publiques au Québec, les bénéficiaires de CAAF doivent rendre compte de ce qu'ils font et de la qualité de leurs travaux au ministère des Ressources naturelles et de la Faune. Afin de guider les bénéficiaires dans ces démarches, le MRNF publie et met à jour régulièrement deux documents de référence : les « Instructions relatives à l'application du règlement sur la valeur des traitements sylvicoles admissibles en paiement des droits » (MRNF, 2004a) et les « Méthodes d'échantillonnage pour les inventaires d'interventions… et pour les suivis des interventions forestières… » (MRNF, 2004b). Il s'agit de documents ayant force de loi décrivant les procédures à suivre pour l'évaluation de la qualité des interventions forestières. Pour s'entendre, les bénéficiaires et les agents du Ministère doivent d'abord connaître et comprendre les démarches suggérées et savoir interpréter les résultats des compilations réalisées.

Cette formation vise à présenter les notions requises à l'évaluation et au contrôle de la qualité des interventions forestières. Elle s'attarde aux éléments de base de la théorie de l'échantillonnage et à la réalisation des compilations statistiques nécessaires pour les prescriptions sylvicoles, le suivi des interventions et le suivi après intervention. Il est vrai que les références utilisées dans la présente formation sont inspirées des besoins d'information relatifs aux activités réalisées dans les forêts du domaine public. Toutefois, il est important de noter que les principes présentés sont universels et que ceux-ci peuvent être mis en application dans différents contextes et champs d'expertise.

OBJECTIFS

Cette formation a comme objectif général de :

Comprendre les principes fondamentaux de l'échantillonnage et de l'estimation afin de les appliquer judicieusement lorsqu'une situation l'exige.

De façon plus spécifique, cette formation poursuit les objectifs suivants :

- mettre en application les concepts d'échantillonnage pour la réalisation d'un plan de sondage
- justifier le choix du type de point d'observation (grandeur de placette)
- justifier le nombre de placettes à réaliser (calculer la taille d'un échantillon)
- utiliser les statistiques pour l'évaluation de la fiabilité des résultats
- connaître les limites de l'estimation des événements rares
- comparer 2 échantillonnages.

FORMATION COMPLÉMENTAIRE

Si vous désirez poussez plus loin votre formation en statistique en raison des mandats que vous confient vos clients ou votre employeur, nous vous invitions à communiquer avec l'OIFQ de façon à ce qu'il soit en mesure de connaître approximativement le nombre de membres pouvant être intéressés et la nature de vos besoins de formation. À la suite de quoi l'Ordre pourrait organiser un cours en salle, et avec professeur (s), afin de répondre à vos besoins.

MISE EN CONTEXTE

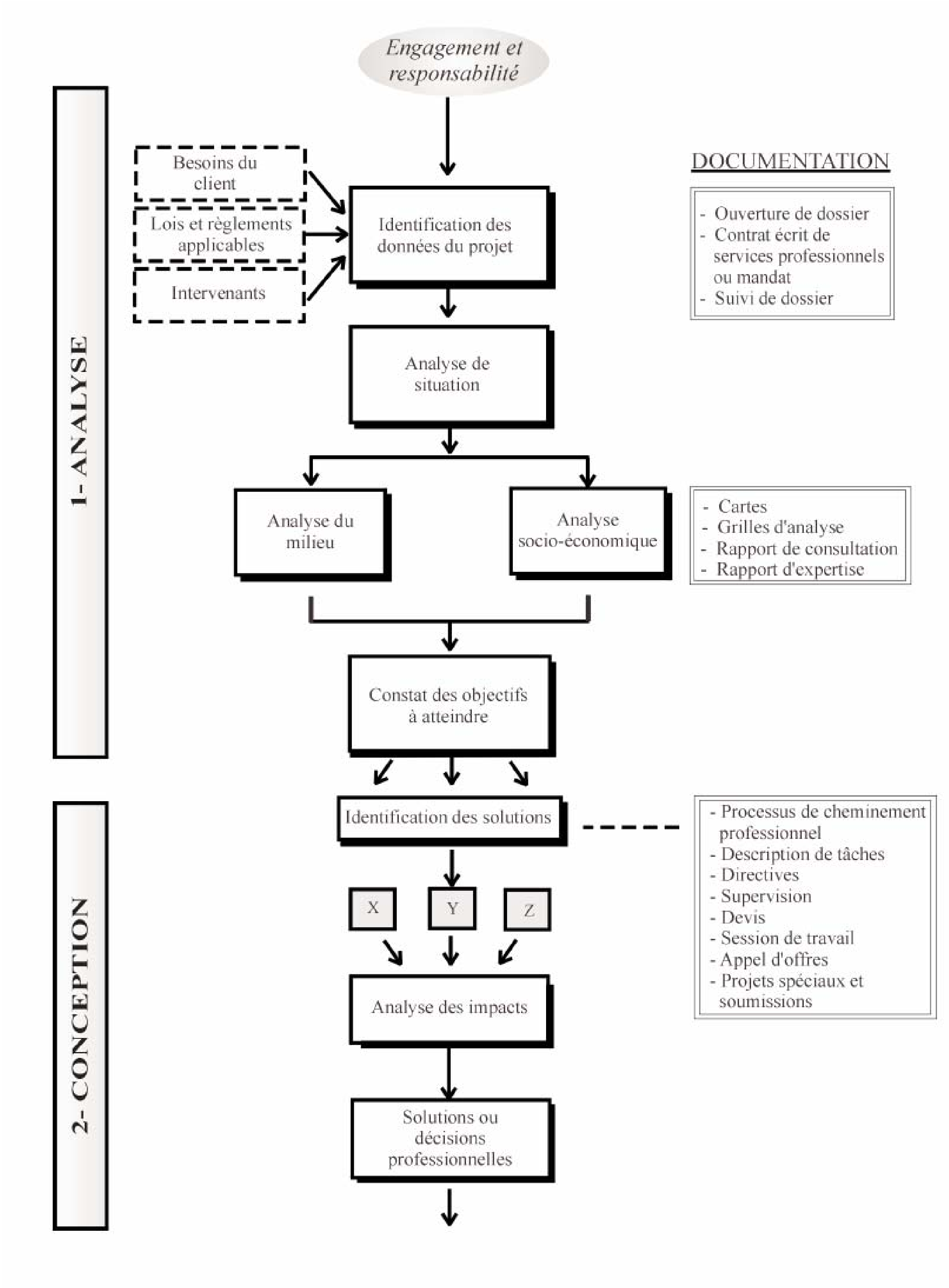
Dans le cadre du système professionnel à l'intérieur duquel l'ingénieur forestier exerce sa profession, l'engagement de conformité à la législation, ainsi que la mise en application du principe d'amélioration continue constituent des exigences essentielles (OIFQ, 2001). Par ailleurs, l'ingénieur forestier est responsable de tout avis ou conseil qu'il donne et, à cet égard, il doit évidemment posséder les connaissances et l'expertise adéquates pour être en mesure de se prononcer sur un sujet donné. De plus, il doit éviter d'émettre une opinion sans avoir vérifié tous les faits et il doit aussi indiquer les limites d'interprétation de son avis (OIFQ, 2001).

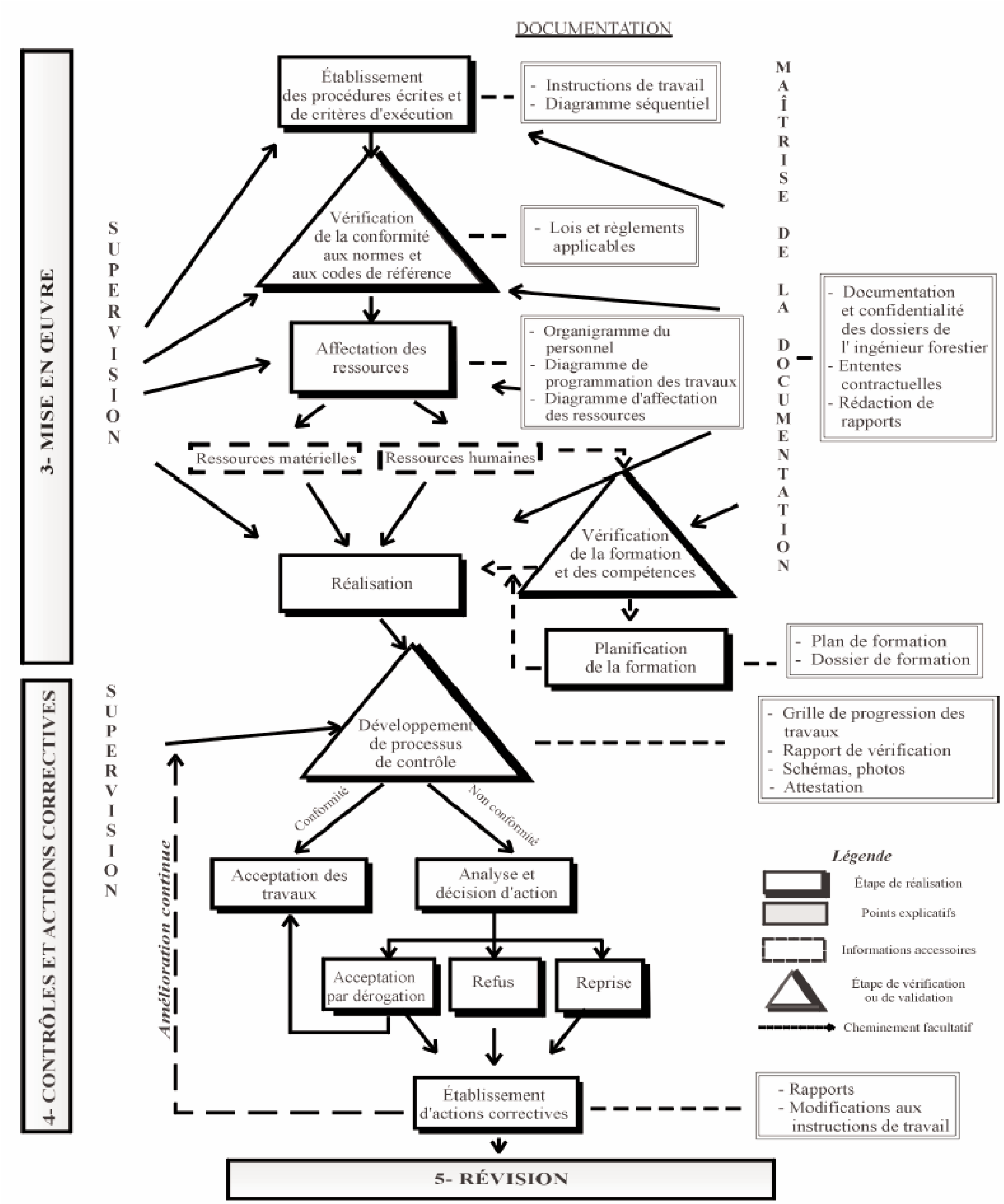
Afin d'assurer l'engagement à la conformité à la législation en matière de suivi des interventions forestières, l'ingénieur forestier doit être en mesure de planifier et superviser la réalisation de l'échantillonnage requis par la réglementation, puis procéder à la compilation des données et à l'analyse des résultats.

Pour faciliter la réalisation de ces tâches, plusieurs professionnels ont exprimé le désir d'avoir accès à un complément d'information en statistiques appliquées qui soit adapté aux exigences spécifiques relatives au suivi des interventions forestières. C'est dans une optique d'amélioration continue que cette formation est présentée.

L'utilisation des statistiques appliquées constitue un exemple concret de la nécessité de bien maîtriser la documentation nécessaire aux étapes de conception, de mise en œuvre et de contrôle et actions correctives, du cheminement professionnel illustré à la figure 1, appliqué au suivi des interventions forestières. Dans ce contexte particulier, les notions de statistiques doivent intervenir aux étapes de mise en œuvre et de contrôle, autant au niveau de l'établissement des procédures et critères d'exécution, qu'à celles de la vérification de la conformité aux normes et du développement de processus de contrôle conduisant soit à la validation de la conformité des travaux, soit à la définition des actions requises en cas de non-conformité.

Figure 1 Description du cheminement professionnel




DOCUMENTATION
3- MISE EN ŒUVRE
SUPERVISION
Établissement des procédures écrites et de critères d'exécution
- Instructions de travail
- Diagramme séquentiel
MAÎTRISE DE LA DOCUMENTATION
Vérification de la conformité aux normes et aux codes de référence
- Lois et règlements applicables
Affectation des ressources
- Organigramme du personnel
- Diagramme de programmation des travaux
- Diagramme d'affectation des ressources
- Documentation et confidentialité des dossiers de l' ingénieur forestier
- Ententes contractuelles
- Rédaction de rapports
Ressources matérielles
Ressources humaines
Réalisation
Vérification de la formation et des compétences
Planification de la formation
- Plan de formation
- Dossier de formation
4- CONTRÔLES ET ACTIONS CORRECTIVES
SUPERVISION
Développement de processus de contrôle
- Grille de progression des travaux
- Rapport de vérification
- Schémas, photos
- Attestation
Conformité
Non conformité
Acceptation des travaux
Analyse et décision d'action
Acceptation par dérogation
Refus
Reprise
Amélioration continue
Établissement d'actions correctives
- Rapports
- Modifications aux instructions de travail
5- RÉVISION
Légende
Étape de réalisation
Points explicatifs
Informations accessoires
Étape de vérification ou de validation
Cheminement facultatif

CONTENU

1. Présentation des principes liés à l'acquisition de connaissances du territoire forestier

1.1 Statistiques descriptives

Les statistiques descriptives permettent de résumer un ensemble de données. Généralement les résultats se limitent aux sujets étudiés. Les statistiques descriptives incluent :

- les distributions de fréquence (discrètes et continues)
- les mesures de tendance centrale (moyenne, médiane, mode)
- les mesures de dispersion (étendue, écart-type, variance).

1.1.1 Définition de constante

Une constante est une caractéristique qui a la même valeur pour toutes les observations d'une population alors qu'une variable est une caractéristique pouvant prendre différentes valeurs pour différentes observations d'une population. Pour chaque observation, on peut observer une ou plusieurs variables, selon le but fixé.

1.1.2 Définition de variable

Une variable peut être **QUANTITATIVE** ou **QUALITATIVE**.

- **VARIABLE QUANTITATIVE** : variable pouvant être exprimée sous forme numérique et possédant un ordre de gradation naturel. Ex : le diamètre, la hauteur, le nombre de tiges par peuplement.

  Les variables quantitatives peuvent être **continues** ou **discrètes** :

  - ***Variable continue :*** variable quantitative pour laquelle toutes les valeurs à l'intérieur d'un intervalle donné sont possibles sur une échelle numérique. Ex. : la hauteur, le volume;

  - ***Variable discrète :*** variable quantitative dont les valeurs ne peuvent être que des nombres entiers; il n'est donc pas possible d'observer toutes les valeurs sur une échelle numérique, puisque certaines valeurs n'existent tout simplement pas. Ex : le nombre de tiges dans un peuplement.

- **VARIABLE QUALITATIVE :** variable qui ne peut être exprimée sous forme numérique. Les variables qualitatives ne possèdent pas d'ordre naturel. Ex : l'essence, la qualité des tiges, le type de sol.

1.2 Statistiques inférentielles

Les statistiques inférentielles permettent la généralisation des résultats obtenus à partir d'un échantillon à toute la population d'où l'échantillon est tiré. Basé sur la loi des probabilités, l'inférence statistique consiste, par exemple, à extrapoler pour une population donnée les comportements étudiés auprès d'un échantillon de celle-ci. L'inférence statistique inclut :

- les tests d'hypothèse (comparaison de moyennes)
- l'estimation de paramètres.

1.2.1 Définition de population

La population consiste en toutes les observations possibles d'une variable. La population représente l'ensemble des individus, des éléments ou unités de même nature auquel s'adresse l'échantillonnage (Rondeux, 1999). Dans le cas d'un peuplement ou d'un massif forestier, la population pourrait être représentée par l'ensemble des arbres, alors que dans le cas d'un inventaire forestier, il peut être plus aisé de considérer la forêt comme une population de placettes (dont on a échantillonné une certaine proportion). Une population est dite finie lorsque le nombre d'observations qu'elle comporte est dénombrable ou limité alors qu'elle est dite infinie dans le cas inverse.

> Lors de la planification d'interventions sylvicoles, l'unité d'échantillonnage est la référence utilisée par le ministère des Ressources naturelles, de la Faune et des Parcs (MRNF, 2004b). Il s'agit d'un territoire ayant une superficie maximale de 250 ha, le plus homogène possible, sur lequel un seul type d'intervention sylvicole sera pratiqué. L'unité d'échantillonnage définie par le MRNF correspond à la population en termes mathématiques.

1.2.2 Acquisition de connaissances : Recensement ou échantillonnage

Lorsque toutes les valeurs d'une population sont connues, il est possible de décrire cette population sans ambiguïté. Ce type d'information s'obtient à l'aide d'un recensement ou d'un inventaire[1] qui correspond à une évaluation portant sur l'ensemble de la population. Par contre, lorsqu'il est techniquement impossible de mesurer l'ensemble des individus de la population, on peut tout de même caractériser la population en ayant recours à l'échantillonnage.

L'échantillonnage vise à estimer la valeur d'une (ou plusieurs) caractéristique(s) d'une population à partir d'un échantillon (Rondeux, 1999). L'échantillon réfère donc à une partie de la population (un sous-ensemble) et, de façon générale, les informations obtenues à partir de l'échantillon sont utilisées pour estimer une caractéristique de la population.

L'échantillon est constitué d'un certain nombre d'observations. L'observation est le relevé d'une variable. En foresterie, l'observation peut prendre différentes formes : un arbre, une placette, une micro-placette, un point de prisme, etc.

De façon générale, on désigne sous le nom de « paramètre » une caractéristique d'une population, et sous le nom de « statistique » une caractéristique d'un échantillon. On désigne les paramètres par des lettres grecques (ex : μ, σ), et les statistiques par des lettres romaines (ex : x, s).

1.2.3 Théorie de l'échantillonnage

Afin d'être considéré comme valable, l'échantillon doit être choisi en respectant certaines règles de base. Ces règles visent à éliminer toute subjectivité dans le processus de positionnement et d'établissement des points d'observation et à faire en sorte que les résultats obtenus soient les plus représentatifs possibles de la population.

1.2.3.1 Représentativité de l'échantillon

Pour être en mesure d'estimer une caractéristique de la population de façon adéquate, l'échantillon doit être représentatif de cette population. Pour ce faire, l'échantillonnage doit être effectué de manière aléatoire (i.e. les points d'observation mesurés doivent être sélectionnés au hasard dans la population), ce qui permet d'éviter les biais. Dans ce cas, et dans ce cas seulement, les lois de la probabilité s'appliquent et les résultats obtenus peuvent être utilisés pour estimer une caractéristique de la population dans son ensemble.

[1] Selon le Petit Robert, l'inventaire est une opération qui consiste à énumérer et décrire les éléments composant un ensemble. Dans le domaine forestier, la terminologie « inventaire » est utilisée de façon généralisée pour désigner l'action de dénombrer les arbres sur une surface donnée (une placette), ce qui en terme mathématique correspond à l'échantillonnage.

Le principe du choix aléatoire des observations signifie que chaque point d'observation a la même chance d'être sélectionné et de faire partie de l'échantillon. De plus, le processus servant à déterminer la localisation des points d'observation doit être automatisé et le déplacement d'un de ces points d'observation n'est pas autorisé, et ce, pour éviter l'introduction de biais.

1.2.3.2 Taux d'échantillonnage

Le taux d'échantillonnage correspond au rapport entre la taille de l'échantillon et la taille de la population (Rondeux, 1999). Ce taux est aussi désigné par les appellations suivantes : intensité d'échantillonnage, taux de sondage ou fraction sondée.

En présence d'une population finie composée de *N* observations de même taille, pour laquelle *n observations* ont été échantillonnées, le taux d'échantillonnage *f* est égal à :

$$f = n/N \qquad \textbf{(Formule 1)}$$

À titre d'exemple, la réalisation de 50 placettes circulaires d'un rayon de 11,28 m (400 m^2) sur une superficie de 250 hectares correspond à effectuer un échantillon d'une superficie de 20 000 m^2 (50 x 400 m^2) dans une population de 2 500 000 m^2, soit un taux d'échantillonnage de 1/125 de la superficie de la population.

Pour l'évaluation des bois de dimension précommerciale, l'utilisation de placettes de 100 m^2 résulte en un taux d'échantillonnage équivalant à 1/500 de la superficie.

1.3 Types d'échantillonnage

Il existe plusieurs types d'échantillonnage; nous ferons ici la description de l'échantillonnage aléatoire simple et de l'échantillonnage systématique aléatoire.

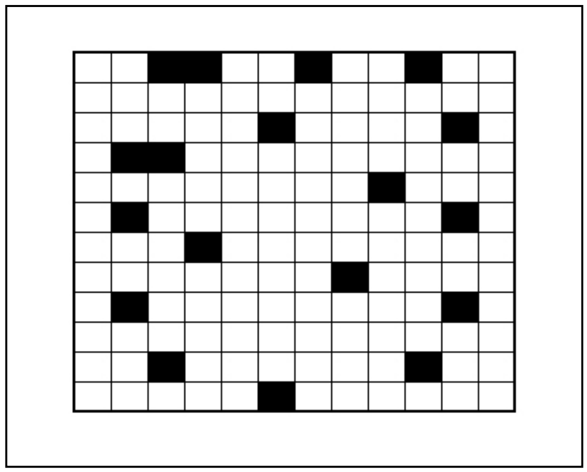
1.3.1 Échantillonnage aléatoire simple

L'échantillonnage aléatoire simple consiste à choisir au hasard la localisation des points d'observation (Figure 2). Sur une superficie de 250 ha avec des points d'observation de 400 m^2, en supposant que l'on puisse quadriller la population parfaitement, il y a 6 250 points d'observation disponibles dans la population. En appliquant un taux d'échantillonnage de 1/125, il s'agira de choisir au hasard 50 points d'observation.

a) Principes et pratique

Ce type d'échantillonnage s'applique pour une population finie et est caractérisé par le fait que tous les échantillons de *n observations* de la population ont une probabilité égale d'être choisis.

Figure 2 Illustration de l'échantillonnage aléatoire simple



Chaque carreau noir représente un point d'observation
(Source : OIFQ, 1996)

b) Avantages et inconvénients

Les avantages de l'échantillonnage aléatoire sont les suivants (tiré de OIFQ, 1996) :

- il élimine toute subjectivité dans le choix des points d'observation à mesurer, car le choix n'est nullement influencé par le personnel qui les sélectionne;
- il permet de calculer l'erreur d'échantillonnage.

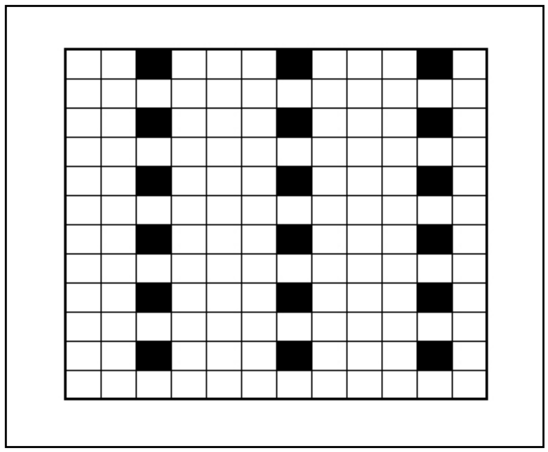
Il comporte cependant les désavantages suivants (tiré de OIFQ, 1996) :

- il est parfois difficile de se rendre sur les points d'observation choisis aléatoirement;
- les points d'observation sont distribués de façon irrégulière et certains secteurs peuvent donc être sur ou sous-échantillonnés;
- les distances qui séparent les points d'observation peuvent prolonger la durée des travaux;
- cette méthode peut entraîner des coûts prohibitifs si les points d'observation sont concentrés dans des zones difficilement accessibles.

1.3.2 Échantillonnage systématique aléatoire

Pour la méthode d'échantillonnage systématique, la localisation du premier point d'observation est choisie aléatoirement, alors que les autres points sont tous équidistants (Rondeux, 1999) (Figure 3).

Figure 3 Illustration de l'échantillonnage systématique aléatoire



Chaque carreau noir représente un point d'observation
(Source : OIFQ, 1996)

a) Principes et pratique

On a plutôt recours à l'échantillonnage systématique aléatoire dans la pratique. L'échantillonnage aléatoire simple représente davantage un concept théorique et est très peu utilisé dû aux contraintes d'application qu'il exige.

b) Avantages et inconvénients

Les avantages de l'échantillonnage systématique sont les suivants (tiré de OIFQ, 1996) :

- l'échantillon est plus facile à sélectionner, car on a besoin d'un seul nombre aléatoire;
- les points d'observation sont généralement mieux répartis dans la population, on peut donc estimer la moyenne plus justement qu'avec un échantillon aléatoire comparable;
- les coûts d'accès aux points d'observation sont moindres, car le nombre de points d'observation d'une virée est fixé en fonction de ce qu'une équipe peut faire dans une journée normale de travail;
- les équipes risquent moins de se tromper dans l'emplacement des points d'observation;
- la localisation des points d'observation est plus facile.

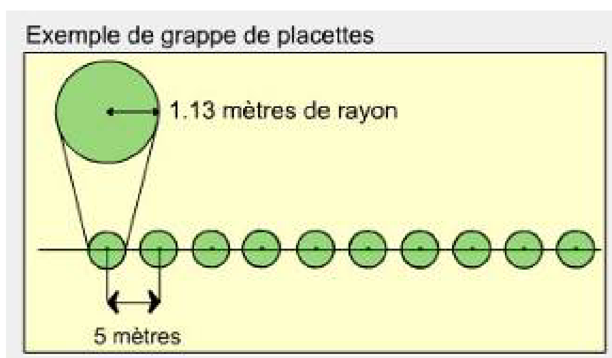
Les principaux inconvénients de cette méthode sont (tiré de OIFQ, 1996) :

- la non–possibilité théorique de calcul valable de l'erreur d'échantillonnage, car elle n'est pas fondée sur les lois du hasard. Seulement le premier point d'observation est tiré au hasard. Le choix du premier point détermine automatiquement le choix des autres, qui y deviennent dépendants. En pratique, les formules utilisées pour l'échantillon aléatoire peuvent aussi servir à calculer l'erreur d'échantillonnage d'un sondage systématique (Rondeux,1993; Cochran, 1977);
- la périodicité, problème qui se pose lorsque la population présente un caractère cyclique qui coïncide avec le pas de sondage. L'échantillon ainsi obtenu n'est pas représentatif de la population.
- Supposant qu'en forêt naturelle, il n'y ait pas de répartition cyclique des arbres ou des conditions dans lesquelles ils se trouvent, l'application de l'échantillonnage systématique est reconnue et les compilations statistiques qui lui sont associées sont tout à fait justifiées.

1.4 Type de points d'observation

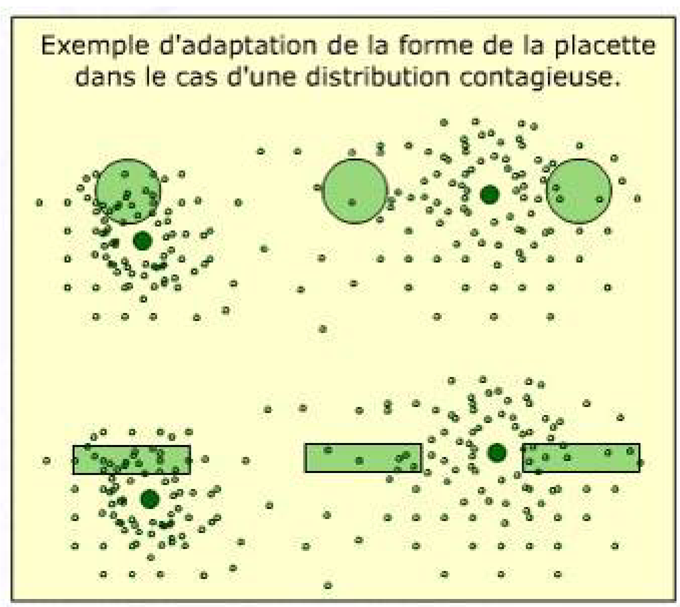
L'unité de base de mesure est le point d'observation. Le point d'observation peut se matérialiser de différentes manières. Par exemple, pour l'estimation du diamètre des tiges d'un peuplement, c'est l'arbre individuel qui sert de point d'observation. Pour l'estimation d'un paramètre comme le volume de bois, le point d'observation prend la forme d'une placette. Pour l'estimation de la distribution d'une essence, le point d'observation prend la forme d'une grappe de placettes, soit le regroupement des placettes (habituellement par groupe de 10) selon une disposition préalablement déterminée.

Les types de placettes utilisées dans le milieu forestier sont soit à surface définie ou à rayon variable. La taille des placettes dépend de l'objectif poursuivi et du paramètre que l'on cherche à estimer.

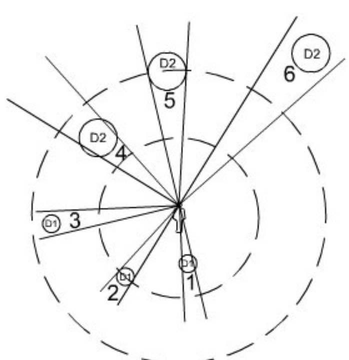


1.4.1 Principes et facteurs déterminant la forme de la placette

Selon Rondeux (1999), les placettes à surface définie peuvent être de formes circulaire, carrée, rectangulaire ou par bande. La forme circulaire est généralement utilisée parce qu'elle est la plus simple à installer sur le terrain. Elle permet une délimitation rapide à condition que la placette n'ait pas une trop grande surface. Cette forme possède aussi le plus petit ratio de périmètre par rapport à la surface, ce qui permet de réduire le nombre d'arbres situés en bordure de placettes qui sont susceptibles de représenter des cas litigieux. Les placettes carrées ou rectangulaires nécessitent l'utilisation d'un matériel approprié pour mesurer les angles; et les risques d'erreurs associés à ces mesures sont relativement grands.



Il arrive toutefois qu'une autre forme que la forme circulaire soit plus appropriée. C'est le cas, par exemple, de l'inventaire de régénération où l'on vise principalement à caractériser la densité d'une essence ayant une distribution contagieuse (i.e. où la régénération se retrouve principalement en îlots autour de l'arbre mère), tel que le sapin. Pour ce cas particulier, il est possible de chercher à mieux couvrir la variabilité existante sur le terrain par l'emploi de placettes rectangulaires qui permettent de ne pas concentrer l'échantillonnage seulement dans les taches densément régénérées ou dans les vides présents entre ces taches (Côté et Bélanger, 1988). Cette disposition permet de couvrir les différentes conditions présentes sur le terrain dans chacune des placettes, tout en minimisant les différences entre celles-ci et conséquemment de favoriser une réduction de la variance du nombre de tiges par grappe de placettes.

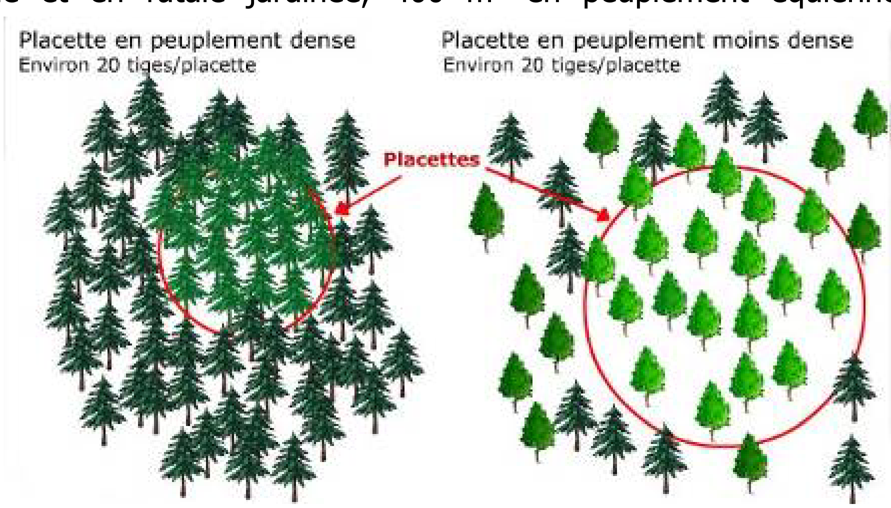


Les placettes à rayon variable correspondent aux placettes effectuées à l'aide d'un instrument projetant un angle constant, tels que le prisme, la jauge d'angle ou le relascope. Dans ces placettes, la probabilité de sélection des arbres est proportionnelle à leur surface terrière (Rondeux, 1999). Les placettes effectuées à l'aide de cette méthode sont rapides à exécuter lorsque les peuplements sont de densité moyenne à faible et que la pente est faible (OIFQ, 1996). Par contre, en peuplements denses, il y a un risque d'oublier des arbres qui peuvent être masqués par d'autres, et sur les terrains en pente, il est nécessaire soit d'effectuer des corrections, avec les risques d'erreur qui y sont associés, soit de procéder à des mesures supplémentaires, ce qui ralentit l'exécution. Étant donné que la superficie de la placette dépend du diamètre mesuré, cette méthode se prête mal à la fixation d'un taux d'échantillonnage exact (OIFQ, 1996).

Dans une étude réalisée par Blouin et al., (2001), l'utilisation de placettes à rayon fixe (11,28 m) et à rayon variable (prisme facteur 2) a été comparée en utilisant les mêmes centres de 36 placettes installées dans un peuplement mixte à dominance résineuse. La compilation des données obtenues à partir des placettes à rayon variable démontre qu'il n'y a pas de différence significative avec les moyennes obtenues des placettes à rayon fixe pour l'estimation du volume, de la surface terrière et du nombre de tiges. Les résultats de précision sont également similaires entre les deux méthodes à l'exception de l'évaluation du nombre de tiges à l'hectare qui présente des précisions inférieures pour la méthode à rayon variable (prisme facteur 2).

1.4.2 Principes et facteurs déterminant la grandeur de la placette

La dimension des placettes est déterminée de façon à obtenir un nombre d'arbres suffisant pour constituer un échantillon représentatif du peuplement. La grandeur des placettes sera donc fonction de la disposition des tiges et de la variabilité retrouvée dans le peuplement évalué. Il faut cependant garder en tête qu'à taux d'échantillonnage égal, il est préférable de disposer d'un nombre de placettes qui soit le plus élevé possible, donc de réduire la surface des placettes, en autant qu'elles demeurent représentatives. Tel que mentionné dans Rondeux (1999), le nombre d'arbres idéal que l'on doit retrouver dans une placette, qui varie selon les auteurs et le type de forêt, se situerait entre 10 et 30 et serait entre 15 et 20 pour les futaies résineuses. En terme de superficie, la placette en futaie[2] âgée devrait mesurer entre 800 et 1 000 m^2, environ 500 m^2 en futaie pleine et en futaie jardinée, 400 m^2 en peuplement équienne homogène et pas trop âgé et moins de 400 m^2 dans les très jeunes peuplements. Ajoutons que la nécessité de couvrir une plus grande superficie en forêt âgée est reliée à la plus grande variabilité retrouvée dans ces forêts qui est attribuable à l'ouverture du couvert



[2] La futaie réfère à un peuplement fermé composé d'individus issus de graines et représente le dernier stade de développement dans la séquence fourré, gaulis, perchis, futaie.

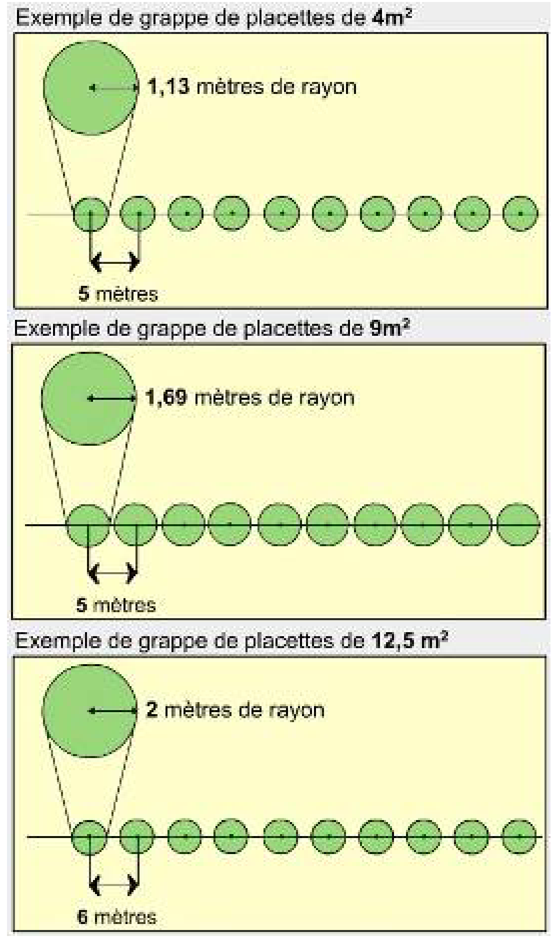
liée à la sénescence, associée à la présence de gros individus. Signalons aussi au passage que la dimension des placettes sur le terrain correspond à la projection horizontale du rayon pour l'obtention de la surface désirée.

Une autre étude a été réalisée, afin de comparer les surfaces terrières initiales, martelées et résiduelles mesurées pour le suivi d'interventions de jardinage en forêt feuillue dans 5 types de placettes, soit des placettes à rayon fixe de 20, 14,56 ou 11,28 m et des placettes à rayon variable effectuées à l'aide d'un prisme de facteur 1 et 2 (Desaulniers, 2004). Les résultats obtenus, pour l'essai en question, indiquent que l'utilisation d'une placette fixe de 20 m de rayon permettait de réduire les variations observées pour les trois types de surface terrière échantillonnés, donc de minimiser le nombre de placettes requis pour atteindre un niveau de précision donné. Pour l'évaluation de la surface terrière initiale, la stabilité obtenue avec des placettes de 20 m de rayon permettrait de réduire le nombre de placettes requis à un niveau tel que la superficie totale échantillonnée serait inférieure à celle nécessaire avec des placettes de rayon inférieur (14,56 et 11,28 m). En revanche, pour l'évaluation des tiges martelées et résiduelles, la superficie totale à inventorier est moins importante avec les placettes de 11,28 m, ce qui indique que ce type de placette serait plus performant pour l'évaluation de l'effet du traitement puisq'aucune différence significative n'a été observée entre les différents types de placettes à rayon fixe. Pour la surface terrière initiale, seule celle obtenue à l'aide du prisme F1 est significativement différente des autres qui ne diffèrent pas entre elles. Enfin, pour obtenir une précision équivalente, l'effectif (donc la superficie totale échantillonnée) doit être plus grand pour évaluer la surface terrière des tiges martelées, que celui nécessaire à l'obtention du niveau recherché pour estimer la surface terrière initiale.

Dans les jeunes peuplements résineux avec un objectif de 2500 tiges/ha, il est approprié d'utiliser des placettes de 100 m^2 (rayon de 5,64 m) pour évaluer la densité de tiges, considérant qu'une placette devrait théoriquement contenir 25 tiges.

La présence de plus de 50 tiges par placette signifie que la dimension de celle-ci n'est pas bien adaptée au milieu échantillonné. En effet, un nombre élevé de tiges dans la placette ne permet généralement pas d'améliorer la qualité de l'information recherchée, étant donné que le travail devient plus complexe à réaliser sur le terrain et que les risques d'erreurs augmentent en conséquence. Dans ce cas, l'effort supplémentaire fourni aurait été mieux mis à profit en échantillonnant davantage de placettes ayant une plus petite superficie, ce qui aurait permis d'augmenter la précision.

Pour estimer la distribution des tiges, la dimension de la placette est fixée en fonction de l'espacement théorique entre les arbres à un âge donné. Cet espacement varie en fonction de l'essence ou du groupe d'essences considéré. À titre d'exemple, pour les résineux en régénération, la densité visée est de 2 500 tiges/ha, ce qui correspond à une tige par 4 m^2. Conséquemment, la dimension de la placette pour le suivi de la régénération résineuse est de 4 m^2 (1,13 m de rayon), ce qui, sur la base de la densité théorique, permettrait d'avoir une tige par placette.



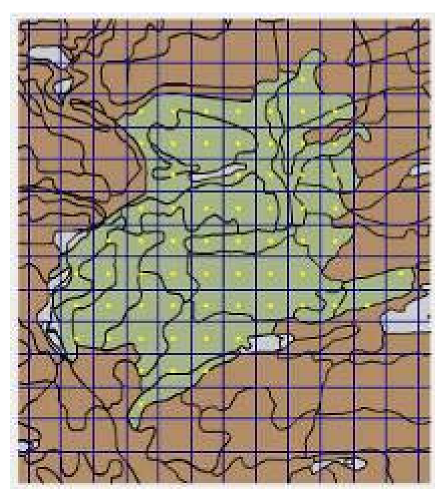
Si la densité visée pour le même type de peuplement résineux est fixée à 2 000 tiges/ha, les placettes devraient avoir une superficie de 5 m^2, soit 1,26 m de rayon. Selon le même principe, dans un peuplement résineux mature contenant 800 tiges/ha, un coefficient de distribution de 100 % signifie que l'on retrouve en moyenne une tige à tous les 12,5 m^2. C'est donc cette dimension de placette qui devrait être utilisée pour évaluer le coefficient de distribution des peuplements récoltés, ce qui correspond à des placettes de 2 m de rayon.

Pour de plus amples détails quant aux normes en vigueur (forêt du domaine public) relativement aux types de placettes requis pour les différents travaux sylvicoles, le lecteur est invité à se référer au document portant sur les méthodes d'échantillonnage (MRNF, 2004b).

1.5 Plan de sondage

Le plan de sondage fournit la localisation des placettes à mesurer sur la carte forestière pour un territoire à échantillonner.

1.5.1 Planification et réalisation – principes et pratique



Le plan de sondage permet de visualiser la méthode d'échantillonnage et la répartition géographique des placettes. Le taux d'échantillonnage variera en fonction de la superficie de l'aire d'intervention et de la variabilité du paramètre à estimer.

Le plan de sondage doit couvrir l'ensemble de la superficie. Il doit présenter des virées équidistantes couvrant l'ensemble de la superficie et les placettes doivent être distribuées systématiquement de long de chacune de ces virées. Pour les plantations ou regarnis, les virées doivent être orientées à ±75° par rapport aux lignes de plantation. Dans le cas des coupes commerciales, les virées doivent être perpendiculaires aux chemins de débardage.

Le nombre de placettes dépend de leur type et de leur dimension, qui sont définis en fonction des objectifs poursuivis, du type d'intervention et du type de peuplement à évaluer.

Afin de respecter les principes statistiques, la localisation des virées équidistantes et la distance de la première placette doivent être fixées de manière aléatoire.

De façon générale, il n'est pas nécessaire d'utiliser le même plan de sondage pour les estimations avant et après les interventions.

1.5.2 Délimitation du territoire

Le territoire à évaluer correspond généralement à une aire d'intervention d'une superficie maximale de 250 ha. Par définition, il s'agit de superficies devant faire l'objet d'un même traitement, réalisé au cours de la même année, dans la même unité d'aménagement forestier. De plus, ces superficies doivent être relativement homogènes. Il n'est cependant pas nécessaire que ce territoire soit d'un seul tenant.

Le point de départ de chacune des virées doit être indiqué sur la carte en précisant l'azimut, la distance entre les placettes, etc. De même, le point central de chaque placette, ou le point de départ de chaque grappe de placettes, ou toute autre information pertinente doivent être indiqués sur la carte. Si le GPS est utilisé, les coordonnées de localisation doivent accompagner la carte.

1.5.3 Implantation des placettes

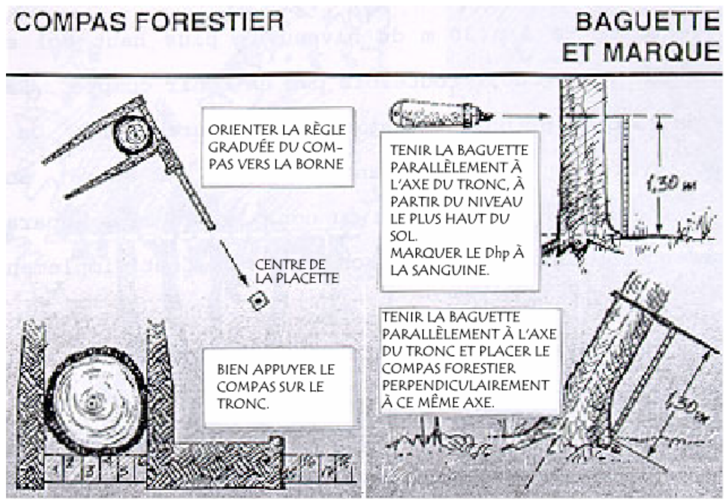
Les placettes doivent être localisées sur le terrain conformément au plan de sondage. Les distances entre les placettes doivent être respectées. Les placettes ne devraient être déplacées en aucun cas, puisque ceci pourrait introduire un biais dans les résultats. L'azimut devrait aussi être respecté.

Dans le cas où les placettes sont utilisées pour plusieurs estimations successives, la localisation des placettes ne devrait pas être matérialisée sur le terrain pour éviter de biaiser la réalisation des travaux.

1.5.4 Outils

Tous les instruments de mesure doivent être soigneusement vérifiés afin de s'assurer de la rectitude des mesures effectuées et éviter l'introduction de biais.

1.5.5 Importance de la qualité de la cueillette des observations



Il est essentiel de réduire le plus possible les erreurs systématiques qui sont différentes de l'erreur d'échantillonnage qui elle, est reliée à plusieurs facteurs dont notamment : la dimension de l'échantillon, la variabilité des unités d'échantillonnage et le ty pe d'échantillonnage utilisé (Rondeux, 1999). Les principales erreurs systématiques à éviter sont les erreurs de localisation des placettes, souvent liées à l'utilisation d'appareils de mesure présentant des erreurs systématiques différentes, les erreurs reliées à la réalisation des mesures ou la collecte d'informations pouvant être attribuables à des appareils défectueux, un manque de soin, ou un défaut de la vue, ainsi que les erreurs de calculs ou de procédures. Pour s'assurer de la validité des résultats, ce type d'erreurs doit faire l'objet de précautions attentives, puisqu'une fois qu'elles sont commises, il est ensuite difficile de les détecter, de les apprécier et de les éliminer.

2. Présentation de la mathématique liée à l'échantillonnage

Afin d'illustrer les différentes compilations statistiques, divers exemples seront utilisés.

Dans un secteur forestier feuillu mature de 150 ha destiné à une coupe partielle, deux échantillons (A et B) de 30 placettes (prisme de facteur 2), ayant fait l'objet de plans de sondage distincts, serviront de référence pour le calcul des variables suivantes :

- Surface terrière totale
- Surface terrière par essence
- Surface terrière par essence par classe de vigueur
- Pourcentage de prélèvement total
- Pourcentage de prélèvement par essence

Dans un jeune peuplement résineux de 230 ha destiné à des travaux dits précommerciaux, l'échantillon C est constitué de 460 placettes de 4 m^2 regroupées en grappe de 10 placettes. Cet échantillon sert de référence pour le calcul des variables :

- Nombre de tiges/ha
- Nombre de tiges résineuses/ha
- Distribution des tiges toutes essences
- Distribution des tiges résineuses

2.1 Lexique des symboles statistiques

Où : y = Une variable (volume, nombre de tiges, surface terrière, etc)

x = Une autre variable (volume, nombre de tiges, surface terrière, etc)

y_i = La i^e observation de la variable y

x_i = La i^e observation de la variable x

$\overline{Y}$ = Moyenne de l'échantillon de la variable y

$\hat{R}$ = Estimation quotient de 2 variables indépendantes

S^2 = Variance de l'échantillon

S = Écart-type de l'échantillon

$S_{\bar{Y}}^2$ = Variance de la moyenne

$S_{\bar{Y}}$ = Erreur standard de la moyenne

CV = Coefficient de variation (%)

L_i = Limite inférieure de l'intervalle de confiance

L_s = Limite supérieure de l'intervalle de confiance

e = Erreur relative (imprécision) (%)

P = Précision (%)

$t_{\nu,1-\alpha/2}$ = Quantile de la loi de t de « Student » avec une probabilité $(1-\alpha/2)$ 100$^{\text{ième}}$ percentile et ν degrés de liberté; si $\nu \rangle 120$ utilisé t = 1,96

n = Nombre d'observations (placettes) constituant l'échantillon

N = Nombre d'observations possible dans la population

ν = Degré de liberté ($n-1$)

α = Niveau de signification, habituellement égale à 0,05

$1-\alpha$ = Niveau de probabilité, habituellement égale à 0,95

$\bar{x}$ = Moyenne de l'échantillon de la variable x

d = Erreur d'échantillonnage

f = Taux d'échantillonnage

2.2 Distribution de fréquences

Normalement, avant toute compilation statistique, les données récoltées devraient être examinées afin de détecter des problèmes éventuels. Un diagramme de fréquences permet de vérifier visuellement la distribution des observations. Ce diagramme, qui montre le nombre d'observations par valeur mesurée, peut prendre la forme d'un histogramme tel qu'illustré à la figure 4 pour le nombre de pins blancs par classe de hauteur dans un peuplement de 80 ans. Pour ce faire, les données sont d'abord classées, puis regroupées dans un nombre relativement restreint de classes.

Les caractéristiques de l'histogramme sont les suivantes :

- Il est constitué d'un ensemble de rectangles juxtaposés;
- La largeur de chacun des rectangles correspond à la largeur des classes et est identique pour toutes les classes;
- La hauteur des rectangles correspond à la fréquence des classes;
- Il est très utile pour condenser les données lorsque le nombre d'observations est important;
- Puisque le nombre de classes (barres verticales) retenu pour tracer l'histogramme peut dans certaines conditions affecter la forme générale de la figure, l'histogramme peut être quelques fois trompeur et donner une représentation distordue de la distribution des données.

Si le nombre d'observations est assez grand, le diagramme de distribution de fréquence de l'échantillon sera représentatif de celui de la population.

Si le diagramme de fréquence présente deux formes de cloches (figure 5), cela signifie que l'on est en présence de deux populations distinctes ayant fait l'objet du même échantillonnage. Dans ce cas, il serait nécessaire d'examiner les résultats pour tenter d'identifier la cause de ces différences, vérifier si elles correspondent à des portions du territoire, et si c'est le cas, refaire la stratification du territoire, puis refaire les compilations. En pareilles circonstances, il arrive cependant que la taille d'un des échantillons, voire les deux, soit insuffisante, ce qui peut imposer alors la nécessité d'ajouter des points d'observation pour compléter l'échantillonnage.

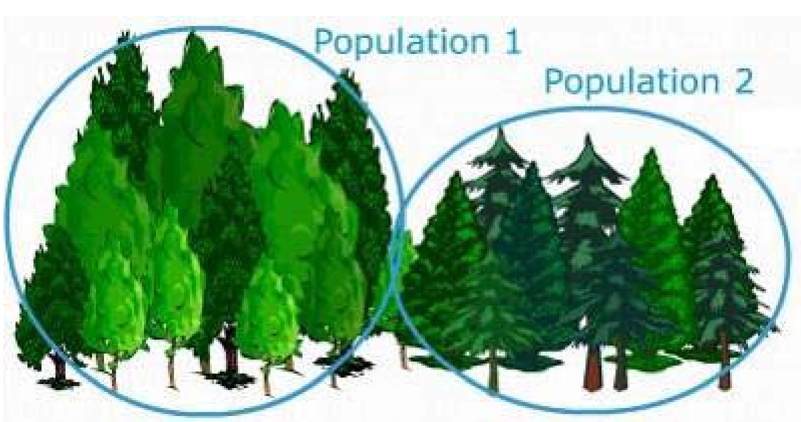


Figure 4 Exemple d'histogramme de distribution de la hauteur des pins blancs (adapté de Baillargeon et Rainville, 1976)

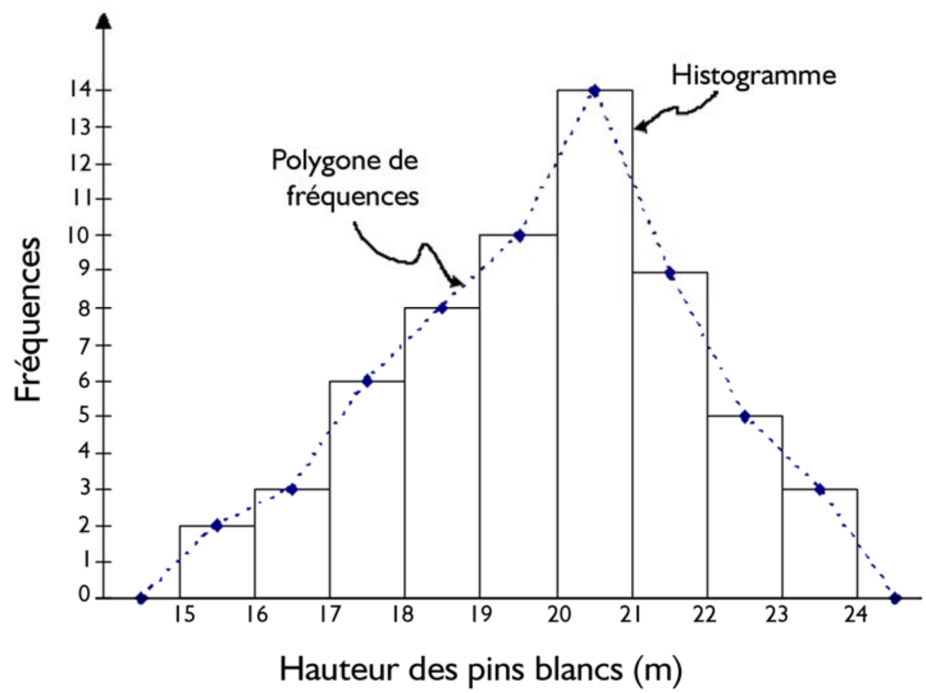
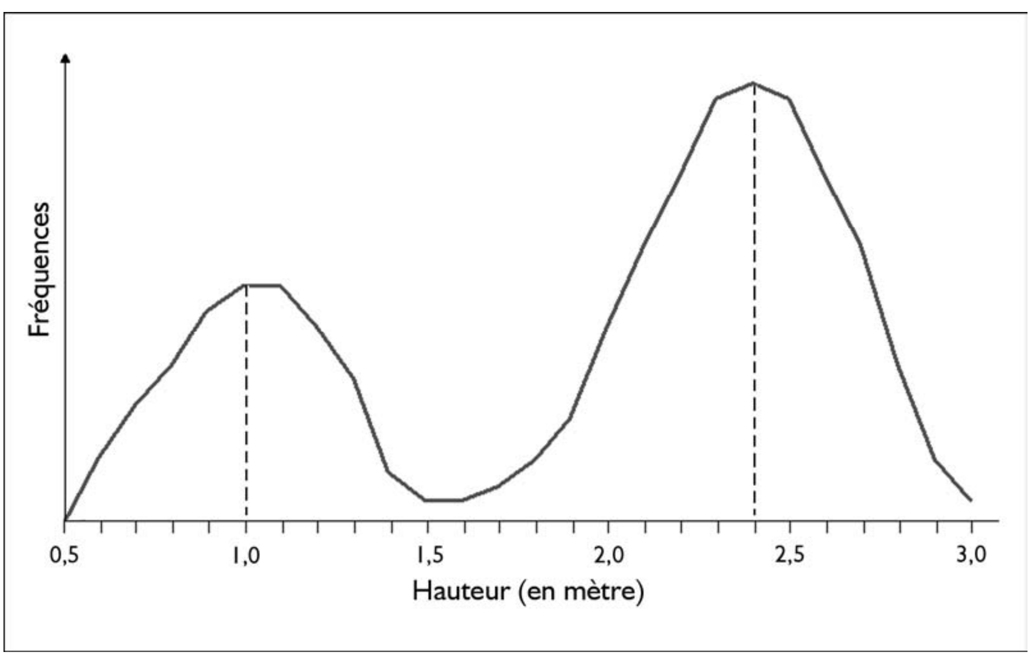
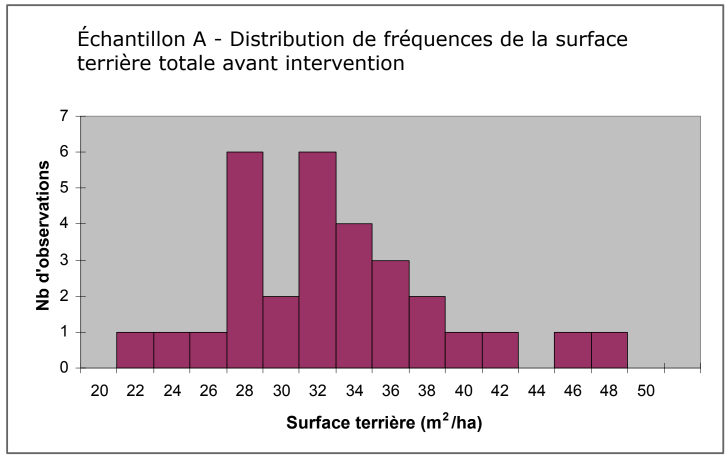


Figure 5 Exemple de distribution de deux populations (tiré de OIFQ, 1996)

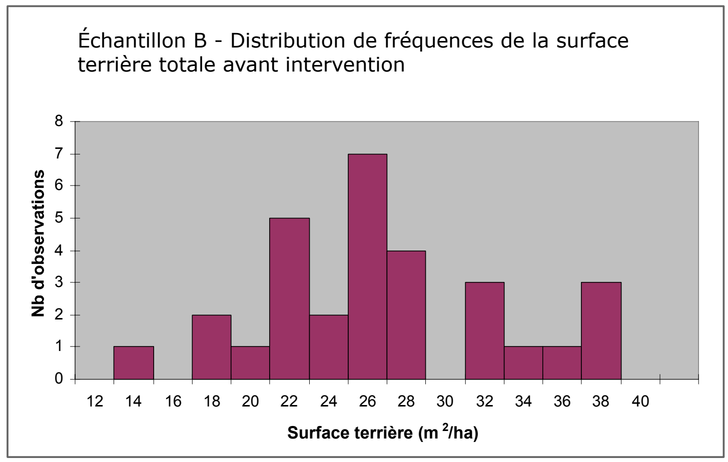


2.2.1 Exemple de distribution de fréquences pour les échantillons A, B et C

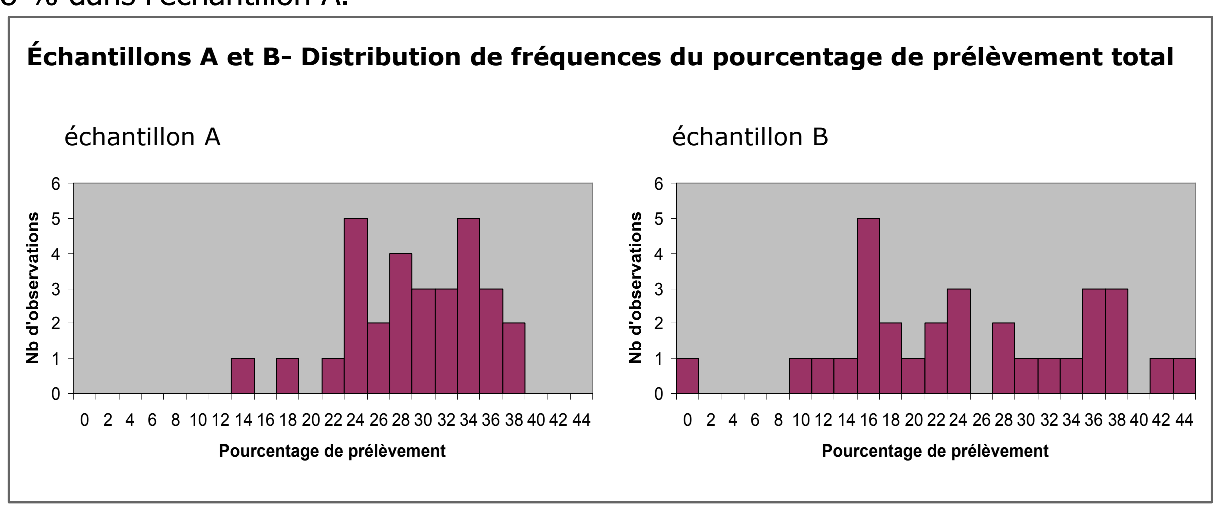
On constate, pour l'échantillon A, un allongement de l'aile droite de la distribution et une surabondance de placettes ayant une surface terrière égale à 28 qui compense une sous-abondance de placettes de 26 et de 30.



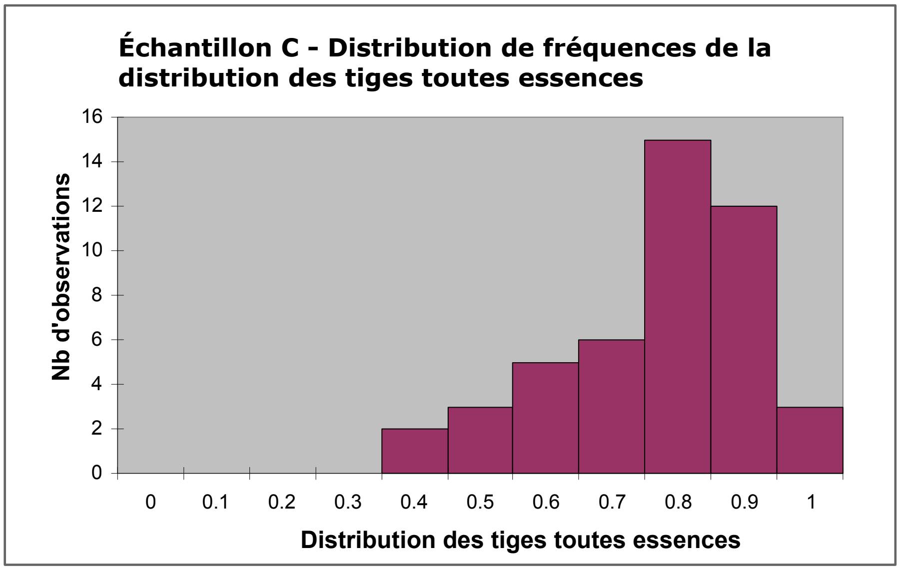
Dans le cas de l'échantillon B, on constate que l'aile de la distribution est plus longue à gauche qu'à droite, 4 placettes ont une surface terrière < 22 m^2/ha et aucune placette avec une surface terrière ≥ 40 m^2/ha.



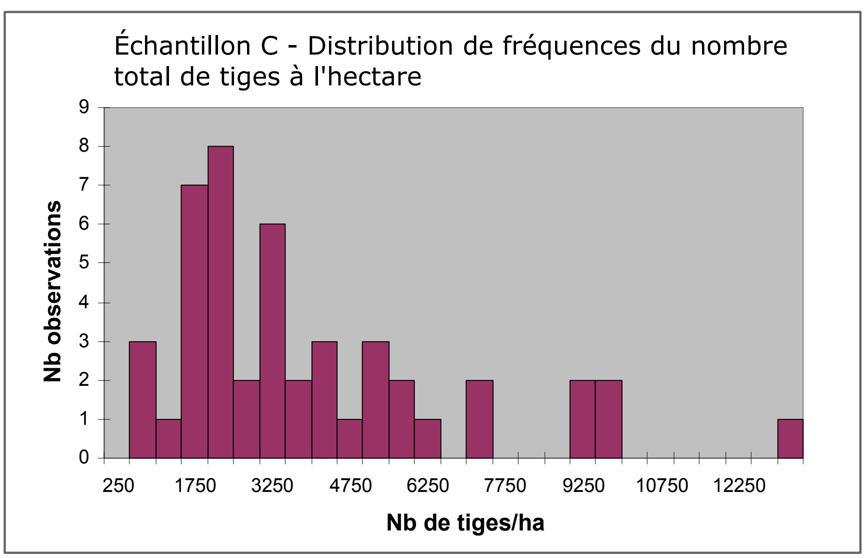
Pour ce qui est du pourcentage de prélèvement total, on constate que l'étendue des données est beaucoup plus importante dans l'échantillon B, puisque les pourcentages de prélèvement totaux obtenus varient entre 0 et 44 %, alors qu'ils varient entre 14 et 38 % dans l'échantillon A.



Dans le cas de l'échantillon C (jeune peuplement résineux), on observe pour la distribution des tiges toutes essences que la majorité des grappes de placettes ont une valeur de 80 et 90 %.



Dans le cas du nombre totales de tiges/ha de l'échantillon C, il y a une fréquence élevée de faibles densités (de moins de 4 000 tiges/ha), alors qu'à l'autre bout du graphique, on constate une fréquence anormalement élevée de densités supérieures à 9 250 tiges/ha. Dans ce cas, la forme de la distribution suggère l'existence d'une hétérogénéité dans les conditions du milieu qui constitue un facteur susceptible d'induire de telles différences. Pour vérifier cette hypothèse, il faudrait examiner la répartition spatiale des données, afin de confirmer si nous sommes en présence de 2 populations différentes. Il est possible qu'une subdivision du territoire, qui aurait pu prendre en compte les variables écologiques, ait conduit à une meilleure caractérisation des aires en régénération, étant donné que le potentiel de régénération est relié aux conditions du milieu.

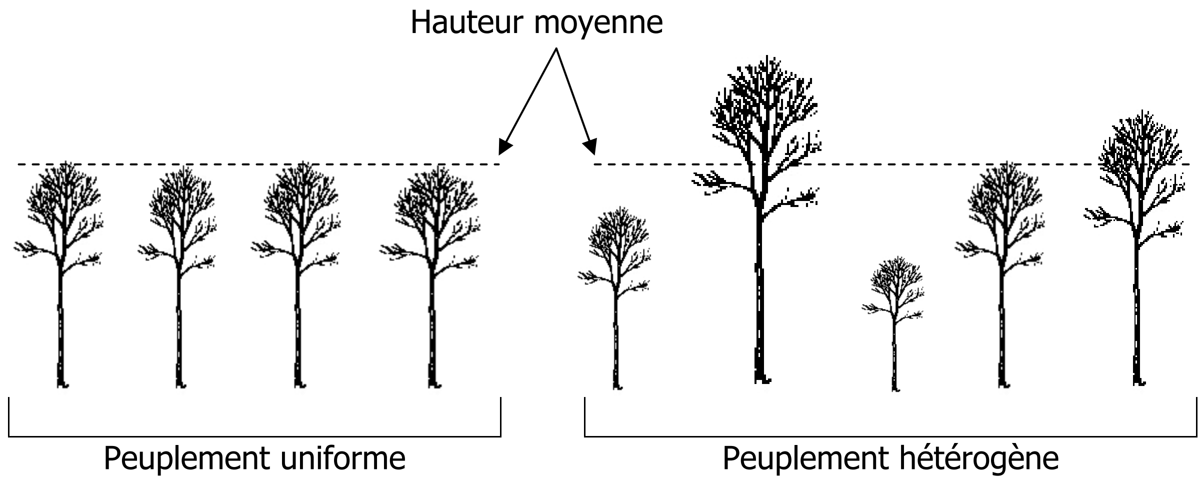


2.3 Mesures de tendance centrale et de dispersion

2.3.1 Moyenne

La moyenne arithmétique est la somme de toutes les observations relatives à une même variable divisée par le nombre total d'observations. La moyenne est la statistique la plus communément utilisée pour caractériser un échantillon. Toutefois, une moyenne ne suffit pas à elle seule pour qualifier correctement une variable (Bédard, 1996). La figure 6 illustre bien que des moyennes similaires de la hauteur de deux peuplements peuvent caractériser deux paysages tout à fait différents et ainsi, des écarts-types différents. Il est donc nécessaire d'utiliser d'autres compilations statistiques afin de qualifier la valeur moyenne de la variable étudiée.

Figure 6 Hauteurs moyennes identiques de deux peuplements différents



La moyenne d'un échantillon aléatoire se calcule de la façon suivante (se référer au début du chapitre pour la légende des symboles mathématiques) :

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{(formule 2)}$$

2.3.2 Quotient de deux variables aléatoires

L'utilisation du quotient est nécessaire pour l'évaluation d'un ratio, d'un pourcentage ou d'une proportion, lorsque l'élément de référence (dénominateur) est différent pour chacune des placettes. À titre d'exemple, l'évaluation du pourcentage de prélèvement, comme dans le cas d'une coupe partielle en peuplement feuillu, exige que l'on connaisse pour chacune des placettes, le nombre de tiges prélevées et le nombre total de tiges, ces deux nombres pouvant être différents dans chacune des placettes, c'est l'évaluation par quotient qui devra être utilisée.

Pour un échantillon aléatoire, le quotient moyen se calcule selon :

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \qquad \text{(formule 3)}$$

2.3.3 Écart-type et variance de l'échantillon

Ce paramètre exprime la dispersion de la valeur des observations par rapport à la moyenne de l'échantillon. Tel qu'illustré sur la figure 6, le peuplement constitué d'arbres uniformes a un écart-type petit, alors que le peuplement constitué d'arbres hétérogènes possède un écart-type plus élevé.

a) Variance et écart-type de l'échantillon de taille n :

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{\mathrm{Y}})^2}{n-1} \qquad \text{(formule 4)}$$

$$S = \sqrt{S^2} \qquad \text{(formule 5)}$$

b) Exemple de calcul et illustration

À titre d'exemple, voici deux échantillons d'arbres dont on a mesuré les hauteurs.

Tableau 1 Fréquence observée de la hauteur des arbres de deux échantillons (tiré de OIFQ, 1996)

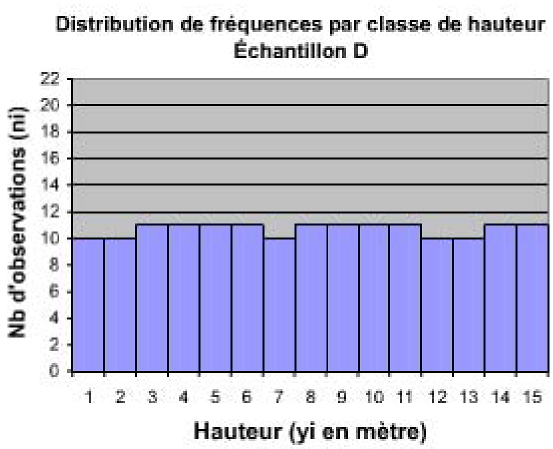
Hauteur (y_i) en mètre	Fréquence observée (n_i)	
	Échantillon D	Échantillon E
1	10	18
2	10	9
3	11	6
4	11	2
5	11	10
6	11	19
7	10	14
8	11	6
9	11	4
10	11	10
11	11	11

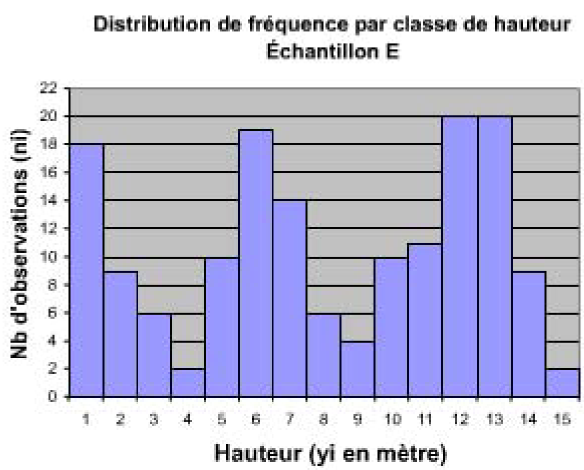
Hauteur (y_i) en mètre	Fréquence observée (n_i)	
	Échantillon D	Échantillon E
12	10	20
13	10	20
14	11	9
15	11	2

Les échantillons D et E ont :

- le même effectif ($n = 160$)
- la même moyenne ($\overline{Y} = 8{,}03$)
- le même écart-type ($S = 4{,}31$)
- les mêmes valeurs extrêmes (1 et 15), donc la même étendue (14).

Si l'on ne considère que la moyenne et l'écart-type, on peut conclure que ces deux échantillons d'arbres sont semblables du point de vue de la hauteur. Or, cette conclusion est erronée, comme le démontre leur graphique de distribution de fréquences.





Il est toujours intéressant de commencer l'analyse des données par l'examen des fréquences ou d'un diagramme, car cela permet de repérer les irrégularités, d'orienter la suite de l'analyse ou même de tirer certaines conclusions. De plus, l'analyse des fréquences permet souvent de savoir si la moyenne est vraiment représentative de la population.

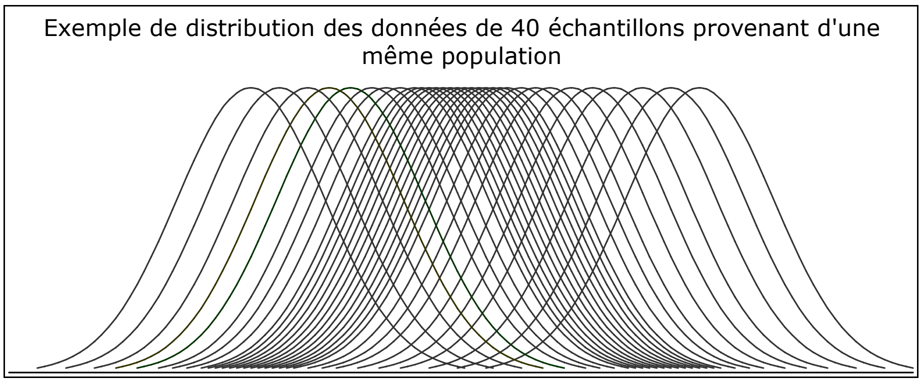
c) Variance et écart-type du quotient de deux variables aléatoires dans un échantillon de taille n :

$$S^2 = \frac{1}{\bar{x}^2}\left[\frac{\sum_{i=1}^{n} y_i^{\,2} + \hat{R}^2 \sum_{i=1}^{n} x_i^{\,2} - 2\hat{R}\sum_{i=1}^{n} x_i y_i}{n-1}\right] \qquad \text{(formule 6)}$$

$$S = \sqrt{S^2}$$

2.3.4 Erreur standard et variance de la moyenne

Cette statistique exprime la variation de la valeur moyenne de la population étudiée. Elle mesure la dispersion des moyennes d'échantillons autour de la moyenne de la population.



Variance de la moyenne $$S_{\bar{Y}}^2 = \frac{S^2}{n}(1-f) \qquad \text{(formule 7)}$$

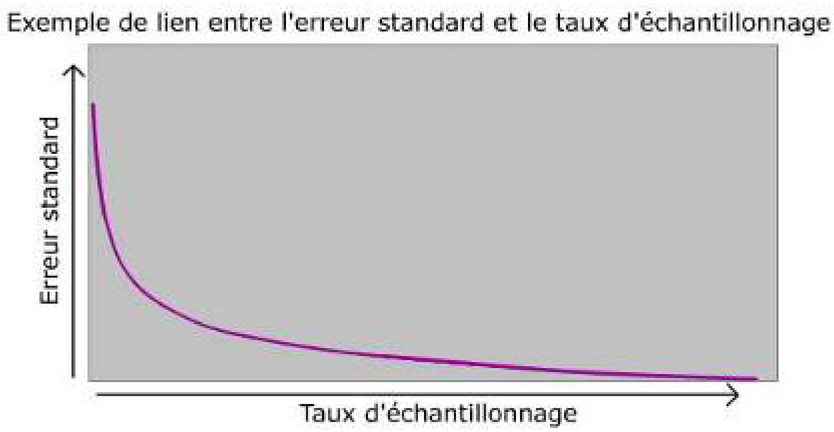
Erreur standard de la moyenne $$S_{\bar{Y}} = \sqrt{S_{\bar{Y}}^2} = \sqrt{\frac{S^2}{n}(1-f)} = S\sqrt{\frac{(1-f)}{n}} \qquad \text{(formule 8)}$$

Dans l'hypothèse d'une population finie, il y a lieu de corriger la variance de la moyenne en la multipliant par le facteur $[(N-n)/N]$, *N* représentant le nombre total possible d'observations composant la population. Ce facteur de correction tient compte du fait qu'une estimation a une signification différente ou un poids différent selon l'intensité d'échantillonnage adoptée. Une estimation réalisée à partir d'une unité sur 100, par exemple, contient en principe plus d'informations que l'estimation réalisée à partir de 10 unités sur 10 000. Il vient donc, pour une population finie, que :

$$S_{\bar{Y}}^2 = \frac{S^2}{n}\left[\frac{N-n}{N}\right] = \frac{S^2}{n}(1-f) \qquad \text{(formule 7)}$$

On observera que la différence entre cette variance et celle d'une population infinie est d'autant plus importante que la fraction sondée $f = n/N$ est élevée. En pratique, cependant, ce facteur de correction sera souvent très petit par rapport au nombre total possible N. On pourra l'ignorer pour des valeurs de $(N - n) / N \geq 0{,}95$ ou de $n \leq N / 20$, soit pour des taux d'échantillonnage inférieurs ou égaux à 5 % (Mendenhall *et al.* 1971).

$$S^2_{\bar{Y}} = \frac{S^2}{n} \qquad \text{(formule 9)}$$



La taille de l'échantillon (n) a un effet très significatif sur l'erreur standard, au fur et à mesure que la taille de l'échantillon augmente, l'erreur standard diminue. Ainsi, la dispersion de la distribution d'échantillonnage de la moyenne s'amenuise quand n augmente.

Afin d'aller plus loin dans l'utilisation et l'interprétation des statistiques de l'écart-type de l'échantillon et de l'erreur standard de la moyenne, il est nécessaire d'aborder la notion de probabilité.

2.4 Notion de probabilité

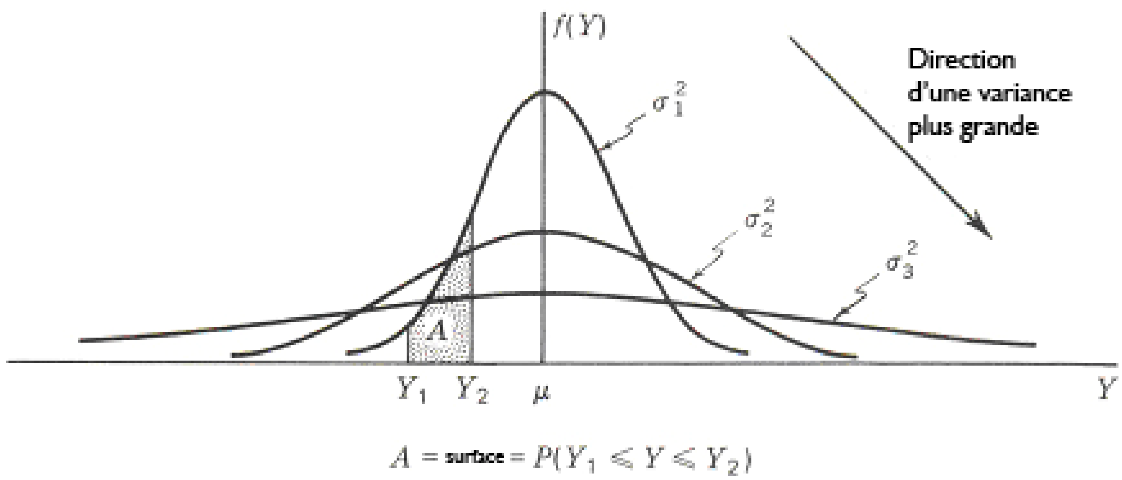
On peut dire que les évènements communs ont de fortes chances de survenir, alors que ceux plus rares ont moins de chances de se produire. Les statisticiens traduisent ce risque de se concrétiser, en utilisant un nombre situé entre 0 et 1 (ou entre 0 % et 100 %) pour indiquer avec précision jusqu'à quel point cet évènement est probable ou improbable (Steel and Torrie, 1980). Dans le cas où une évaluation est faite à partir d'un échantillon, donc sur la base d'une information incomplète, il peut arriver que l'on soit amené à tirer des conclusions erronées, puisque le hasard joue un rôle. Les statistiques nous fournissent les outils permettant de dire dans quelle mesure, en moyenne, les conclusions tirées sont correctes.

De façon universelle, le niveau de confiance de 95 % est utilisé comme référence.

2.4.1 Distribution normale

La distribution normale s'applique aux variables continues et possède une courbe symétrique en forme de cloche (voir figure 7). La distribution consiste en la fréquence des observations en fonction des valeurs d'un paramètre donné. La localisation du centre de la courbe est fournie par la moyenne de la population μ et l'étalement de la cloche dépend de la variance de la population σ^2. La courbe possède deux points d'inflexion situés à $\mu \pm \sigma$. L'importance de la superficie A entre Y_1 et Y_2 indique la probabilité qu'un individu choisi au hasard se situe entre Y_1 et Y_2, ce qui se traduit en termes mathématiques par $P(Y_1 \leq Y \leq Y_2) = A$. Il est possible de comparer différentes courbes normales en standardisant les valeurs observées. C'est la loi normale centrée réduite qui permet l'utilisation d'une table unique (table de distribution normale). Les caractéristiques d'une variable standardisée sont une moyenne (μ) égale à 0 et une variance (σ^2) égale à 1. On dira alors que la variable z suit une N(0,1).

Figure 7 Courbes de distribution normale (tiré de Steel et Torrie, 1980)



De façon générale, on utilise la table de Z plutôt que la courbe pour obtenir les probabilités associées à une distribution normale (voir tableau 2). Cette table est bâtie pour connaître la probabilité associée à une moyenne μ = 0 et une variance σ^2 = 1. Toutefois, cette table peut être utilisée pour connaître la probabilité associée à n'importe quelle variable aléatoire normale en autant que l'on connaisse la moyenne et la variance de cette variable. Il est possible de standardiser une variable aléatoire normale en calculant la valeur de z :

$$Z = \frac{Y - \mu}{\sigma} \qquad \text{(Formule 10)}$$

Tableau 2 Probabilité que la variable aléatoire de $Z=(Y - \mu)/\sigma$ soit plus grande que la valeur indiquée dans la marge de gauche à laquelle il faut ajouter les centièmes inscrits dans le haut de chaque colonne.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,9	0,0000									

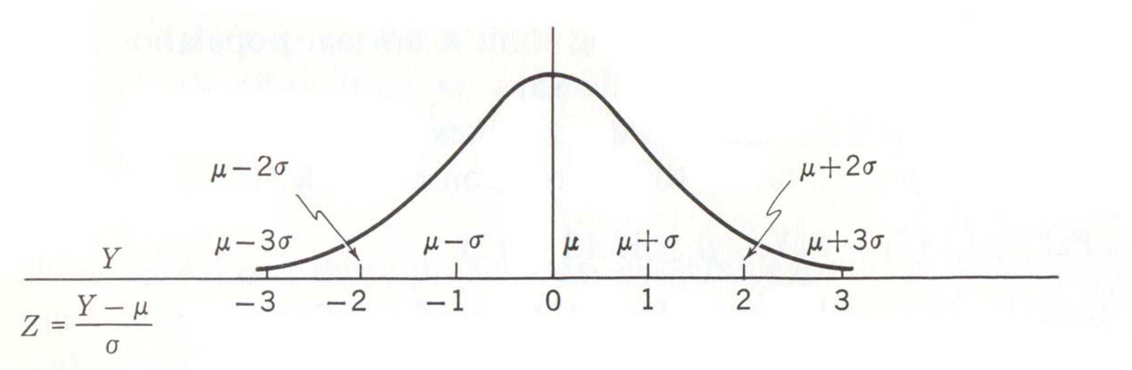
Selon la loi de la distribution normale, en utilisant la table de Z ou en se référant au graphique de la figure 8, il est possible d'affirmer que :

la probabilité qu'une observation se situe à plus ou moins une fois l'écart-type autour de la moyenne est de 68 %; $P(-1\sigma \leq \mu \leq +1\sigma)$ = 1 – 2 (0,1587) = 0.6826

la probabilité qu'une observation se situe à plus ou moins deux fois l'écart-type autour de la moyenne est de 95 %; $P(-2\sigma \leq \mu \leq +2\sigma)$ = 1 – 2 (0,0228) = 0.9544

la probabilité qu'une observation se situe à plus ou moins trois fois l'écart-type autour de la moyenne est de 99 %; $P(-3\sigma \leq \mu \leq +3\sigma)$ = 1 – 2 (0,0013) = 0.9974

Figure 8 Relation entre *Y* et *Z* pour calculer la probabilité associée à la distribution normale (tiré de Steel et Torrie, 1980)



2.4.2 Distribution normale des observations de l'échantillon

En fonction de la distribution normale, nous pouvons par exemple, répondre à la question suivante : quelle est la probabilité d'obtenir une observation plus grande qu'une valeur donnée? Considérons un échantillon pris au hasard dans une population ayant une distribution normale dont la moyenne μ = 5 et la variance σ^2 = 4 (tiré de Steel and Torrie, 1980). Quelle est la probabilité d'obtenir une valeur $Y \geq 7{,}78$?

$\mu = 5$

$\sigma^2 = 4$ et $\sigma = 2$

Il faut ensuite utiliser la formule 10, avec :

$$Z = \frac{Y - \mu}{\sigma} = \frac{7{,}78 - 5}{2} = 1{,}39$$

On obtient, en consultant le tableau 1, $P(Z \geq 1{,}39) = 0{,}0823$, ce qui correspond à $P(Y \geq 7{,}78) = 0{,}0823$, soit 8 chances sur 100 que l'on obtienne une valeur plus grande que 7,78.

Il est aussi possible de répondre à la question : quelle est la probabilité d'obtenir une observation pour laquelle la valeur mesurée se situerait entre deux valeurs données? Considérons un échantillon pris au hasard dans une population ayant une distribution normale dont la moyenne μ = 5 et la variance σ^2 = 4. Quelle est la probabilité d'obtenir une valeur entre 2,22 et 7,78 ($2{,}22 \leq Y \leq 7{,}78$)?

Il est possible d'affirmer que :

$$\begin{aligned} P(2{,}22 \leq Y \leq 7{,}78) &= 1 - P(Y \text{ soit en dehors de cet intervalle}) \\ &= 1 - [\, P(Y \leq 2{,}22) + P(Y \geq 7{,}78) \,] \\ &= 1 - [\, P(Z \leq -1{,}39) + P(Z \geq 1{,}39) \,] \\ &= 1 - [\, P(Z \geq 1{,}39) + P(Z \geq 1{,}39) \,] \\ &= 1 - (0{,}0823 + 0{,}0823) = 0{,}835 \end{aligned}$$

Dans ce cas, il y a donc 84 % des chances que la valeur d'une observation se situe entre 2,22 et 7,78.

2.4.3 Distribution normale des moyennes d'échantillonnages

Le même principe utilisé pour la distribution des observations s'applique pour la distribution des moyennes des échantillons tirés d'une population normale. Le calcul de la valeur de Z devient :

$$Z = \frac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} \qquad \text{(Formule 11)}$$

où : $\overline{Y}$ est la moyenne de l'échantillon et $\sigma_{\overline{Y}}$ son écart-type.

Ainsi, nous pouvons par exemple, répondre à la question suivante : quelle est la probabilité d'obtenir une moyenne plus grande qu'une valeur donnée? Considérons un échantillon pris au hasard de *n* = 16 observations, dans une population ayant une distribution normale dont la moyenne μ = 10 et la variance σ^2 = 4 (tiré de Steel and Torrie, 1980). Quelle est la probabilité d'obtenir une moyenne $\overline{Y} \geq 11$?

$$\mu_{\overline{Y}} = \mu = 10$$

$$\sigma^2_{\overline{Y}} = \frac{\sigma^2}{n} = 4/16 = 1/4 \qquad \text{et} \qquad \sigma_{\overline{Y}} = \sqrt{1/4} = 1/2$$

Il faut ensuite utiliser le tableau 1, avec :

$$Z = \frac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}} = \frac{11-10}{1/2} = 2$$

On obtient alors $P(Z \geq 2) = 0{,}0228$, ce qui correspond à $P(\overline{Y} \geq 11) = 0{,}0228$, ce qui signifie que dans cette population, il y a 2,3 % (0,0228) de chance qu'un échantillon de taille n=16 ait une moyenne supérieure ou égale à 11.

Compte tenu de la symétrie de la distribution normale, à partir du même exemple, il est possible de considérer la valeur absolue de *Z* symbolisée par |*Z*|. Donc, P(|*Z*| ≥ 2) = 2 X 0,0228 = 0,0456. Ainsi :

$$P(-2 \leq Z \leq 2) = P(9 \leq \overline{Y} \leq 11) = 1 - 0{,}0456 = 0.9544$$

En conséquence, une moyenne inférieure à 9 ou supérieure à 11 ne sera obtenue qu'environ 5 fois sur 100 (0,0456), ce qui est considéré comme inhabituel, et 95 % (0,9544) du temps, la moyenne se situera dans l'intervalle entre 9 et 11.

2.4.4 Distribution *t* de Student

Étant donné qu'à partir d'un petit échantillon, l'utilisation de s pour estimer σ ne s'avère pas fiable pour le calcul de Z (Steel et Torrie, 1980), une table alternative, soit la table de *t*, a été construite. Ainsi, la variable *t* de Student, qui représente les statistiques pour un échantillon aléatoire tiré d'une population ayant une distribution normale est devenue :

$$t_{\nu,1-\alpha/2} = \frac{\overline{Y} - \mu}{S_{\overline{Y}}} = \frac{\overline{Y} - \mu}{\sqrt{\frac{S^2}{n}(1-f)}} \qquad \text{(Formule 12)}$$

Cette distribution s'applique lorsque l'échantillon de taille n est petit (*n*<30) et que la variance de la population est inconnue (Baillargeon et Rainville, 1976). Pour des échantillons dont la taille se situe entre 30 et 50, σ^2 est estimé par S^2 et la distribution de la statistique *t* se rapproche d'une loi normale, alors que lorsque la taille de l'échantillon est supérieure à 50 le *t* de Student sera similaire au *Z* de la loi normale.

Lorsqu'on procède par échantillonnage, σ^2 est estimé par S^2. L'utilisation du *t* de « Student » implique l'utilisation de deux statistiques, $\bar{Y}$ et $s_{\bar{Y}}$.

La densité de probabilité d'une loi de *t* de « Student » avec $\nu = 15$ degrés de liberté est symétrique, telle qu'illustrée à la figure 9 Elle a aussi une forme de cloche, mais elle est plus aplatie que celle de *Z*. Elle se situe sous celle de *Z* dans le centre, et est au-dessus de cette dernière dans les extrémités. Elle approche de la courbe normale (celle de *Z*) à mesure que *n* augmente.

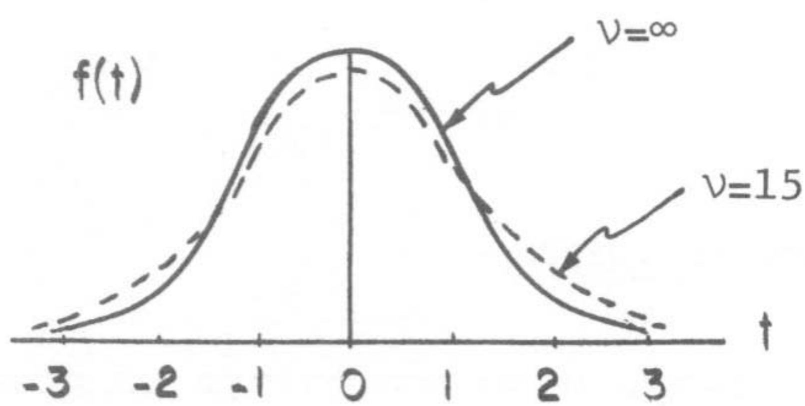


Figure 9. Courbe de *t* $(\nu = 15)$ et courbe de *Z* $(\nu = \infty)$ (tiré de Baillargeon et Rainville, 1976)

La table de t est fournie au tableau 3. Cette table s'utilise de la façon suivante : par exemple, pour un échantillon aléatoire de taille *n* égale à 16, donc (16-1)=15 degrés de liberté (ν), et un niveau de probabilité d'avoir une valeur numérique plus grande que celle de *t* fixée à 95 % (0,05 de la ligne du haut), *t* prend la valeur de 2,131. Il s'agit de la valeur absolue (positive ou négative). Pour évaluer la probabilité d'avoir une valeur positive seulement ou négative seulement, il faut utiliser la ligne du bas. Ainsi, $P(t \leq -2{,}131) = 0{,}025 = P(t \geq 2{,}131)$.

Tableau 3 Valeur de *t* (adapté de Steel et Torie, 1980)

	Niveau de signification bilatéral d'une valeur numérique plus large que t								
df	**0,5**	**0,4**	**0,3**	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**	**0,001**
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,689
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,660
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291
df	**0,25**	**0,2**	**0,15**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,0005**
	Niveau de signification unilatéral d'une valeur positive plus large que t								

2.4.5 Autres distributions

En plus de la distribution normale, la distribution exponentielle et la distribution de Weibull servent à décrire des variables continues. Telle que décrite dans le manuel de foresterie (OIFQ, 1996), la distribution exponentielle est parfois utile pour décrire la fréquence des diamètres à hauteur de poitrine (DHP) ou celle de durées de vie. Cette distribution est asymétrique. La densité de probabilité d'une variable exponentielle atteint son maximum à $x = 0$. La distribution de Weibull est une généralisation de la distribution exponentielle. Elle est utilisée pour décrire les DHP qui présentent généralement une distribution fortement asymétrique.

Telles que décrites dans le manuel de foresterie (OIFQ, 1996), les variables discrètes peuvent être décrites par différentes distributions. La distribution de Bernoulli est obtenue lorsque la variable aléatoire X prend la valeur 1 (succès) avec la probabilité p, et la valeur 0 (échec) avec la probabilité 1 - p. (Ex.: au lancer d'une pièce de monnaie, $x = 0$ si l'on obtient face, et $x = 1$ si l'on obtient pile ; $p = 0{,}5$.). La distribution binomiale survient lorsque n expériences de Bernoulli sont réalisées, n étant un entier positif. La variable X correspond au nombre de succès obtenus parmi n essais indépendants avec probabilité constante à chaque essai (Ex. : au lancer de trois pièces de monnaie équilibrées, $p = 0{,}5$, $n = 3$, et X est le nombre de piles, donc $X(S) = (0,1,2,3)$.). Quant à la distribution de Poisson, elle s'applique parfois à des variables aléatoires dont la valeur est un nombre d'objets ou d'évènements par unité de surface, de volume ou de temps, quand ces objets ou évènements sont indépendants les uns des autres.

Nous vous référons au Manuel de foresterie (OIFQ, 1996), Baillargeon et Rainville (1976) et à Steel et Torrie (1980) pour plus d'information à ce sujet.

2.4.6 Test de normalité

Afin de vérifier la normalité, il est possible de procéder aux tests de symétrie (« skewness ») et d'aplatissement (« kurtosis ») tels que décrits dans diverses références (Sokal et Rohlf, 1969; Baillargeon et Rainville, 1976). Le premier de ces tests (« skewness ») s'attarde sur la symétrie, particulièrement au niveau des bouts de la courbe (figure 10), alors que le second (« kurtosis ») considère plutôt l'aplatissement de la courbe (figure 11).

Figure 10 Exemples de symétrie de la courbe (tiré de Baillargeon et Rainville, 1976)

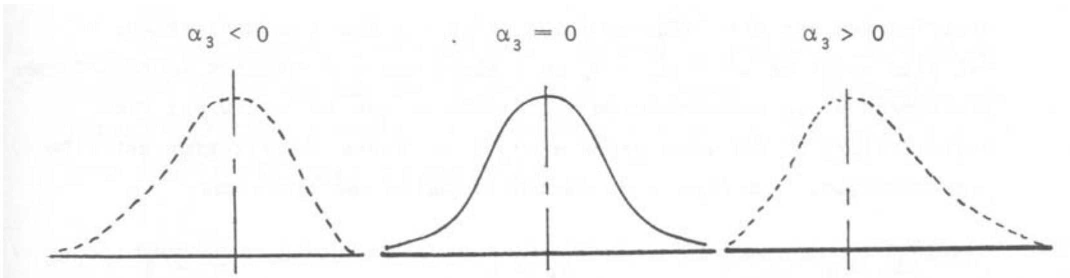
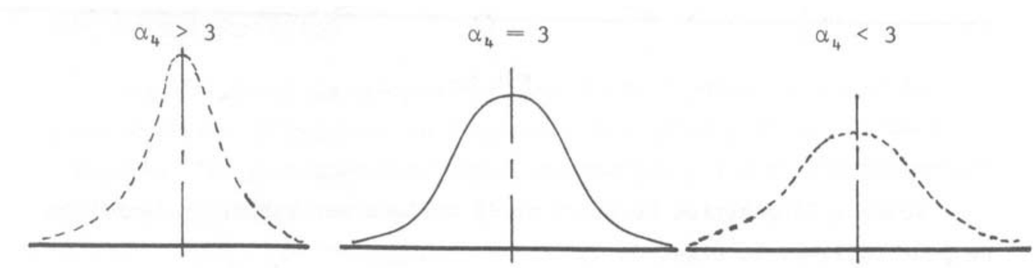


Figure 11 Exemples d'aplatissement de la courbe (tiré de Baillargeon et Rainville, 1976)



2.4.7 Application du théorème central-limite

Le théorème central-limite est la pierre angulaire de la théorie de l'échantillonnage. Le théorème central-limite stipule que la distribution d'échantillonnage des moyennes ($\bar{Y}$) tend vers une loi normale **quelle que soit la distribution des observations (Y) dans la population**, même si cette distribution a une forme en J. C'est précisément là la force du théorème. Bien sûr, quand la distribution d'Y est normale dans la population, la distribution d'échantillonnage d'$\bar{Y}$ est normale, quelle que soit la taille n de l'échantillon. Mais quand la distribution d'Y dans la population n'est pas normale, la distribution d'échantillonnage d'$\bar{Y}$ s'approche d'une distribution normale quand n augmente, et elle le fait d'autant plus vite que la distribution d'Y dans la population est proche d'une normale.

À titre d'exemple, une variable aléatoire Y qui suit une loi exponentielle de paramètre λ = 1 a une densité en J: $f(y) = e^{-y}$ où y est un nombre réel positif (Figure 12). Dans les Figures 13 a-d, la distribution d'échantillonnage de la moyenne de 1000 échantillons de tailles 5, 15, 30 et 50, respectivement, est représentée, ainsi que la densité d'une loi normale de même moyenne et de même écart-type que ceux de la distribution empirique des 1000 moyennes. On voit que les distributions d'échantillonnage de la moyenne des échantillons de taille 30 et 50 se rapprochent d'une normale, alors que la distribution des moyennes d'échantillons de taille 5 est encore assez asymétrique. De plus, un examen attentif des abscisses des Figures 13 a-d montre que la dispersion de la distribution d'échantillonnage d'$\overline{\mathrm{Y}}$ s'amenuise quand n augmente, comme le prévoit la théorie.

Figure 12 **Densité exponentielle de moyenne** $\lambda^{-1} = 1$

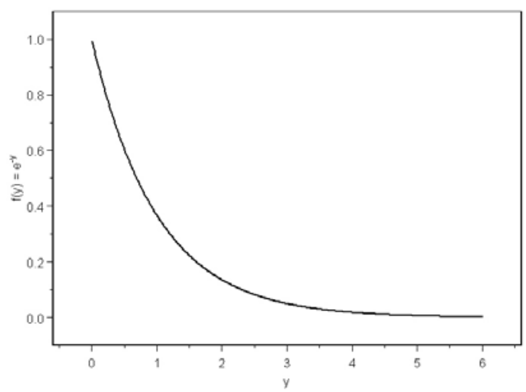
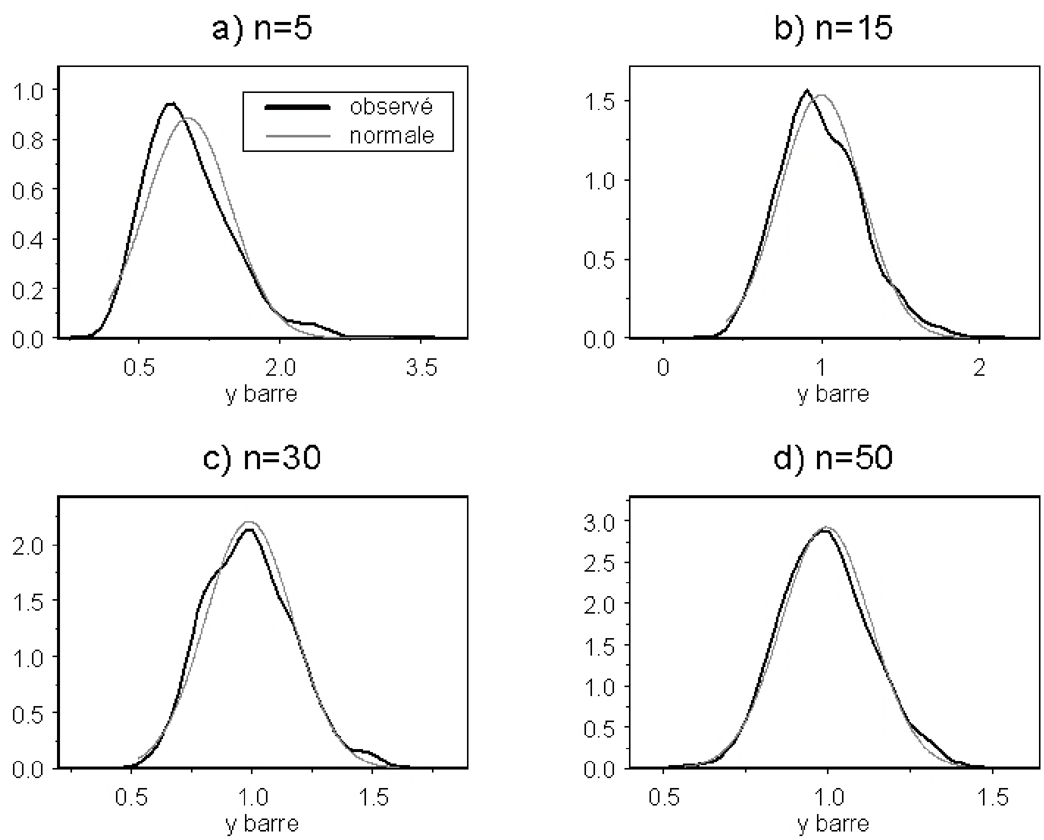


Figure 13 Distribution de la moyenne de 1000 échantillons de taille n

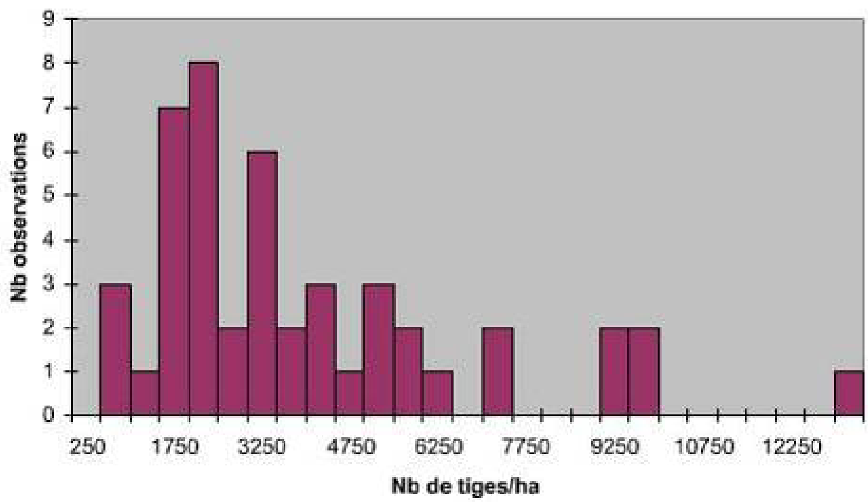


2.4.8 Exemple de distribution de fréquences de l'échantillon C

Pour le nombre de tiges totales/ha de l'échantillon C, la distribution de fréquences des observations s'éloigne de la normalité, ce qui suggère d'augmenter la taille de l'échantillon pour faire en sorte que la distribution des moyennes d'échantillon s'approche de la normalité.

Il est important de rappeler que l'application du théorème central limite (loi de la distribution normale) fait en sorte que, même si la distribution de la population-mère n'est pas normale, la distribution d'échantillonnage des moyennes provenant d'échantillonnages aléatoires, tend vers la normale à mesure que le nombre d'observations augmente (Baillargeon et Rainville, 1976; Steel and Torrie, 1980). Pour appliquer ce théorème, il n'y a aucune restriction sur la forme de la distribution de la population-mère, en autant que la taille de l'échantillon soit assez grande ($n \geq 30$).

Échantillon C - Distribution de fréquences du nombre total de tiges à l'hectare

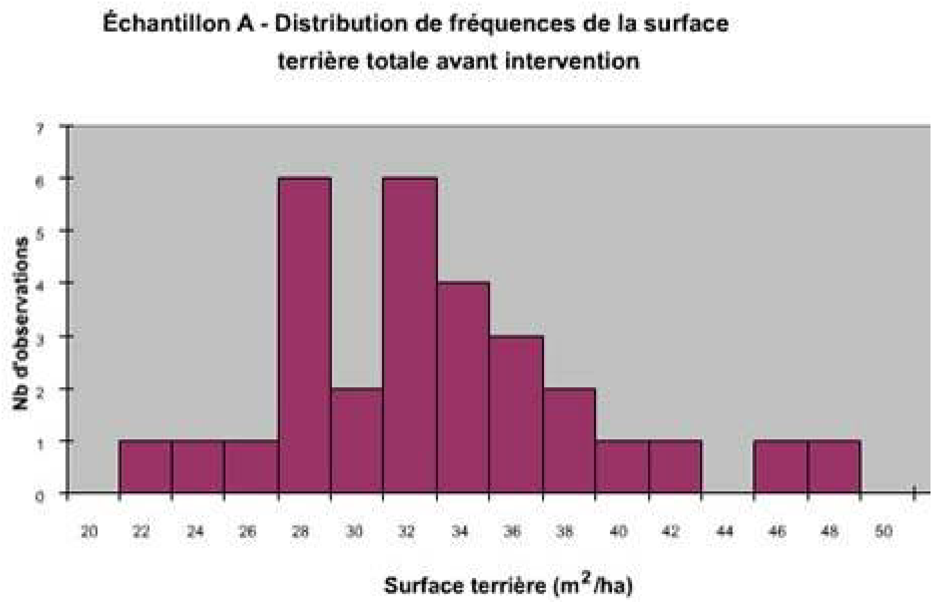


2.5 Autres statistiques

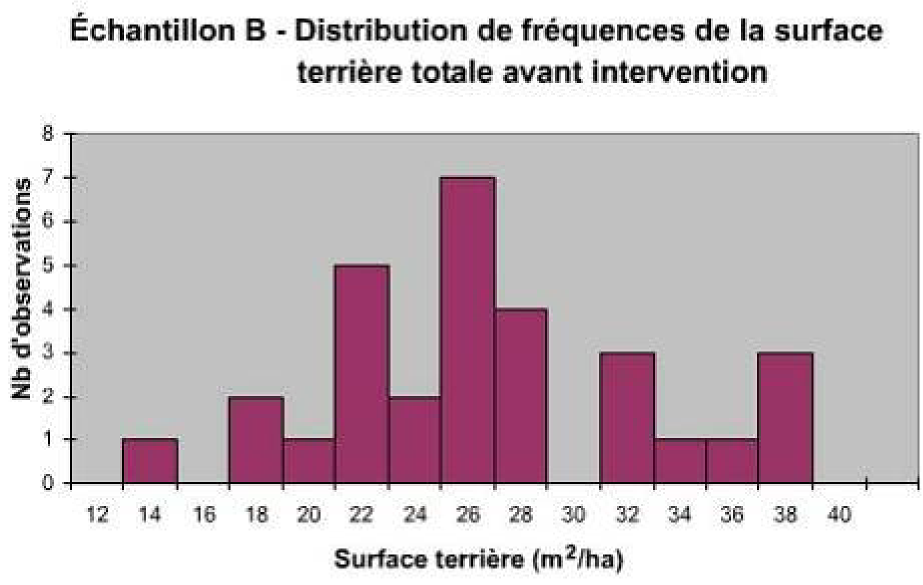
2.5.1 Retour sur l'écart-type de l'échantillon

Si l'on fait un retour sur la notion d'écart-type combinée à la notion de probabilité, pour l'échantillon A (moyenne de 32,93 m^2/ha et écart-type de 5,96 m^2/ha), on peut vérifier, en observant le graphique de distribution de fréquence, qu'environ 95 % des placettes (28 sur 30) se trouvent dans un intervalle de 2S (2 écarts-types), soit entre 21,0 et 44,9 m^2/ha.

En présence d'un échantillon constitué d'un nombre d'observation suffisamment élevé pour être considéré bien représentatif de la population, cela signifie que si l'on retourne sur le terrain prendre la mesure d'un point d'observation additionnel dans ce peuplement, il y a **approximativement** 95 % de chance que la valeur de cette observation se retrouve à l'intérieur de l'intervalle compris entre deux écarts-types autour de la moyenne.



Dans le cas de l'échantillon B, l'intervalle de $2S$ (2 écarts-types) s'étend de 14,4 et 39,1 m^2/ha, ce qui couvre l'étendue des 30 placettes. Ceci permet de démontrer **que la probabilité à 95 % est approximative** et fonction de la distribution des placettes dans la population et de la taille de l'échantillon.



2.5.2 Coefficient de variation

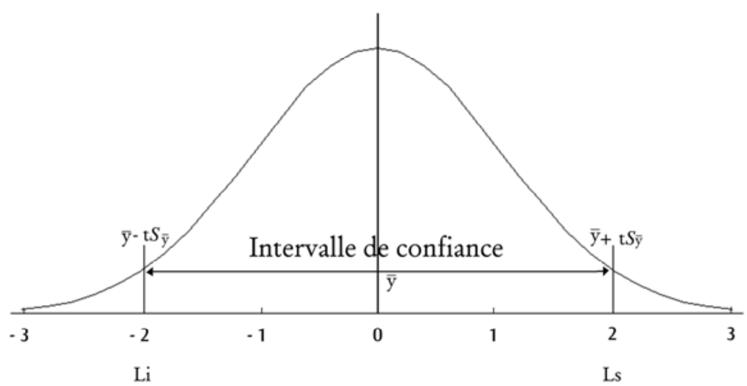
C'est le rapport entre l'écart-type et la moyenne. Cette valeur est exprimée en pourcentage. Il est estimé par :

$$CV = \frac{S}{\overline{Y}} \times 100\% \qquad \text{(formule 13)}$$

2.5.3 Intervalle de confiance

L'intervalle de confiance correspond à l'intervalle à l'intérieur duquel devrait se trouver la vraie valeur de la moyenne, et ce avec un certain niveau de confiance (habituellement 95 %). Comme les résultats sont obtenus au moyen d'un échantillon, c'est la table t de Student qui est utilisée (figure 14).

Figure 14 Distribution de fréquence selon la loi de t de student



Intervalle de confiance (L_i, L_s)

$$L_i = \overline{Y} - t_{\nu,1-\alpha/2} \times S_{\overline{Y}} \qquad \text{(formule 14)}$$

$$L_s = \overline{Y} + t_{\nu,1-\alpha/2} \times S_{\overline{Y}} \qquad \text{(formule 15)}$$

En pratique, pour un nombre d'observations supérieur à 30, la variable *t* de Student intervenant dans le calcul est souvent simplifiée et remplacée par la variable normale réduite μ, ce qui conduit pour $\alpha = 0{,}05$, à l'expression :

$$\overline{Y} \pm \mu_{1-\alpha/2} \times S_{\overline{Y}} \cong \overline{Y} \pm 2S_{\overline{Y}}$$

Concernant l'intervalle de confiance, si l'on échantillonnait la population un très grand nombre de fois et que l'on construisait, pour chaque échantillon, l'intervalle de confiance (deux erreurs standard autour de la moyenne), à peu près 95 % de ces intervalles contiendraient la moyenne de la population (Good et Hardin, 2003).

À titre d'exemple (adapté de Steel and Torrie, 1980), suite à une expérience, le poids de 4 individus choisis au hasard a été enregistré : 28, 65, 35 et 36 kg. Nous voulons estimer la moyenne de la population. Nous supposons que cette population a une distribution normale, de moyenne et de variance inconnues. Les statistiques de l'échantillon sont les suivantes : $\overline{Y} = 41$ kg, $S^2 = 269$ et $S_{\overline{Y}} = 8{,}2$ kg. Le $t_{.025}$ pour 3 degrés de liberté est de 3,18. L'estimation de l'intervalle pour la moyenne de la population est donnée par $\overline{Y} \pm t_{\nu,1-\alpha/2} S_{\overline{Y}} = 41 \pm (3{,}18) \times (8{,}2) = (15{,}67)$ kg. Ainsi, on assigne 95 % comme mesure de notre confiance dans l'énoncé que la moyenne de la population μ se situe dans l'intervalle de 15 à 67 kg. Toutefois, il est possible que la moyenne de la population se situe en dehors de cet intervalle si l'échantillon prélevé est exceptionnel, ce qui peut survenir 1 fois sur 20 (soit 5 % du temps). Dans cet exemple, il y a donc un risque de faire une affirmation fausse (niveau de signification) dans 5 % des cas.

Enfin, si les points d'observation de la population ne sont pas distribués selon une courbe normale, l'intervalle de confiance pour la moyenne de l'échantillon sera approximatif. Cette approximation sera meilleure si $n \geq 30$ et lorsque la distribution des points d'observation s'approche de la normale (OIFQ, 1996) comme démontré dans la section 2.4.7.

2.5.4 Taille de l'échantillon (effectif)

La taille de l'échantillon (ou nombre de placettes requis) dépend normalement de la précision recherchée (Rondeux, 1999). Dans ce contexte, la moyenne de la population peut être estimée avec une erreur d'échantillonnage *d* pour un degré de confiance (niveau de probabilité) de $1-\alpha$. L'évaluation du nombre de placettes requis nécessite donc la connaissance de la variance du paramètre estimé. Cette variance provient soit d'un inventaire antérieur ou d'un prééchantillonnage. Le calcul de la taille de l'échantillon dépend aussi du type d'échantillonnage employé. En premier lieu, nous verrons ici les formules qui s'appliquent à l'échantillonnage aléatoire simple.

$$n = \frac{t^2_{\nu,1-\alpha/2} \times S^2}{d^2} \qquad \text{(formule 16)}$$

Si l'on désire tenir compte du taux d'échantillonnage, alors on peut calculer :

$$\hat{n} = \frac{n}{1 + \dfrac{n}{N}} \qquad \text{(formule 17)}$$

Où : n = Nombre d'observations requis

N = Nombre d'observations possibles dans la population

$\hat{n}$ = Nombre d'observations requis en tenant compte du taux d'échantillonnage

S^2 = Variance de l'échantillon

$t_{\nu,1-\alpha/2}$ = Quantile de la loi de t de « Student » avec une probabilité $(1-\alpha/2)$ 100ième percentile et ν degrés de liberté; si $\nu \rangle 120$ utiliser t = 1,96; ν représente le nombre de degrés de liberté du prééchantillonnage

d = Erreur d'échantillonnage, correspond à la demi-largeur de l'intervalle de confiance. Il s'agit de l'erreur relative permise par rapport à la moyenne (pour une précision de 90 %, l'erreur relative est égale à 10 % de la moyenne ($d = e \times \overline{Y} = 10\,\% \times \overline{Y}$)). On multiplie *e* par la moyenne pour que l'erreur soit exprimée dans les mêmes unités que la variance de l'échantillon

Étant donné que l'erreur standard de la distribution de la moyenne est égale à $S_{\overline{Y}} = S\sqrt{(1-f)/n}$, et que l'écart-type ne peut être contrôlé, le seul élément susceptible de permettre l'obtention d'une plus grande précision est la taille de l'échantillon *n*. Ainsi, une augmentation de la taille de l'échantillon amènera une diminution de l'erreur standard et du même coup une augmentation de la précision. Par contre, une augmentation de la taille de l'échantillon provoquera une augmentation des coûts. Il faudra donc chercher à optimiser la taille de l'échantillon de façon à obtenir une précision acceptable, tout en considérant les coûts associés à la réalisation de l'échantillonnage.

Il n'existe pas de règle absolue quant à la détermination du pourcentage d'erreur relative acceptable pour déterminer la taille de l'échantillon. Toutefois, de façon pratique, il peut être jugé parfois raisonnable de tolérer une erreur relative allant de 5 à 30 %.

2.6 Notions supplémentaires

Les notions d'erreur relative et de précision, présentées plus bas, sont des estimations relatives de la fiabilité des résultats, obtenues par échantillonnage, sans fondement statistique. Ils sont une forme simplifiée d'interprétation de la grandeur de la demi-largeur de l'intervalle de confiance par rapport à la moyenne. En cas de doute, c'est l'intervalle de confiance qui doit être utilisé comme étant la référence statistique reposant sur les concepts de probabilité visant à décrire la précision des résultats.

Ces notions sont présentées afin de fournir une référence sur la façon exacte dont elles sont définies et calculées dans le domaine forestier.

2.6.1 Erreur relative (imprécision)

Il s'agit de la demi-largeur de l'intervalle de confiance à 95 % de la moyenne, exprimée en pourcentage de $\bar{Y}$. Plus l'erreur standard est grande, plus grande est l'erreur relative.

$$e = \left[\frac{t_{v,1-\alpha/2} \times S_{\bar{Y}}}{\bar{Y}} \times 100\% \right] \qquad \text{(formule 18)}$$

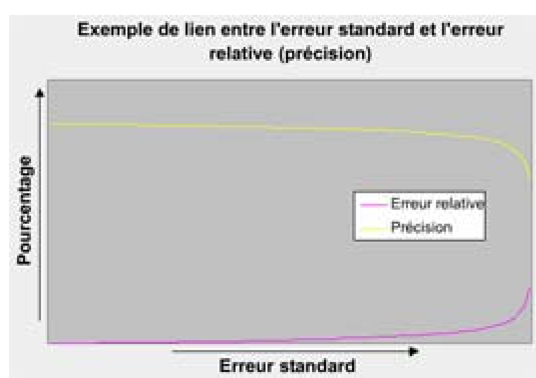
Le concept de précision (ou d'imprécision) d'une estimation couvre l'influence de l'écart-type estimé de la moyenne, qui dépend de la variance et du nombre d'observations. Il n'y a aucune relation avec le biais et les erreurs systématiques. L'imprécision due à l'échantillonnage est d'autant plus faible que le nombre d'observations prises en considération est élevé (Rondeux, 1999).

2.6.2 Précision

Il s'agit du complément de l'erreur relative, soit le complément de la demi-largeur de l'intervalle de confiance à 95 % de la moyenne, exprimée en pourcentage de $\bar{Y}$. Plus l'erreur standard est faible, plus grande est la précision.

$$P = 100\% - \left[\frac{t_{v,1-\alpha/2} \times S_{\bar{Y}}}{\bar{Y}} \times 100\% \right] \qquad \text{(formule 19)}$$

Dans certaines situations, il faut se méfier des calculs de précision car il est possible d'obtenir des résultats aberrants



2.6.3 Exemple de calcul des statistiques pour l'échantillon A - Surface terrière totale et pourcentage de prélèvement total

Surface terrière totale

Pour la variable de la surface terrière totale, la moyenne se calcule au moyen de la formule 2, l'écart-type avec la formule 4(5), l'erreur standard avec la formule 7(8), l'intervalle de confiance à l'aide des formules 14 et 15, l'effectif requis avec la formule 16 et la précision avec la formule 19.

Statistiques

Pour l'échantillon A, la moyenne est de 32,93 m^2/ha, l'écart-type est de 5,96 m^2/ha, l'erreur standard est de 1,09 m^2/ha et l'intervalle de confiance à un niveau de probabilité de 95 % va de 30,70 et 35,15 m^2/ha, la précision calculée est de 93,2 %. Pour cette variable, il aurait suffi d'approximativement 14 observations pour atteindre une précision de 90 %.

Pourcentage de prélèvement

Le pourcentage de prélèvement est le ratio de la surface terrière martelée par rapport à la surface terrière totale. Étant donné qu'il s'agit d'un ratio, le rapport moyen est estimé à l'aide de la formule 3 et l'écart-type avec la formule 6. Les autres statistiques se calculent comme illustré précédemment; ce sont seulement les moyennes et les écarts-types qui doivent être calculés de façon différente pour les rapports.

Statistiques

Pour l'échantillon A, le pourcentage de prélèvement moyen est de 29,4 % avec une précision de 93,3 %. L'intervalle de confiance à 95 % de probabilité est de 27,4 et 31,3 % de prélèvement.

2.6.4 Exercice de calcul des statistiques pour l'échantillon B - Surface terrière totale et pourcentage de prélèvement total

Surface terrière totale

Pour la variable de la surface terrière totale de l'échantillon B, la moyenne est de 26,73 m^2/ha, l'écart-type est de 6,16 m^2/ha, l'erreur standard est de 1,12 m^2/ha et l'intervalle de confiance va de 24,43 et 29,03 m^2/ha, la précision obtenue de 91,4 % et l'effectif requis pour atteindre une précision de 90 % d'approximativement de 22 observations.

Pourcentage de prélèvement

Pour l'échantillon B, le rapport moyen obtenu est de 24,7 %, avec une précision de 84 %. L'intervalle de confiance va de 20,7 et 28,6 % de prélèvement.

2.6.5 Exemple supplémentaire de calculs statistiques – résultats aberrants de précision

Après la récolte d'un peuplement mature sur pied dans un secteur résineux, il est nécessaire de vérifier la présence de régénération résineuse établie afin de confirmer les besoins de remise en production du site par la préparation de terrain et le reboisement.

L'échantillon F est l'évaluation de la distribution des tiges résineuses présentes après la récolte d'un peuplement résineux. Il s'agit d'un échantillon constitué de 600 placettes de 4 m^2 regroupées en 60 grappes de 10 placettes. Les données des cet échantillon sont présentées à la page suivante.

Ainsi, lorsque l'on estime les statistiques de la distribution de la présence de résineux, on obtient une moyenne de 13 % de distribution, une erreur d'échantillonnage de 3,1 % de distribution et une précision de 76,2 %.

Suite à l'analyse de ces statistiques, il a été possible de faire la démonstration qu'il n'est pas correct d'utiliser la statistique de la précision pour estimer la fiabilité des résultats de distribution de faible valeur (i.e. distribution inférieure à 50 %). En effet, le calcul de l'erreur relative (imprécision) correspond au ratio de l'erreur d'échantillonnage (demi-largeur de l'intervalle de confiance) sur la moyenne; ainsi, plus la moyenne est petite, plus l'erreur relative est grande.

Afin de faire cette démonstration, si l'on calcule la distribution de l'absence de résineux (complément de la présence), on obtient une moyenne de 87 % de distribution et la même erreur d'échantillonnage de 3,1 % de distribution. Toutefois, puisque la moyenne est plus grande, on obtient maintenant une précision de 96,4 %.

Ce résultat amène à se questionner sérieusement sur la justesse de l'utilisation de la statistique de la précision pour l'évaluation de la fiabilité des données de distribution telles que la présence de la régénération.

# placette	Présence résineuse Nbr/10	Absence résineuse Nbr/10	Présence résineuse	Absence résineuse	Numéro de	Présence résineuse Nbr/10	Absence résineuse Nbr/10	Présence résineuse	Absence résineuse
00071	0	10	0.0	1.0	00136	0	10	0.0	1.0
00072	1	9	0.1	0.9	00137	1	9	0.1	0.9
00075	1	9	0.1	0.9	00138	1	9	0.1	0.9
00076	1	9	0.1	0.9	00139	1	9	0.1	0.9
00077	2	8	0.2	0.8	00140	0	10	0.0	1.0
00078	1	9	0.1	0.9	00141	4	6	0.4	0.6
00079	3	7	0.3	0.7	00142	2	8	0.2	0.8
00080	1	9	0.1	0.9	00143	1	9	0.1	0.9
00081	2	8	0.2	0.8	00144	1	9	0.1	0.9
00082	0	10	0.0	1.0	00147	2	8	0.2	0.8
00116	1	9	0.1	0.9	00148	1	9	0.1	0.9
00117	3	7	0.3	0.7	00151	5	5	0.5	0.5
00118	1	9	0.1	0.9	00152	3	7	0.3	0.7
00119	3	7	0.3	0.7	00163	1	9	0.1	0.9
00120	0	10	0.0	1.0	00164	0	10	0.0	1.0
00121	3	7	0.3	0.7	00165	2	8	0.2	0.8
00122	2	8	0.2	0.8	00166	0	10	0.0	1.0
00123	3	7	0.3	0.7	00169	2	8	0.2	0.8
00124	1	9	0.1	0.9	00170	0	10	0.0	1.0
00125	0	10	0.0	1.0	00172	0	10	0.0	1.0
00126	2	8	0.2	0.8	00173	1	9	0.1	0.9
00127	1	9	0.1	0.9	00174	1	9	0.1	0.9
00128	0	10	0.0	1.0	00175	0	10	0.0	1.0
00129	1	9	0.1	0.9	00176	1	9	0.1	0.9
00130	0	10	0.0	1.0	00178	0	10	0.0	1.0
00131	2	8	0.2	0.8	00179	0	10	0.0	1.0
00132	0	10	0.0	1.0	00180	1	9	0.1	0.9
00133	0	10	0.0	1.0	00233	2	8	0.2	0.8
00134	1	9	0.1	0.9	00237	3	7	0.3	0.7
00135	4	6	0.4	0.6	00238	2	8	0.2	0.8

	Présence résineuse Nbr/10	Absence résineuse Nbr/10	Présence résineuse	Absence résineuse
Nb placette	60	60	60	60
Moyenne	1.300	8.700	0.130	0.870
Écart-type	1.197	1.197	0.120	0.120
Erreur standard	0.155	0.155	0.015	0.015
Erreur d'échantillonnage	0.309	0.309	0.031	0.031
Limite inférieure de l'intervalle de confiance	0.991	8.391	0.099	0.839
Limite supérieure de l'intervalle de confiance	1.609	9.009	0.161	0.901
Précision	0.762	0.964	0.762	0.964
	Nombre de placette sur 10 occupé par au moins une tige résineuse		Pourcentage des placettes occupées par au moins une tige résineuse	

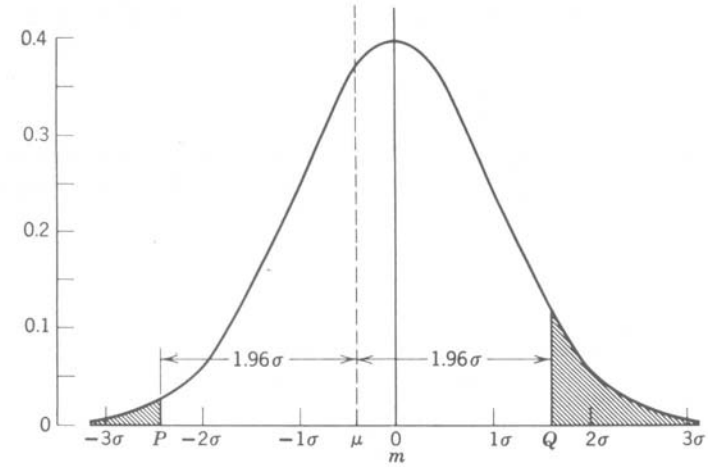
2.7 Limite d'application des résultats d'échantillonnage

Les résultats d'un échantillonnage ne s'appliquent qu'à la population qui a été échantillonnée. **Ainsi, les résultats ne peuvent s'appliquer à une population donnée, si seulement une partie de cette population a été échantillonnée, ou encore, lorsque ces résultats proviennent d'une population différente** (OIFQ, 1996).

Pour que les résultats soient représentatifs de la vraie valeur d'une variable donnée d'une population, l'estimation de cette variable doit être non biaisée. **Le fait de choisir les points d'observations au hasard est essentiel pour éviter l'introduction de biais**. Il existe aussi d'autres sources potentielles de biais pouvant être associées à des instruments de mesure en mauvais état, ou au manque de précaution lors de la prise de données. L'existence d'un biais fournira une mesure de la tendance centrale (moyenne, médiane ou centre de l'intervalle) qui sera décalée d'un côté ou de l'autre par rapport à la vraie valeur de la variable observée pour la population (voir figure 15).

Le biais correspond donc à une erreur qui est systématiquement dans le même sens. Par exemple, une mesure du diamètre à hauteur de poitrine (DHP) qui serait toujours effectuée à un niveau inférieur au 1,30 m prescrit produirait un biais positif conduisant à une surestimation du diamètre moyen. Dans la figure 15, l'ampleur du biais correspondrait à $m - \mu$. En présence d'un biais, on constate que les extrémités de la courbe sont affectées différemment et que les probabilités associées à la moyenne de l'échantillon sont distordues.

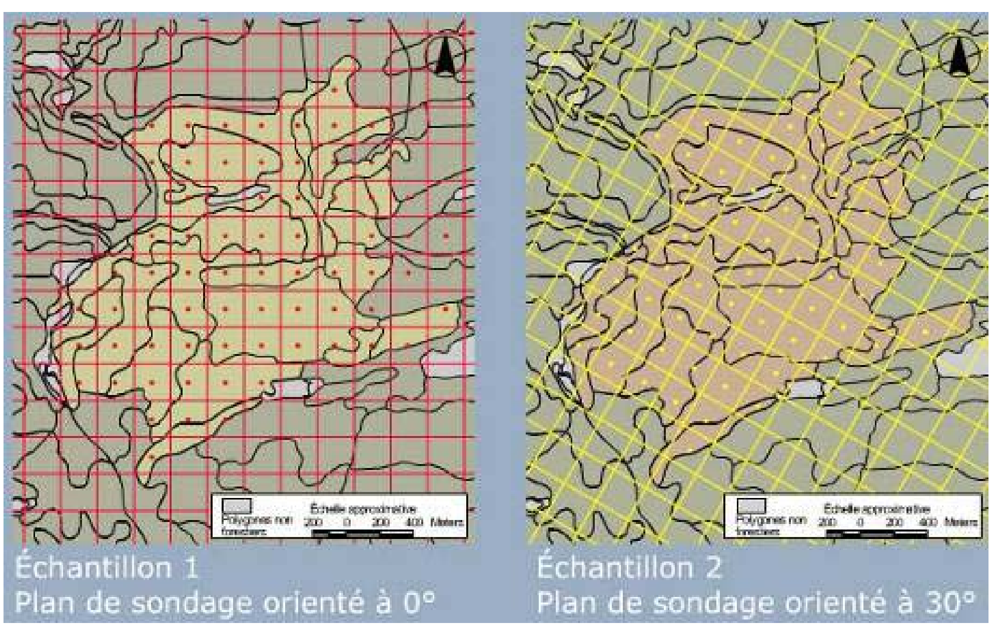
Figure 15 Effet d'un biais sur les erreurs d'estimation (tiré de Cochran, 1977)



2.8 Comparaison de deux échantillons

La comparaison de moyennes provenant de deux échantillons se fait au moyen d'un test d'hypothèses. Le test de *t* de Student constitue le principal critère utilisé dans le cadre de tests d'hypothèses concernant une ou deux moyennes (Steel and Torrie, 1980).

Souvent, nous voulons vérifier si les moyennes provenant de deux échantillons indépendants, tirés de la même population, sont égales. En théorie dans ce cas, les variances dans les échantillons, s_1^2 et s_2^2, devraient être égales puisqu'ils estiment la variance σ^2 de la même population. Il est tout de même possible de vérifier l'égalité des variances au moyen d'un test de F.



2.8.1 Test d'égalité des variances

Afin d'utiliser le test de t approprié, il y a lieu de vérifier l'égalité des variances au moyen d'un test de F. Il est possible de tester l'hypothèse suivante :

$H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$

σ_1^2 : variance de l'échantillon 1 de la population
σ_2^2 : variance de l'échantillon 2 de la population

En utilisant les variances des échantillons, nous pouvons calculer la valeur de F :

$$F = \frac{S^2 \ la \ plus \ grande}{S^2 \ la \ plus \ petite} \qquad \text{(formule 20)}$$

La valeur de F calculée est comparée à celle de la table de F (tableau 4) en utilisant comme référence de la ligne à travers le haut de la table, le nombre de degré de liberté de la variance de l'échantillon la plus grande et, en utilisant comme référence de la colonne à gauche de la table, le nombre de degré de liberté de la variance de l'échantillon la plus petite, et ce avec un niveau de signification unilatéral $\alpha / 2 = 0{,}025$. Rappelons que le niveau de signification unilatéral $\alpha / 2 = 0{,}025$ est équivalant à un niveau de signification bilatéral $\alpha = 0{,}05$.

Dans le cas où l'on observe une différence significative de la variance des échantillons, il y a lieu de s'interroger sur les raisons ayant causé cette situation et l'introduction possible d'un biais puisque nous sommes en présence de deux échantillons provenant du même secteur forestier servant à estimer la même variance σ^2 de la population. En théorie, les variances d'échantillon d'une même population sont égales avec un niveau de probabilité de 95 %.

Tableau 4 Valeur de F pour un niveau de signification unilatéral $\alpha / 2 = 0{,}025$

	Numérateur df																		
Dénominateur df	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	647,7	799,	864,	899,	921,	937,	948,	956,	963,	968,	976,	984,	993,	997,	1001,	1005,	1009,	1014,	1018,0
2	38,51	39,0	39,1	39,2	39,3	39,3	39,3	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,46	39,47	39,48	39,49	39,50
3	17,44	16,0	15,4	15,1	14,8	14,7	14,6	14,5	14,4	14,4	14,3	14,2	14,1	14,1	14,08	14,04	13,99	13,95	13,90
4	12,22	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

2.8.2 Test t de Student (variances égales)

En supposant que les variances d'échantillons de la variable aléatoire dans la population sont égales, il existe une statistique qui suit une loi de t exacte. Il est possible d'utiliser le test de t pour vérifier cette hypothèse. Il s'agira de tester l'hypothèse suivante :

$H_0 : u_1 = u_2$ versus $H_1 : u_1 \neq u_2$

u_1 : moyenne de l'échantillon 1 de la population
u_2 : moyenne de l'échantillon 2 de la population

En utilisant les moyennes et variances des échantillons, nous pouvons calculer la statistique t_0 :

$$t_0 = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{(n_1 - 1)S_1^2(1 - f_1) + (n_2 - 1)S_2^2(1 - f_2)}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ (formule 21)

Dans le cas où le taux d'échantillonnage est faible (< 1 %), il est possible d'utiliser la formule simplifiée :

$$t_0 = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ (formule 22)

avec $v_0 = n_1 + n_2 - 2$ degrés de liberté (formule 23)

où t_0 est distribuée selon une distribution de t de Student.

On rejette donc $H_0 \text{ si} \left|t_0\right| \rangle\, t_{\nu, 1-\alpha/2}$ de la table

Où : $\overline{Y}_1$ = Moyenne de l'échantillon 1

$\overline{Y}_2$ = Moyenne de l'échantillon 2

S_1^2 = Variance de l'échantillon 1

S_2^2 = Variance de l'échantillon 2

n_1 = Nombre d'observations de l'échantillon 1

n_2 = Nombre d'observations de l'échantillon 2

f_1 = Taux d'échantillonnage de l'échantillon 1

f_2 = Taux d'échantillonnage de l'échantillon 2

2.8.3 Test t de Student (variances inégales)

Si on ne peut pas assumer que les deux variances sont égales, il est tout de même possible de vérifier l'égalité des moyennes des deux échantillons selon la loi de *t* approximative.

Ainsi, en supposant que les deux variances d'échantillon sont différentes, il s'agira de tester l'hypothèse suivante :

$H_0 : u_1 = u_2$ versus $H_1 : u_1 \neq u_2$

u_1 : moyenne de l'échantillon 1 de la population
u_2 : moyenne de l'échantillon 2 de la population

En utilisant les moyennes et variances des échantillons, nous pouvons calculer la statistique $t_0^{'}$:

$$t_0^{'} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2(1-f_1)}{n_1} + \frac{S_2^2(1-f_2)}{n_2}}} \qquad \text{(formule 24)}$$

avec $$v_0^{'} = \frac{\left(\frac{S_1^2(1-f_1)}{n_1} + \frac{S_2^2(1-f_2)}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2(1-f_1)}{n_1}\right)^2}{(n_1-1)}\right] + \left[\frac{\left(\frac{S_2^2(1-f_2)}{n_2}\right)^2}{(n_2-1)}\right]}$$ degrés de liberté (formule 25)

Dans le cas où le taux d'échantillonnage est faible (< 1 %), il est possible d'utiliser les formules simplifiées :

$$t_0^{'} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad \text{(formule 26)}$$

avec $\nu_0^{'} = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\left[\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{(n_1 - 1)}\right] + \left[\dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{(n_2 - 1)}\right]}$ degrés de liberté (formule 27)

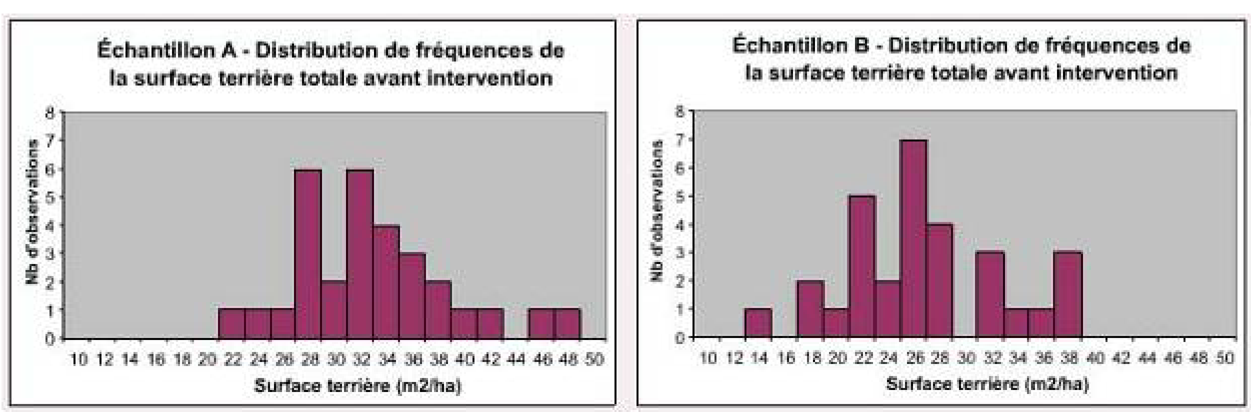
où $t_0^{'}$ est distribuée selon une distribution de t de Student.

On rejette donc $H_0 \text{ si} \left|t'_0\right| \rangle\, t_{\nu,1-\alpha/2}$ de la table

Lors de la comparaison de deux échantillons, l'utilisation du test de t est justifié parce qu'il permet de calculer une « erreur standard pondérée » des deux échantillons comparés. Ainsi, les résultats de l'un ne servent pas de référence pour l'autre mais sont utilisés tous les deux de façon complémentaire.

2.8.4 Exemple de calcul du test de t pour les échantillons A et B - Surface terrière totale

Une première comparaison visuelle (graphiques de distribution de fréquences) permet tout d'abord de constater que non seulement les moyennes sont distantes, mais de plus, leurs intervalles de confiance respectifs ne se chevauchent pas, et ce en dépit d'une précision semblable et très bonne (> 90 %) dans les deux cas.



Après vérification que les variances des deux échantillons sont égales (formule 20), le test de t (formule 22) a permis de faire ressortir la différence significative entre les moyennes mesurées des deux échantillons. L'obtention d'une différence significative entre deux échantillons provenant du même secteur forestier suscite des interrogations. Cette différence peut être soit le fruit du hasard (1 chance sur 20), soit résulter de l'existence d'un biais dans la prise de données. Pour vérifier qu'un biais n'a pas été introduit, il faudrait s'assurer que le plan de sondage initial a été bien conçu et respecté dans les deux cas, qu'aucun déplacement de placettes ne soit toléré et que les données ont été correctement récoltées, avec un équipement en bon état.

2.8.5 Exercice de calcul du test de t pour les échantillons A et B – Pourcentage de prélèvement total

Avant d'effectuer la comparaison des moyennes, la vérification de l'égalité des variances a permis de détecter une différence significative au seuil de probabilité de 95 %. Il y a lieu de s'interroger sur les raisons ayant causé cette différence significative. Cette différence peut soit être le fruit du hasard (1 chance sur 20), soit résulter de l'existence d'un biais dans la réalisation des travaux suggérant un comportement différent dans chacun des cas.

L'utilisation du test de t approprié (formule 26) permet d'estimer que le pourcentage de prélèvement total est significativement différent entre les deux échantillons. Comme dans le cas de l'inégalité des variances, cette différence significative est soit le fruit du hasard (1 chance sur 20), soit le résultat de l'existence d'un biais dans la réalisation des travaux.

2.9 La stratification

2.9.1 Échantillonnage stratifié aléatoire

Dans le cas de l'échantillonnage aléatoire stratifié, la population est subdivisée en groupes relativement homogènes appelés strates. La sélection des points d'observation se fait de manière aléatoire dans chaque strate (Figure 16).

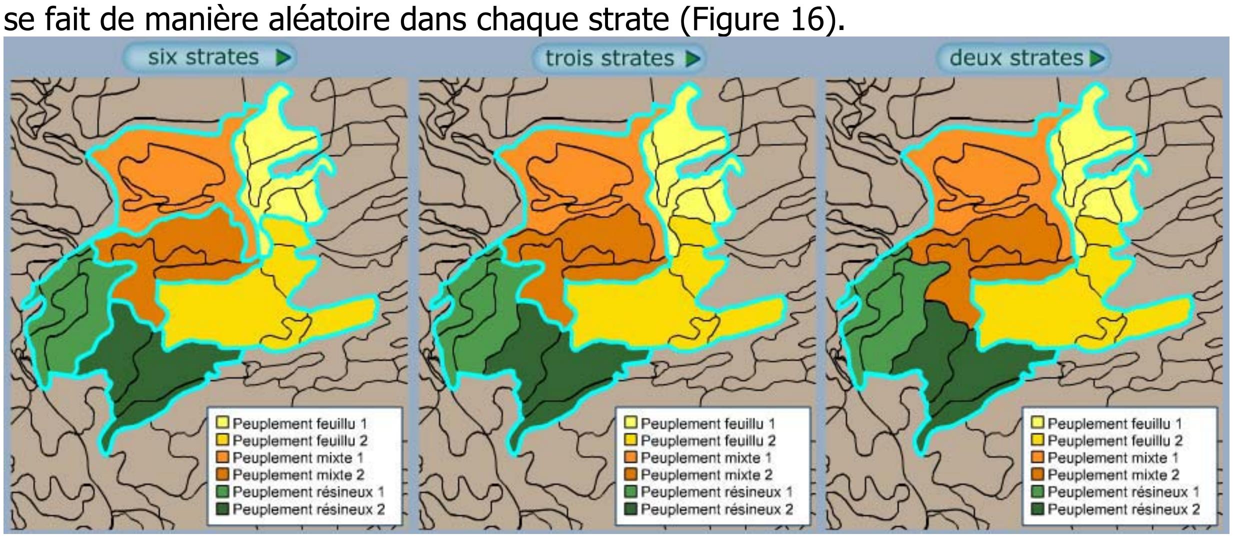
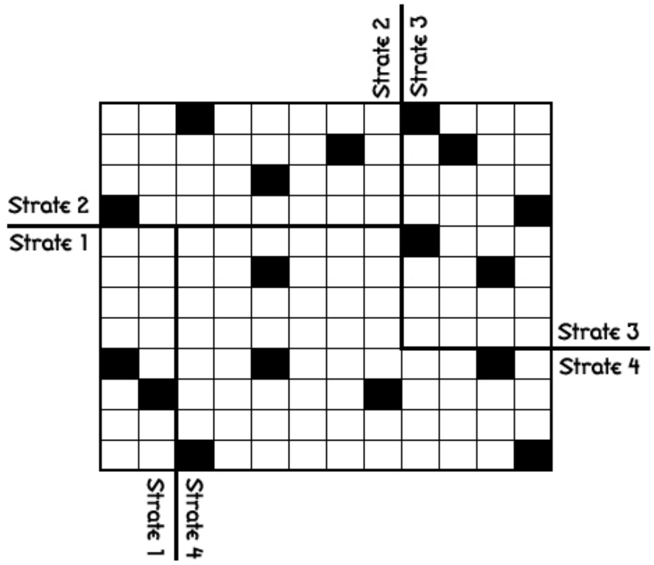


Figure 16 Illustration de l'échantillonnage aléatoire stratifié



Chaque carreau noir représente un point d'observation (Source : OIFQ, 1996)

a) Principes et pratique

Cette méthode d'échantillonnage permet de diminuer l'effet de la variabilité de la variable à estimer et conséquemment, de réduire l'erreur d'échantillonnage (Rondeux, 1999). La création de strates vise à minimiser la variabilité intrastrate et maximiser la variabilité interstrate. Ainsi, le nombre total de points d'observation requis pour atteindre un niveau de précision donné avec cette méthode devrait être moins élevé que le nombre requis avec l'échantillonnage aléatoire simple.

Les principes de l'échantillonnage stratifié peuvent être utilisés pour toute variable pouvant faire l'objet d'une pondération, telle que l'âge du peuplement ou le diamètre moyen obtenus à partir d'études d'arbres de différentes essences.

La répartition des points d'observation par strate peut se faire de plusieurs façons, entre autres, d'une façon proportionnelle[3] ou optimale[4]. Dans le premier cas, la proportion de l'échantillon attribué dans chaque strate sera établie sur la base de l'importance relative de chaque strate. Dans le cas de la répartition optimale, la portion de l'échantillon attribué à chaque strate est basée sur le poids de la strate et son écart-type, ce qui

[3] La stratification avec allocation proportion-nelle signifie que le taux d'échantillonnage (f) est le même dans chaque strate.
[4] La stratification avec allocation optimale signifie que l'on cherche à intensifier l'échantillonnage dans les strates ayant les plus fortes variabilités (les plus grandes variances de l'échantillon).

permet d'accroître l'échantillonnage dans les strates ayant une variance élevée. Cette méthode d'allocation des points d'observation requiert cependant la connaissance préalable de la variance dans chaque strate, qui peut être obtenue soit grâce à des échantillonnages antérieurs, ou au moyen d'un prééchantillonnage.

b) Avantages et inconvénients

Les principaux avantages de l'échantillonnage aléatoire stratifié sont les suivants (tiré de OIFQ, 1996) :

- il donne des estimations généralement plus précises, car il réduit la variabilité d'une variable dans la population. Si les strates sont homogènes, la variable mesurée varie peu d'un point d'observation à l'autre à l'intérieur d'une même strate et l'estimation est alors plus précise.
- il permet d'obtenir des estimations distinctes pour certaines strates ou sous-populations;
- il assure une représentation adéquate des sous-populations dans l'échantillon;
- il permet d'appliquer des techniques d'échantillonnage différentes à diverses catégories ou strates de la population. Par exemple, si l'objectif de l'inventaire est d'évaluer le volume marchand, on pourra avoir recours, pour les strates classées en voie de se régénérer, à une méthode de sondage où l'on pourra déterminer la densité de la régénération.

Les principaux inconvénients de cette méthode sont (tiré de OIFQ, 1996) :

- la dimension de chaque strate doit être connue avant l'échantillonnage ou estimée aussi exactement que possible;
- chaque point d'observation doit figurer dans une seule strate. Cela suppose que les strates ne peuvent se chevaucher ou se recouper. Leurs limites doivent donc être bien connues.

2.9.2 Gains en précision

Selon Cochran (1977), s'il était possible de stratifier sur la base de la variable mesurée, il n'y aurait pas de chevauchement entre les strates et la variance à l'intérieur de chaque strate serait beaucoup plus petite que la variance avant stratification. Évidemment, il est impossible de stratifier selon les valeurs de la variable à mesurer. Mais, lorsque les conditions font que la stratification effectuée s'approche de cette situation, les gains en précision deviennent particulièrement importants. Pour ce faire, trois conditions doivent être respectées :

- La population est composée d'éléments qui varient considérablement entre eux;
- Les principales variables à mesurer sont fortement corrélées aux critères de stratification;
- Une bonne mesure de la dimension de la strate est disponible.

Un échantillon stratifié avec allocation optimale sera généralement plus performant en termes de réduction de variance, qu'un échantillon stratifié avec allocation proportionnelle. Et, généralement, un échantillon correctement stratifié sera plus performant en terme de réduction de variance qu'un échantillon aléatoire simple.

2.9.3 Méthodes de stratification

La stratification peut se faire *a priori* ou *a posteriori*, selon qu'elle ait lieu avant ou après la sélection des points d'observation. Il est à noter que les deux méthodes impliquent l'utilisation d'estimateurs différents. Nous traiterons ici de la stratification *a priori*.

2.10 Statistiques relatives à l'échantillonnage aléatoire stratifié

Le calcul des moyennes et variances doit tenir compte du type d'échantillonnage utilisé lors de la prise de données.

2.10.1 Moyenne pondérée

Dans le cas d'une stratification, les estimateurs doivent être pondérés en fonction de l'élément de pondération de chaque strate échantillonnée. Les éléments de pondération peuvent être de natures diverses, telles que la superficie, le nombre d'observations, la surface terrière par essence, etc.

$$\overline{Y}_p = \frac{\sum_{h=1}^{L}(N_h \times \overline{Y}_h)}{N} \qquad \text{(formule 28)}$$

2.10.2 Variance de la moyenne pondérée

$$S^2_{\overline{Y}_p} = \frac{\sum_{h=1}^{L}(N_h{}^2 \times S^2_{\overline{Y}_h})}{N^2} \qquad \text{(formule 29)}$$

où : $\overline{Y}_p$ = Moyenne pondérée

$S^2_{\overline{Y}_p}$ = Variance de la moyenne pondérée

L = Nombre de strates

N_h = Poids de la strate h de l'élément de pondération (superficie, nombre d'observations, surface terrière par essence, etc)

$\overline{Y}_h$ = Moyenne de la strate h

$S^2_{\overline{Y}_h}$ = Variance de la moyenne de la strate h

N = Poids total de l'élément de pondération (superficie, nombre d'observations, surface terrière par essence, etc)

Les autres statistiques (intervalle de confiance et précision) sont calculées de la même façon que précédemment, en utilisant la moyenne et la variance de la moyenne pondérée.

2.10.3 Taille de l'échantillon (effectif) pour l'échantillonnage aléatoire stratifié

Lorsqu'il est prévu de procéder par échantillonnage aléatoire stratifié, la formule à utiliser afin de déterminer la taille de l'échantillon qui servira à calculer une moyenne, doit aussi prendre en considération le type d'allocations (proportionnelle ou optimale) que l'on compte utiliser.

Pour une allocation proportionnelle :

$$n = \frac{t^2_{\nu,1-\alpha/2} \sum_{h=1}^{L} W_h S_h^2}{d^2} \qquad \text{(formule 30)}$$

Pour une allocation optimale :

$$n = \frac{t^2_{\nu,1-\alpha/2} \left(\sum_{h=1}^{L} W_h S_h \right)^2}{d^2} \qquad \text{(formule 31)}$$

Dans les deux cas, l'effectif par strate est obtenu par :

$$n_h = n\left[\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h}\right] \qquad \text{(formule 32)}$$

Où : n = Nombre d'observations requis

n_h = Nombre d'observations requis dans la $h^{\text{ième}}$ strate

S_h^2 = Variance de l'échantillon de la $h^{\text{ième}}$ strate

S_h = Écart-type de l'échantillon de la $h^{\text{ième}}$ strate

L = Nombre de strates

d = Erreur d'échantillonnage

$t_{\nu,1-\alpha/2}$ = Quantile de la loi de t de « Student » avec une probabilité $(1-\alpha/2)$ 100$^{\text{ième}}$ percentile et ν degrés de liberté; si $\nu \rangle 120$ utiliser t = 1,96; ν représente le nombre de degrés de liberté du prééchantillonnage

W_h = Proportion du total dans la $h^{\text{ième}}$ strate ($w_h = N_h / N$) où *N* = nombre total d'observations

Ajoutons que les formules à employer pour le calcul de l'effectif requis peuvent être différentes selon que le paramètre à évaluer soit une moyenne, un total ou une proportion. Les formules ici présentées s'appliquent au calcul d'une moyenne. Le lecteur qui désire en savoir davantage sur le sujet est invité à consulter les références pertinentes à ce sujet (Cochran, 1977; Rondeux, 1999).

2.10.4 Exemple de calcul pour des strates regroupées des échantillons B et G - Surface terrière totale

Dans le voisinage du peuplement ayant fait l'objet de l'échantillon B, un autre peuplement a fait l'objet d'une évaluation similaire. Il s'agit d'un peuplement feuillu mature couvert par un échantillonnage constitué de 19 placettes au prisme de facteur 2 (échantillon G). En vue de réunir les informations obtenues pour ces deux peuplements différents, il est possible de procéder au regroupement de ces strates au moyen de la pondération proportionnelle.

La moyenne et la variance pondérées se calculent à l'aide des formules 28 et 29, l'effectif à partir de la formule 30, alors que les autres statistiques sont calculées telles qu'illustrées précédemment en utilisant la moyenne pondérée et la variance de la moyenne pondérée.

Le modèle d'échantillonnage stratifié permet d'obtenir une moyenne de 30,41 m^2/ha, une erreur standard de 0,81 m^2/ha et nous indique que seulement 14 placettes auraient été nécessaires pour atteindre une précision de 90 %.

En revanche, si l'on ne fait pas de distinction entre les deux strates et que l'on regroupe l'ensemble des 49 placettes comme si elles avaient été réalisées selon un plan d'échantillonnage aléatoire simple, on obtient toujours une moyenne de 30,41 m^2/ha mais une erreur standard de 1,04 m^2/ha et on estime que 23 placettes auraient été nécessaires pour atteindre 90 % de précision selon ce plan d'échantillonnage.

Il est donc possible de constater que la réduction de la variance (erreur standard) est spécifique à l'utilisation de l'échantillonnage stratifié aléatoire.

L'échantillon H présente des mesures d'âge effectuées sur des tiges de trois essences différentes. Il est possible de considérer que ces observations aient été obtenues selon un échantillonnage aléatoire simple ou selon un échantillonnage stratifié aléatoire.

Les statistiques de l'échantillonnage aléatoire simple démontrent une moyenne de 53,5 ans et une erreur standard de 2,61 ans et un effectif requis de 13 arbres pour atteindre une précision de 90 %. En assumant que l'on conserve l'identification de l'essence des arbres comme élément de stratification, on obtient des résultats statistiques semblables, ce qui suggère que dans ce cas, l'identification de l'essence ne contribue pas à la distinction de groupe d'âge entre les tiges. On peut remarquer des besoins en effectif plus réduits en considérant le type d'allocation optimale.

Il est également possible de tenir compte de l'identification de l'essence, non pas en fonction du nombre d'arbres étudiés, mais en fonction de la surface terrière respective occupée par chacune de ces essences dans le peuplement. Dans ce cas, on observe une moyenne de 54,5 ans et un besoin en effectif légèrement supérieur pour atteindre une précision de 90 %.

2.11 Limite des méthodes utilisées

Il est essentiel de garder en tête le niveau de confiance associé aux intervalles de confiance autour des moyennes qui sont calculées. En effet, lorsque l'on dit qu'une moyenne et une variance sont calculées à un niveau de probabilité de 95 %, cela signifie que l'intervalle de confiance contient la vraie moyenne de la population 19 fois sur 20. Il peut donc arriver, approximativement une fois sur vingt, que l'intervalle de confiance calculé ne contient pas la moyenne de la population. Il est donc nécessaire de toujours prendre du recul par rapport aux résultats obtenus et se demander s'ils sont bien représentatifs de la population considérée et, en cas de doute, un examen attentif des données permettra parfois d'identifier la source des variations observées (ex : présence d'un individu exceptionnellement gros). En pareilles circonstances, la stratification, l'ajustement de l'effectif et l'examen des paramètres estimés sont des moyens permettant de contrebalancer l'effet de la grande variabilité observée ou la présence d'une donnée aberrante.

3. Exemples de calcul et d'interprétation des statistiques

3.1 Démonstration des concepts

3.1.1 Effet de la taille de l'échantillon

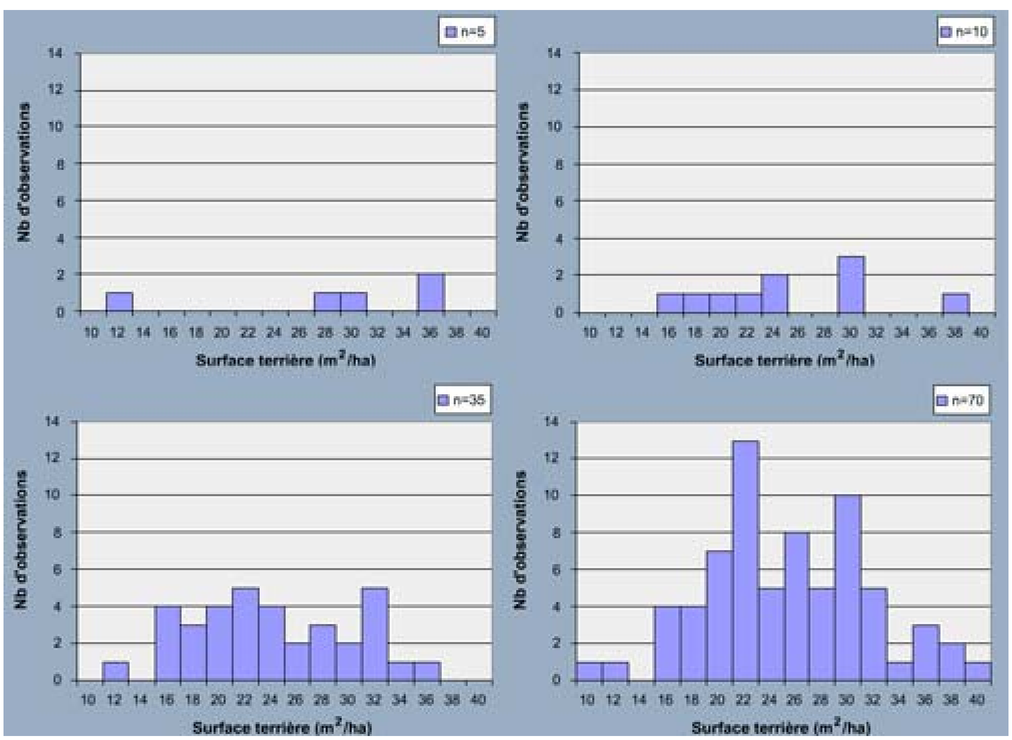
a) Évaluation du paramètre

Supposons un petit peuplement forestier feuillu mature ayant fait l'objet d'un recensement dont la totalité de la superficie a été couverte par la réalisation de 81 placettes. Le paramètre de la surface terrière totale de la population a une moyenne de 25,4 m^2/ha et un écart-type de 6,3 m^2/ha.

Afin d'observer l'effet du taux d'échantillonnage, des échantillons de différentes tailles ont été choisis au hasard parmi l'ensemble de la population des 81 placettes.

Dans un premier temps, des échantillons de 5, 10, 35 et 70 placettes ont été sélectionnés de façon aléatoire, puis les statistiques associées à la surface terrière totale ont été calculées. De plus, pour chacun des échantillons, la distribution de fréquences a été illustrée.

Distribution de fréquence par surface terrière pour 4 échantillons de taille différente:



b) Interprétation des résultats

On constate qu'avec un échantillon de très petite taille (n=5), la moyenne obtenue (28,4 m^2/ha) peut s'éloigner fortement de la moyenne de la population (25,4 m^2/ha). L'examen des statistiques permet de mettre en évidence une variabilité importante, comme le confirme l'écart-type obtenu de 9,8 m^2/ha comparativement à celui de la population égale à 6,3 m^2/ha. Compte tenu de l'importance de l'erreur standard et du faible nombre de placettes, l'intervalle de confiance (16,6 à 40,2 m^2/ha) est grand. À partir de cet échantillonnage, on estime qu'il serait nécessaire de faire 93 placettes pour atteindre une précision de 90 %.

Avec un échantillon constitué de 10 placettes, la moyenne obtenue (25,2 m^2/ha) s'est approchée de celle de la population. L'écart-type (6,7 m^2/ha), qui représente l'étendue des données des placettes individuelles, se rapproche de celle des échantillons de plus grande taille et de la population. On remarque qu'en dépit d'une augmentation de la précision, l'intervalle de confiance est encore grand (20,7 à 29,7 m^2/ha). Sur la base des résultats obtenus avec cet échantillon, il serait nécessaire de réaliser 37 placettes pour atteindre une précision de 90 %.

Avec des échantillons de 35 et 70 placettes, les moyennes obtenues sont, elles aussi, près de celle de la population. Les écarts-types sont semblables à celui obtenu pour la population.

Avec des échantillons contenant 35 et 70 placettes, le nombre de placettes requis pour obtenir une précision de 90 % est du même ordre, soit 27 et 25 placettes respectivement. On observe donc que, pour cette population, le nombre de placettes nécessaires à l'obtention d'une précision de 90 % lors de l'évaluation de la surface terrière totale par échantillonnage se stabilise aux environs de 25.

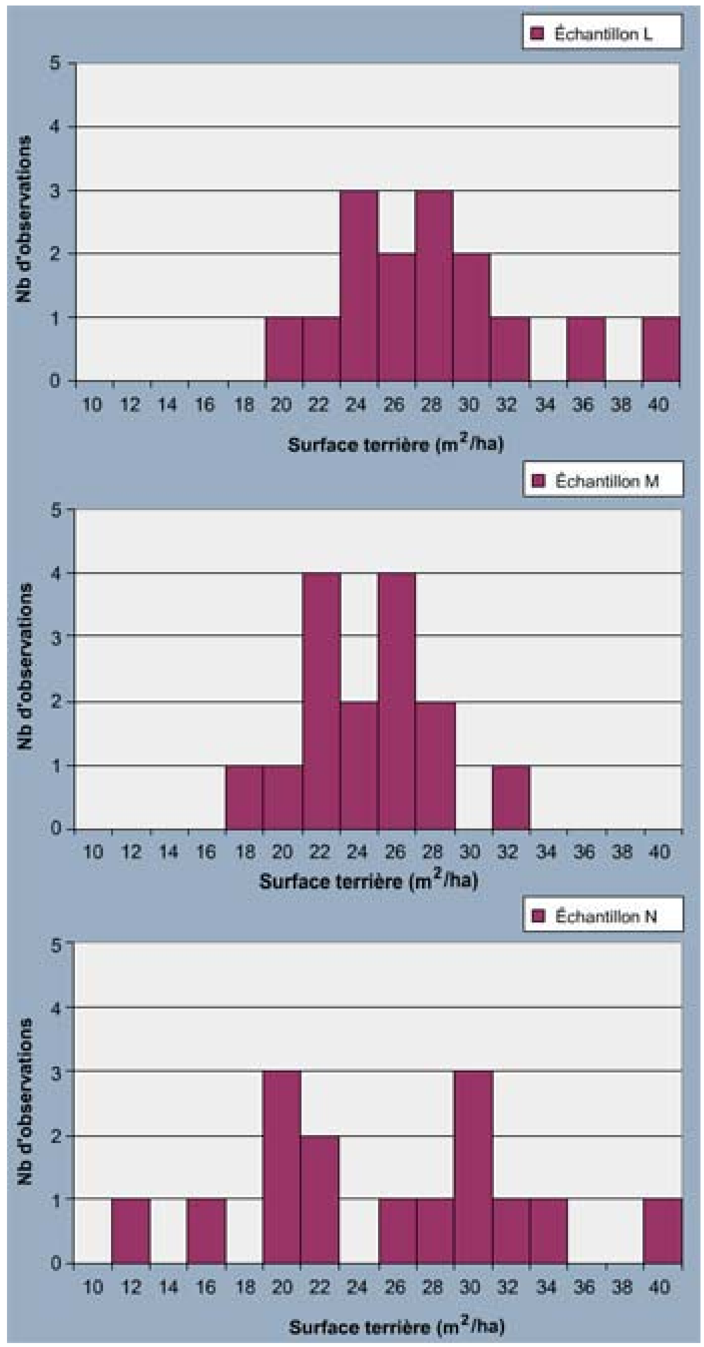
Comme on peut le remarquer suite à l'analyse des statistiques et des graphiques de distribution de fréquences, le fait de passer de 35 à 70 placettes n'apporte pas de gain significatif en termes d'amélioration de la qualité de l'information servant à estimer les paramètres de la population. Par contre, il est facile de visualiser qu'un très faible échantillonnage (exemple de 5 placettes) est plus inquiétant puisque les statistiques estimées sont relativement éloignées des vraies valeurs de la population, justifiant ainsi un nombre plus élevé de placettes à réaliser pour faire l'estimation d'un paramètre de la population.

Lors de l'évaluation de la taille de l'échantillon, l'estimation de l'effectif requis pour une précision donnée est toujours très approximative lorsque calculée à partir d'un faible nombre de placette puisque l'estimation de l'écart-type est très approximative. Il est donc essentiel de refaire ce calcul au fur et à mesure de l'acquisition de nouvelles données.

3.1.2 Effet du hasard

a) Évaluation du paramètre

Afin de démontrer l'effet du hasard, trois tirages aléatoires distincts de 15 placettes ont été effectués parmi la population de 81 placettes. Les statistiques associées à la surface terrière totale ainsi que les graphiques de distribution de fréquence sont présentés.



b) Interprétation des résultats

La moyenne obtenue à partir de l'échantillon L (27,9 m^2/ha) est celle qui s'éloigne le plus de la moyenne de la population. Cette moyenne est plus élevée suite à l'absence de faibles valeurs dans ce tirage. Cette absence a entraîné l'obtention d'un écart-type relativement faible (5,3 m^2/ha), d'où une estimation précise en dépit d'un *n* pas très élevé.

L'échantillon M a conduit à l'obtention d'une faible moyenne (24,4 m^2/ha) compte tenu de l'absence de fortes valeurs. De plus, la variabilité observée au sein de cet échantillon est très faible, comme le confirment le petit écart-type obtenu (3,6 m^2/ha) et l'étroitesse de l'histogramme de fréquence. La précision est par conséquent, très élevée et supérieure à 90 %.

Enfin, l'échantillon N a conduit à l'obtention d'une moyenne (25,5 m^2/ha) se situant entre les deux échantillons précédents, mais l'écart-type est nettement plus élevé (7,5 m^2/ha) comme l'indiquent l'étalement de la courbe de fréquences et la précision plus faible.

Il est ainsi possible de constater que ces trois échantillons de 15 placettes issus de la même population conduisent à l'estimation de statistiques (moyenne, écart-type et intervalle de confiance) très différentes. En effet, l'estimation de l'effectif requis pour atteindre une précision de 90 % est respectivement de 16, 10 et 40 placettes. En observant les limites de l'intervalle de confiance, chacun des intervalles des trois échantillons (L,M,N) contient la moyenne de la population égale à 25,4 m^2/ha. Toutefois, lorsque l'on compare entre elles les moyennes de l'échantillon L et M, le test de t conclut à une différence significative, l'intervalle de confiance de l'échantillon M étant très étroit. Ceci permet de rappeler que l'on travaille en échantillonnage à un niveau de probabilité de 95 % et qu'il existe toujours 5 % de risque de se tromper dans l'interprétation des résultats.

On constate donc que l'effet du hasard peut être important et qu'il est essentiel de bien maîtriser les concepts statistiques afin d'en faire une interprétation correcte.

3.1.3 Échantillonnage et intervalle de confiance

a) Évaluation du paramètre

La définition de l'intervalle de confiance stipule que : « si l'on échantillonnait la population un très grand nombre de fois et que l'on construisait l'intervalle de confiance pour chaque échantillon, 95% de ces intervalles contiendraient la moyenne de la population ».

Afin d'illustrer la mise en application de ce concept, 40 échantillons aléatoires de 15 placettes ont été réalisés à partir de notre population de 81 placettes et les statistiques associées, comprenant les limites supérieures et inférieures des intervalles de confiance, ont été calculées pour chacun.

Malgré le fait qu'un échantillon constitué de 15 placettes peut être considéré comme étant de petite taille, il est important de mentionner que dans cet exemple, il représente un taux d'échantillonnage (15/81) de 18,5 %, ce qui est très élevé comparativement à l'échantillonnage effectué en milieu forestier.

b) Interprétation des résultats

On constate que pour trois échantillons sur un total de 40 (soit 7,5 %), les intervalles de confiance obtenus ne contiennent pas la moyenne de la population. Il s'agit des résultats des échantillons 22, 31 et 35 pour lesquels les limites supérieures de l'intervalle de confiance calculé se situent en-dessous de 25,4 m^2/ha. Toutefois, la limite supérieure obtenue pour le tirage 31 (25,0 m^2/ha) s'approche considérablement de la moyenne de 25,4 m^2/ha. On voit donc qu'en appliquant les concepts de probabilité reliés à l'intervalle de confiance, il existe toujours une probabilité, qui est de l'ordre de 95 %, d'obtenir un résultat d'échantillonnage représentatif de la vraie moyenne et un risque d'approximativement 5 % de se tromper.

3.2. Évaluation de critères de travaux dits commerciaux

3.2.1 Surface terrière par essence

a) Évaluation du paramètre

Nous avons vu que la surface terrière totale avait été mesurée avec un degré de précision élevé (93,2 %). Voyons maintenant, pour l'échantillon A, les statistiques pour la surface terrière par essence. Afin de simplifier la présentation, seules les principales essences sont considérées.

b) Interprétation des résultats

Pour l'érable à sucre, qui est la principale essence avec une occupation de 34 % de la surface terrière totale, la surface terrière moyenne obtenue est de 11,1 m^2/ha avec un intervalle de confiance allant de 8,1 à 14,1 m^2/ha pour une précision estimée de 73,1 %. On remarque que le coefficient de variation (72,1 %) (écart-type par rapport à la moyenne) a augmenté considérablement par rapport au total toutes essences (18,1 %). Pour être en mesure d'estimer la surface terrière en érable à sucre avec une précision de 80 %, il aurait fallu 54 placettes. Pour obtenir une précision de 85 %, 97 placettes auraient été nécessaires, alors que pour une précision de 90 %, il en aurait fallu 218. On réalise ici que l'obtention d'une bonne précision par essence, même pour une essence principale, peut nécessiter la réalisation d'un grand nombre de placettes.

Pour le peuplier à grandes dents, qui est la deuxième essence en importance (16 % de la surface terrière totale), la précision obtenue pour la surface terrière est de 40,0 %, pour le tilleul, elle est de 47,4 % et pour le bouleau jaune, elle est de 40,6 %. Pour ces essences, on constate que l'écart-type est plus grand que la moyenne (coefficient de variation plus grand que 100 %). Pour ces essences de moindre importance, il aurait fallu plus de 200 placettes pour obtenir une précision de 80 % seulement et plus de 800 pour une précision de 90 %. On peut conclure qu'il est illusoire de penser obtenir une bonne précision pour les essences secondaires compte tenu des moyens qui devraient alors être investis. Il faut d'ailleurs toujours garder en tête que les estimations par essence comportent une part d'incertitude qui s'accroît en fonction de la marginalité de l'essence considérée.

3.2.2 Surface terrière par essence par classe de vigueur

a) Évaluation du paramètre

Voyons maintenant ce que deviennent les statistiques lorsque l'on subdivise davantage et que l'on considère la surface terrière par essence, par classe de vigueur.

b) Interprétation des résultats

Considérant l'essence principale, soit l'érable à sucre, la surface terrière moyenne obtenue pour cette essence, par classe de vigueur, est petite et leur écart-type est du même ordre que leur moyenne respective. Le niveau de précision se situe autour de 60 % (entre 58 et 66 %) pour les principales classes de vigueur 1, 3 et 4. Pour obtenir une précision d'au moins 80 % pour les érables à sucre de vigueur 1, 3, ou 4, il aurait fallu faire entre 86 et 130 placettes.

Dans cet échantillon constitué de 30 placettes, une seule tige d'érable à sucre de vigueur 2 a été trouvée, correspondant à une moyenne de 0,07 m^2/ha. Il s'agit d'un évènement beaucoup trop rare pour que l'on puisse considérer quelque statistique que ce soit à ce sujet.

Pour les essences secondaires, les niveaux de précision obtenus sont plus bas et présentent davantage de variations d'une classe de vigueur à l'autre. Ainsi, pour le bouleau jaune de vigueur 1, la précision n'est que de 15,1 % et pour la classe de vigueur 3 de 31,9 %, les autres classes étant trop faiblement représentées pour faire des calculs statistiques. Une évaluation des bouleaux jaunes de vigueur 1 avec une précision de 80 % nécessiterait 541 placettes et 2 164 placettes pour atteindre une précision de 90 %. On constate donc que, plus on subdivise les données, moins les résultats par critères considérés sont précis, jusqu'à atteindre des niveaux de précision si bas que l'utilisation des valeurs moyennes devient hasardeuse. Si l'on désire subdiviser les données et obtenir un niveau de précision acceptable, le nombre de placettes devrait être augmenté de façon irréaliste.

Dans cette base de données, la présence de précisions négatives (érable à sucre de vigueur 2, bouleau jaune de vigueur 2 et 4) est une autre aberration démontrant les limites d'utilisation de cette statistique. Il s'agit également d'un signal d'alarme quant à la non-fiabilité du paramètre estimé.

3.2.3 Prélèvement par essence

a) Évaluation du paramètre

Afin d'observer les effets d'une subdivision de l'échantillon sur les statistiques du prélèvement, voyons maintenant les résultats obtenus pour chacune des principales essences.

b) Interprétation des résultats

Le prélèvement moyen de l'érable à sucre, qui est l'essence principale, est de 0,323 avec un intervalle de confiance allant de 0,258 à 0,389 et une précision de 79,6 %, alors que la précision sur le ratio toutes essences était de 93,3 %. Il aurait fallu 124 placettes pour atteindre une précision de 90 %. Le prélèvement du peuplier à grandes dents est de 0,468 avec une précision de 77,2 % et 156 placettes auraient été nécessaires pour atteindre une précision de 90 %. Pour les autres essences, un peu plus marginales, la précision chute à 59,1 % pour le bouleau jaune avec une moyenne de 0,212 et un intervalle de confiance allant de 0,125 à 0,299, et une précision de 31,8 % pour le tilleul avec une moyenne de 0,222 et un intervalle de confiance allant de 0,071 à 0,374. Pour ces essences, le nombre de placettes pour atteindre un niveau de précision jugé acceptable de 80 % impliquerait la réalisation de 126 placettes pour le bouleau et 349 pour le tilleul, pour atteindre 90 % de précision, il aurait fallu 503 et 1 397 placettes respectivement. On constate donc que l'atteinte d'objectifs établis en termes de prélèvement par essence engendrerait des coûts prohibitifs pour réaliser un échantillonnage avec un niveau de précision acceptable. À ce niveau de détail (prélèvement par essence), il faut donc accepter de fonctionner avec des intervalles de confiance très grands et utiliser ces résultats qu'à titre indicatif.

3.2.4 Capital forestier

L'échantillon I est constitué de 57 placettes au prisme de facteur 2 issues d'un peuplement forestier feuillu mature. Afin de déterminer si ce peuplement est éligible à certains travaux de récolte, on doit évaluer les paramètres de surface terrière totale, de surface terrière du capital forestier (surface terrière des tiges qui ne risque pas de mourir d'ici la prochaine coupe) et de surface terrière du capital forestier en croissance (surface terrière des tiges dont le bois ne risque pas de se dégrader d'ici la prochaine coupe).

a) Évaluation du paramètre

Les moyennes ont été calculées à l'aide de la formule 2, les écart-types à l'aide de la formule 4(5) et l'intervalle de confiance à l'aide des formules 14 et 15.

b) Interprétation des résultats

La surface terrière totale est en moyenne de 24,9 m^2/ha, avec un écart-type de 6,5 m^2/ha et un intervalle de confiance allant de 23,2 à 26,7 m^2/ha. La surface terrière du capital forestier est en moyenne est de 16,2 m^2/ha, avec un écart-type de 6,4 m^2/ha et un intervalle de confiance allant de 14,5 à 17,9 m^2/ha. La surface terrière du capital forestier en croissance est en moyenne de 15,8 m^2/ha avec un écart-type de 6,4 m^2/ha et un intervalle de confiance allant de 14,1 à 17,5 m^2/ha.

À titre d'exemple, si l'on établit pour un peuplement feuillu que pour être admissible à la coupe de jardinage, la surface terrière totale doit être d'au moins 24 m^2/ha, la surface terrière du capital forestier doit être d'au moins 17 m^2/ha et la surface terrière du capital forestier en croissance doit être d'au moins 9 m^2/ha. Alors, selon ces critères, ce peuplement ne rencontre pas les exigences relativement à la réalisation du jardinage, puisque la surface terrière du capital forestier estimé est inférieure à 17 m^2/ha.

À la lumière des résultats obtenus pour l'échantillon I, il est plausible de croire que la réalisation d'un autre échantillon (indépendant de l'échantillon I) pourrait mener à des conclusions différentes quant à la justification de l'admissibilité du traitement.

3.3. Évaluation de critères de travaux dits précommerciaux

Dans un jeune peuplement résineux de 230 ha destiné à des travaux dits précommerciaux, l'échantillon C est constitué de 460 placettes de 4 m^2 regroupées en grappe de 10 placettes, soit 46 points d'observation (ou grappes de placettes). Cet échantillon sert de référence pour le calcul des variables :

- Nombre de tiges total/ha
- Nombre de tiges résineuses/ha
- Distribution des tiges résineuses

3.3.1 Nombre de tiges total/ha

a) Évaluation du paramètre

En fonction du plan de sondage réalisé, le point d'observation correspond à l'information contenue dans une grappe de placette. Les statistiques sont donc calculées sur la base du point d'observation correspondant au nombre de tiges total par 40 m^2.

b) Interprétation des résultats

Pour l'échantillon C, le nombre de tiges total est en moyenne de 12 935 tiges/ha, avec un écart-type de 7 249 tiges/ha, un intervalle de confiance allant de 10 782 à 15 088 tiges/ha pour une précision de 83,4 %. On constate ici que les résultats obtenus comportent une grande variabilité. Pour atteindre une précision de 90 %, il aurait fallu un total de 127 grappes de placettes, soit 81 grappes de placettes supplémentaires.

À titre d'exemple, si l'on établit que pour qu'un jeune peuplement résineux soit admissible à la coupe d'éclaircie précommerciale, le nombre de tiges total doit être au minimum de 4 000 tiges/ha, alors l'admissibilité de l'échantillon C ne pose aucun doute, la limite inférieure de l'intervalle de confiance (10 782 tiges/ha) étant bien au-dessus du seuil minimal d'admissibilité.

Cependant, une fois sur le terrain, il faut s'attendre à une très grande variabilité dans les conditions du peuplement (distribution du nombre de tiges total à l'hectare) puisque nous sommes en présence d'un coefficient de variation élevé de 56 %, l'étendue des données de l'échantillon allant de 2 250 à 30 250 tiges/ha.

3.3.2 Nombre de tiges résineuses/ha

a) Évaluation du paramètre

Tout comme dans le cas des résultats de suivi d'interventions commerciales, la subdivision des données a des répercussions sur la précision des résultats. Nous pouvons utiliser le nombre de tiges résineuses de l'échantillon C pour apprécier cet impact.

b) Interprétation des résultats

Pour le nombre de tiges résineuses, la moyenne est estimée à 4 060 tiges/ha avec un écart-type de 2 768 tiges/ha et un intervalle de confiance allant de 3 238 à 4 882 tiges/ha pour une précision de 79,7 %. Pour évaluer le nombre de tiges résineuses avec une précision de 90 %, il aurait fallu approximativement 189 grappes de placettes. On observe des résultats similaires pour l'épinette qui est la principale essence résineuse avec une précision de 79,2 %. Pour le bouleau et le peuplier, les moyennes sont respectivement de 1 125 et 940 tiges/ha. Pour ces essences, une estimation avec une précision de 90 % aurait nécessité 500 et 783 grappes de placettes. Enfin, pour les autres essences moins abondantes, la précision est inférieure à 50 % et il aurait fallu un nombre considérable de grappes de placettes pour évaluer leur nombre avec précision. Encore cette fois, on peut constater que plus on subdivise la population, moins les données sont précises et l'imprécision s'accroît à mesure que l'on considère des entités dont l'occurrence est de plus en plus marginale, menant même à des résultats aberrants (précision négative) comme dans le cas du pin gris.

3.3.3 Distribution des tiges résineuses

a) Évaluation du paramètre

La distribution est définie par la proportion de la couverture occupée par une tige sur la base de la présence d'une tige à tous les 4 m^2. Même s'il s'agit d'une proportion, la moyenne de la distribution peut être estimée à l'aide de la formule de base de la moyenne (formule 2) parce que le nombre de placettes par grappe est le même pour toutes les grappes. Dans le cas où des grappes de tailles différentes étaient réalisées, c'est l'estimation quotient (formule 3) qu'il serait nécessaire d'utiliser.

b) Interprétation

Pour l'échantillon C, la distribution des tiges résineuses a une moyenne estimée de 0,650 avec un intervalle de confiance allant de 0,593 à 0,707 pour une précision de 91,3 %. On estime à approximativement 35 le nombre de grappes suffisant ayant permis d'atteindre une précision de 90 % pour ce paramètre.

3.3.4 Nombre de tiges résiduelles et éclaircies

a) Évaluation du paramètre

Dans un jeune peuplement résineux de 145 ha dans lequel des travaux d'éclaircie précommerciale ont été réalisés, l'échantillon J est constitué de 290 placettes de 4 m^2 regroupées en grappe de 10 placettes, soit 29 points d'observation (ou grappes de placettes). Cet échantillon sert de référence pour le calcul du nombre de tiges éclaircies/ha et du nombre de tiges résiduelles/ha.

b) Interprétation des résultats

Le nombre de tiges résiduelles a une moyenne estimée à 2 284 tiges/ha avec un intervalle de confiance allant de 2 060 à 2 509 tiges/ha pour une précision de 90,2 %. Pour ce qui est du nombre de tiges éclaircies, la moyenne est estimée à 1 850 tiges/ha, avec un intervalle de confiance allant de 1 657 à 2 043 tiges/ha et une précision de 89,6 %.

À titre d'exemple, supposant que pour être considérées comme bien traitées, le nombre de tiges résiduelles doit être inférieur à 3 125 tiges/ha, l'échantillon J rencontre ce critère aisément. Considérant que pour le nombre de tiges éclaircies, il y ait un seuil, fixé à 1 875 tiges/ha, autour duquel on considère le peuplement comme étant de faible ou de forte densité, l'échantillon J se classe parmi les peuplements de faible densité.

3.3.5 Nombre de plants reboisés et nombre de plants reboisés conformes

a) Évaluation du paramètre

Suite à la récolte d'un peuplement résineux mature de 225 ha, une plantation d'épinette blanche a été réalisée. L'échantillon K est constitué de 75 placettes à rayon fixe de 5,64 m (100 m^2). Cet échantillon sert de référence pour estimer le nombre de plants reboisés ainsi que le nombre de plants reboisés conformes, c'est-à-dire les plants reboisés qui respectent les conditions établies de mise en terre.

b) Interprétation des résultats

Les résultats obtenus pour l'échantillon K indiquent que le nombre de plants reboisés est de 1 515 plants/ha, avec un écart-type de 519 plants/ha, un intervalle de confiance allant de 1 395 à 1 634 plants/ha et une précision de 92,1 %. Dans ce cas, 47 placettes auraient été suffisantes pour atteindre une précision de 90 %. Pour ce qui est des plants reboisés conformes, la moyenne estimée est de 1 472 plants/ha avec un intervalle de confiance allant de 1 352 à 1 592 plants/ha et une précision de 91,9 %.

À titre d'exemple, en supposant que le nombre de plants conformes doit se situer entre 1 500 et 2 200 plants/ha pour être considéré acceptable, l'échantillon K ne rencontre pas le minimum fixé en terme de plants conformes/ha.

3.4. Comparaison de deux échantillons effectuée dans le même peuplement forestier

On doit d'abord vérifier l'égalité des variances afin d'être en mesure de choisir le test t de « Student » approprié.

3.4.1 Surface terrière par essence

Pour cet exemple, nous utiliserons les estimations des échantillons A et B.

a) Évaluation du paramètre

Le test d'égalité des variances est d'abord effectué. Dans le cas où les variances de l'échantillon sont déclarées égales (totale et érable à sucre), ce sont les formules 21 et 22 du test de t de « Student » qui sont utilisées. Dans le cas où les variances de l'échantillon sont significativement différentes (bouleau jaune, peuplier à grandes dents et tilleul), ce sont les formules 23 et 24 du test de t de « Student » qui sont utilisées.

Le bon choix du test de t est important lorsque la taille des échantillons comparés est différente. Les deux tests de t donnent des résultats identiques seulement lorsque les deux échantillons sont de même taille.

b) Interprétation des résultats

Nous avons déjà vu que les surfaces terrières totales des échantillons A et B étaient significativement différentes. La valeur de t_0 calculée étant plus grande que la valeur de t de la table, signifiant que la distance entre les deux moyennes (numérateur de la formule du test de t) est plus grande que l'estimation de la demi-largeur de l'intervalle de confiance moyen pondéré (dénominateur de la formule du test de t) des deux échantillons. Au niveau des essences individuelles, aucune différence significative n'est observée. C'est l'effet cumulatif de la surface terrière par essence toujours plus grande dans l'échantillon A qui se fait ressentir au niveau de la surface terrière totale. De plus, les intervalles de confiance par rapport aux moyennes des estimations par essences sont plus grands, rendant plus difficile la détection de différences significatives.

Il est également possible d'obtenir les résultats du test de comparaison des variances (test de F) et de comparaison des moyennes (test de T) au moyen des formules de excel. Ces formules renvoient la probabilité associée à la valeur de F ou la valeur de T calculée. Ainsi, une probabilité inférieure à 0,05 sera déclarée significativement différente à un niveau de probabilité de 95 %.

3.4.2 Prélèvement par essence

a) Évaluation du paramètre

Le test d'égalité des variances est effectué. Dans le cas où les variances de l'échantillon sont déclarées égales (érable à sucre), ce sont les formules 21 et 22 du test de t de « Student » qui sont utilisées. Dans le cas où les variances de l'échantillon sont significativement différentes (bouleau jaune, peuplier à grandes dents, tilleul et totale), ce sont les formules 23 et 24 du test de t de « Student » qui sont utilisées.

b) Interprétation des résultats

Le test d'égalité des variances démontre qu'au niveau du prélèvement, trois essences sur quatre ainsi que le total ont une différence significative, les variances de l'échantillon A étant toujours plus faibles, suggérant un comportement différent dans ce cas.

Ici aussi, bien que des différences significatives soient présentes au niveau du prélèvement total, aucune différence significative au niveau des essences n'est détectée, due aux fortes variations présentes à ce niveau.

4. Autres considérations relatives à la statistique appliquée à la foresterie

4.1 Statistique appliquée et justification de l'utilisation d'une variable

Dans les forêts du domaine public du Québec, de nombreuses variables servent à évaluer les conditions des peuplements forestiers. Comme le nombre de variables d'évaluation est très élevé, il y a lieu de se demander si toutes les variables ont le même poids dans la prise de décision.

Ainsi, il est correct de fixer comme objectif que les résultats d'évaluation des variables concernant les peuplements forestiers commerciaux atteignent une précision supérieure à 80 % puisque plus la précision est élevée, plus les résultats peuvent être considérés comme fiables. Toutefois, à la lumière des exemples présentés précédemment, on constate qu'il est impossible d'atteindre une précision de 80 % pour toutes les variables évaluées.

Les difficultés surviennent lorsque l'on génère des sous-échantillons provenant de l'échantillon principal. Il faut d'abord rappeler, à titre indicatif, que le taux d'échantillonnage pour les évaluations en forêt feuillue est approximativement 1/125 de la superficie en considérant toutes les tiges présentes dans chacune des placettes. À partir du moment où seulement certaines essences sont considérées (un sous-échantillon), le taux d'échantillonnage est diminué de façon inversement proportionnelle à la surface terrière de ces essences, conduisant selon toute probabilité à une plus grande variabilité dans l'estimation des résultats et une précision inférieure. Si l'on poursuit le raisonnement et que l'on veuille évaluer, par exemple, le prélèvement par essence (un sous-échantillon du sous-échantillon), le taux d'échantillonnage est une fois de plus dilué et il devient beaucoup plus difficile, voire impossible, d'obtenir une estimation fiable comme dans certains exemples de la section 3.

Sur une base statistique, il est raisonnable d'affirmer que **les variables pour lesquelles l'atteinte d'une précision de 80 à 90 % est réaliste, devraient être utilisées pour juger de l'acceptation d'un traitement**. Dans le cas où les variables sont marquées d'un intervalle de confiance moyennement grand, soit une précision estimée entre 51 et 80 %, ces résultats peuvent faire l'objet d'une certaine tolérance mais doivent être utilisés avec prudence, en leur accordant un poids moins important dans l'évaluation d'un traitement. Dans le cas où les variables sont caractérisées par de grands intervalles de confiance, une précision inférieure à 50 %, alors ces variables ne devraient pas être utilisées et ne devraient pas influencées la prise de décision.

L'usage de la statistique permet de distinguer le niveau de confiance que l'on peut accorder à l'estimation d'une variable. En présence de variables non fiables (précision < 50 %), il est du ressort des professionnels de s'interroger sur les objectifs visés et d'identifier d'autres moyens leur permettant de prendre une décision.

4.2 Statistique appliquée et utilisation d'un seuil prédéterminé fixe d'évaluation d'une variable

Pour chacune des variables évaluées, un seuil fixe d'acceptabilité fait en sorte que le résultat est soit accepté, soit refusé, sans zone grise. Une telle méthode est pratique et facile d'application lorsque l'on évalue des éléments ayant une grande précision. Toutefois, comme nous travaillons dans le milieu forestier et vivant, et que les estimations sont empreintes d'une grande variabilité, il arrive que la prise de décision, lorsque les résultats sont près des seuils, exige d'aller plus loin.

Afin de prendre des décisions, l'utilisation de seuil fixe est nécessaire. D'un autre coté, l'utilisation des statistiques comme l'intervalle de confiance permet de réaliser et de comprendre que l'évaluation des variables n'est pas exacte, que l'estimation des paramètres est obtenue au moyen d'un échantillonnage et que les statistiques sont caractérisées par un niveau de probabilité et une erreur d'échantillonnage. Cependant, malgré l'utilisation des statistiques, il faut se rappeler qu'en regard d'un seuil, la moyenne demeure toujours la référence.

À titre d'exemple, on fixe à 24 m^2/ha la surface terrière initiale minimale pour qu'un peuplement soit considéré apte à une intervention de coupe partielle. Si un premier échantillon de 30 placettes fournit une estimation de la moyenne de 24,5 m^2/ha, ce résultat est en haut du seuil et autorise l'intervention. Un deuxième échantillon indépendant effectué dans le même peuplement et comprenant le même nombre de placettes pourrait fournir une estimation de 23,5 m^2/ha qui cette fois est en bas du seuil fixé et n'autorise pas l'intervention.

En regard du seuil minimal fixé et des deux moyennes estimées, les résultats mènent à deux décisions différentes.

Du point de vue statistique, en considérant que ces deux estimations possèdent une précision de 90 % (24,5 ± 2,45 m^2/ha et 23,5 ± 2,35 m^2/ha), un test de *t* de Student confirmerait qu'il n'y a pas de différence significative entre les échantillons sachant que le hasard a une part à jouer dans l'estimation de la moyenne de chaque échantillon.

C'est donc ici que la statistique et l'utilisation d'un seuil se confronte.

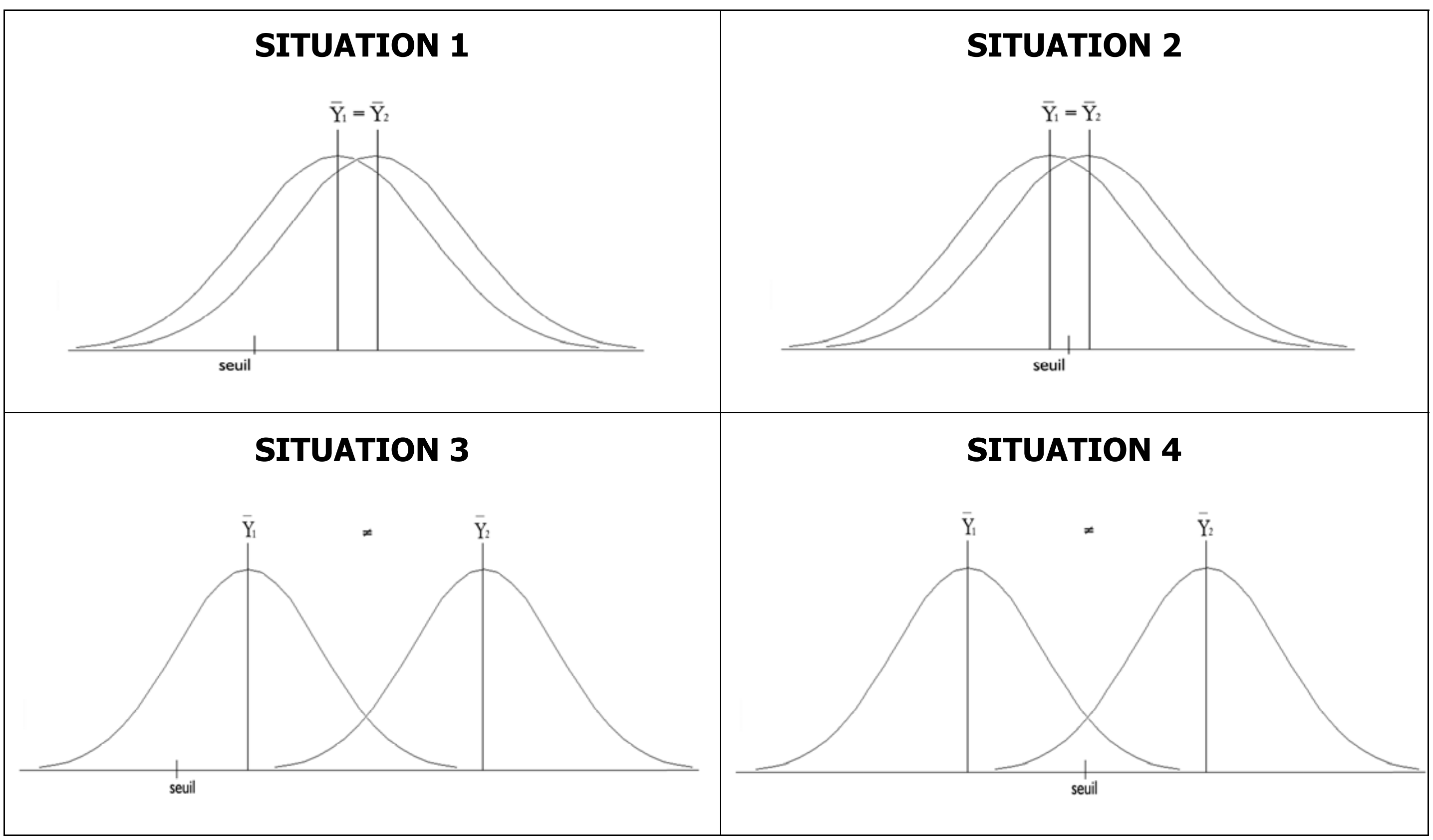
Dans cet exemple, il faut comprendre que **l'objectif est de traiter des peuplements dont la surface terrière est la plus élevée possible**, mais que cette surface terrière doit être supérieure à 24 m^2/ha. La surface terrière devrait donc, en principe, **être le plus loin possible du seuil pour ainsi éviter les conflits et les mauvaises interprétations**. Toutefois, dans le cas où le résultat est près du seuil, comme dans l'exemple ci-dessus, il faut comprendre qu'il est possible de prendre deux décisions différentes sans être en mesure d'affirmer qu'un résultat soit meilleur que l'autre puisqu'on ne connaît pas la vraie moyenne de la population.

4.3 Comparaison des estimations de deux échantillons et utilisation d'un seuil prédéterminé fixe d'évaluation d'une variable

La comparaison des estimations de deux échantillons devrait être effectuée au moyen du test de t de « Student ». L'utilisation du test de t est justifiée parce qu'il permet de calculer une « erreur standard pondérée » des deux échantillons comparés. Ainsi, les résultats de l'un ne servent pas de référence pour l'autre mais sont utilisés tous les deux de façon complémentaire.

Il existe quatre situations possibles de comparaison des moyennes par rapport à la position du seuil d'acceptabilité telles qu'illustrées à la figure 17.

Figure 17 Illustration des quatre situations de comparaison des moyennes par rapport à la position du seuil d'acceptabilité.



Situation 1 : Il n'y a pas de différences significatives entre les moyennes et elles se retrouvent du même côté du seuil d'acceptabilité.

Ce cas est le plus simple et ne pose aucun problème d'interprétation, les résultats sont semblables et mènent à la même décision.

Situation 2 : Il n'y a pas de différences significatives entre les moyennes mais il y en a une de chaque côté du seuil d'acceptabilité.

Lorsque les moyennes sont déclarées comme n'ayant pas de différence significative, mais que celles-ci conduisent à des décisions différentes, diverses procédures peuvent être envisagées.

En terme de solution, il est d'abord possible d'effectuer un nouvel échantillonnage conjoint qui servira à fournir le résultat définitif ou encore, il est possible de convenir à l'avance que dans la situation où le peuplement est trop près de la limite du seuil, les interventions seront abandonnées pour le moment et reportées dans quelques années de façon à ce que le peuplement puisse croître et éventuellement se situer bien au-delà des limites du seuil.

Situation 3 : Les deux moyennes sont significativement différentes, mais du même côté du seuil d'acceptabilité.

Les deux résultats mènent à la même décision et confirment l'acceptation ou le refus du critère. Au niveau de l'interprétation du résultat, on peut cependant se demander pourquoi une différence significative a été observée. À cet égard, il serait nécessaire de vérifier si la différence observée n'est pas le résultat de l'introduction d'un biais et, voir si, le cas échéant, la cause peut être identifiée. Les correctifs nécessaires pourraient alors être apportés.

Situation 4 : Les deux moyennes sont significativement différentes et il y en a une de chaque côté du seuil d'acceptabilité.

Résultats significativement différents qui confirment la décision. Toutefois, au niveau de l'interprétation, lorsque deux estimations indépendantes de la même variable, dans le même peuplement, avec un nombre semblable de placettes sont significativement différentes, il y a lieu de s'inquiéter. D'abord, l'utilisation d'un niveau de probabilité de 95 % signifie que les résultats statistiques nous amèneront à faire une déclaration erronée dans 5 % des cas. Il y a donc 5 % de chance de se tromper lorsque l'on déclare que deux moyennes sont significativement différentes.

En terme de solution, il est possible d'appliquer une stratégie permettant de confirmer sans équivoque la présence de différences significatives. Comme la notion de probabilité nous indique que l'on risque de se tromper dans 5 % des cas, on doit porter notre jugement sur un grand nombre de cas et non sur un seul.

Ainsi, à titre d'exemple, si l'on fait l'estimation d'une variable au moyen de deux échantillons indépendants dans un peuplement (population) et que l'on obtient une différence significative; comme il y a un niveau de probabilité de 95 % d'associé à ce résultat, il est toujours possible d'invoquer le fait qu'il existe 5 % de chance de se tromper dans l'interprétation de ce résultat. Ce fait est important mais peu convaincant au niveau de l'interprétation des résultats.

Afin de s'assurer de faire une bonne interprétation concernant un intervenant, on pourrait faire l'estimation d'une variable au moyen de deux échantillons indépendants dans vingt peuplements (populations) différents au cours d'une saison. Alors, si l'on obtient un cas démontrant une différence significative, cela permet seulement de confirmer l'application de la théorie que l'on risque de se tromper dans un cas sur vingt (5 %) et donc l'ensemble des évaluations couvrant les vingt peuplements devraient être déclarés comme étant non différents. Par contre, si parmi les vingt peuplements, il y a plus d'un cas qui sont estimés comme étant significativement différents, on pourra confirmer que nous sommes vraiment en présence de différences significatives.

5. Outils informatiques disponibles

5.1 Logiciels existants

5.1.1 Chiffrier Microsoft Excel

Le chiffrier Microsoft Excel 2002 comprend 78 fonctions statistiques dans le sous-menu *Fonction* du menu *Insertion*. Ces fonctions couvrent une grande gamme de statistiques, comme : la somme, le dénombrement, la moyenne, la variance de l'échantillon, l'écart-type, la somme des carrés, la somme des produits, la valeur de Z, la valeur de F, la valeur de t de Student, le test de F sur la variance de deux échantillons et le test de t de Student sur la moyenne de deux échantillons. Il s'agit des fonctions statistiques utilisées dans le cadre de la présente formation.

De plus, Microsoft Excel propose un ensemble d'outils d'analyse de données, appelé *Utilitaire d'analyse*, qu'il est possible d'utiliser pour éviter certaines étapes lors du développement d'analyses statistiques ou techniques plus complexes. En fournissant les données et les paramètres pour chaque analyse, l'outil utilise les fonctions macro statistiques ou techniques appropriées, puis affiche les résultats dans une table de sortie. Certains de ces outils génèrent des graphiques en plus des tables de sortie.

Excel propose bien d'autres fonctions statistiques, financières et techniques associées aux feuilles de calcul. Certaines fonctions statistiques sont intégrées, d'autres deviennent disponibles uniquement lorsqu'on installe l'*Utilitaire d'analyse*.

L'*Utilitaire d'analyse* comprend les outils décrits ci-dessous. Pour accéder à ces outils, il suffit de cliquer sur *Utilitaire d'analyse* dans le menu *Outils*. Si la commande *Utilitaire d'analyse* n'est pas disponible, on doit charger la *macro complémentaire Utilitaire d'analyse*.

- Analyse de variance
- Analyse de corrélation
- Analyse de covariance
- Statistiques descriptives
- Lissage exponentiel
- Analyse de Fourier
- Histogramme
- Moyenne mobile
- Rang et centile
- Régression
- Échantillonnage

En plus des fonctions et de l'utilitaire d'analyse compris avec Microsoft Excel, il existe d'autres modules ou « compagnons » (*add-in*) disponibles sur Internet pour augmenter la convivialité et la puissance de Microsoft Excel pour effectuer des analyses statistiques, tels que *WinSTAT*, *XLSTAT, Analyse-it*, etc.

5.1.2 Base de données Microsoft Access

Microsoft Access est un logiciel qui permet de gérer des bases de données. À partir de ce logiciel, il est possible d'exporter les données vers d'autres applications, comme Microsoft Excel ou SAS par exemple, afin d'y calculer des statistiques.

Néanmoins, il est possible de calculer des statistiques dans Access, mais ce n'est pas convivial. Par exemple, pour calculer la moyenne du nombre de tiges par hectare dans un peuplement, il faudrait créer une requête permettant de calculer dans des champs différents le nombre total de tiges de toutes les placettes et le nombre de placettes; puis, calculer le ratio de ces deux champs dans un troisième pour obtenir la moyenne. Aussi, il est possible de programmer en Visual Basic ou SQL dans Microsoft Access afin d'obtenir de tels résultats.

L'utilité la plus pertinente de l'outil Access dans une situation d'analyse statistique est son apport dans la préparation des données. Quelques exemples de traitements possibles étant :

- la fusion de plusieurs champs pour la création de clé;
- la fusion de plusieurs sources de données;
- les corrections de format de données.

5.1.3 Progiciels SAS/STAT ou SPSS

SAS et SPSS offrent une très large gamme de produits spécialisés pour effectuer des analyses et des statistiques (descriptives et inférentielles) et le plus commun pour effectuer des analyses de variance et des régressions est SAS/STAT. Contrairement à Excel, l'environnement SAS est moins convivial, mais plus performant et adéquat (McCallough et Wilson, 1999). C'est pourquoi il est très populaire auprès des chercheurs et des statisticiens qui peuvent se permettre d'investir du temps dans l'étude de sa programmation.

5.2 Outils spécialisés pour la compilation des données des interventions forestières

5.2.1 Logiciel TIGE

Le logiciel TIGE a été développé par le MRN pour permettre la saisie des données des inventaires d'intervention en forêt feuillue et réaliser des compilations sur le terrain en plus de créer une banque de données sur la base des peuplements, des placettes et des tiges (MRN, 2002). « Les principales opérations que peut effectuer ce module [le module MICRO du logiciel TIGE] sont la création et la gestion d'une banque de données, la compilation de données selon divers critères (essence, diamètre, vigueur, etc.), la production de fichiers compilés et la création de rapports permettant une analyse précise et rapide des travaux sylvicoles, afin de parfaire la connaissance du domaine forestier » (Tiré de MRN, 2002). TIGE permet d'obtenir des moyennes et de calculer le facteur q de Liocourt correspondant à la courbe théorique idéale de distribution des tiges en fonction de leur diamètre dans un peuplement inéquienne. Il ne permet cependant pas d'effectuer des statistiques descriptives de l'étendue des données ou des statistiques inférentielles.

5.2.2 Compilateurs de la direction de l'assistance technique (MRNF)

La direction de l'assistance technique (DAT) a mis au point des compilateurs permettant de vérifier certains résultats comme : la moyenne, la précision, le nombre de placettes à réaliser et l'atteinte des seuils de certaines variables associées aux traitements sylvicoles présentés au tableau 5.

Tableau 5 : Liste des compilateurs de la direction de l'assistance technique

Traitements sylvicoles (travaux non commerciaux)

Nom du compilateur	**Nature de traitement du compilateur**
Pre.xls	Préparation de terrain
Pl_reg.xls	Plantation ou regarni de la régénération
Deg_av.xls	Dégagement de la régénération (avant traitement)
Deg_ap.xls	Dégagement de la régénération (après traitement)
Epc_av.xls	Coupe d'éclaircie précommerciale (avant traitement)
Epc_ap.xls	Coupe d'éclaircie précommerciale (après traitement)
Ens.xls	Ensemencement des pins par voie aérienne ou terrestre (excluant les miniserres)
Enr.xls	Enrichissement de pins ou de feuillus (après traitement)

Suivi des interventions des années antérieures

Nom du compilateur	Nature de traitement du compilateur
Siaa_pl.xls	Suivi dans une plantation ou dans un regarni pour obtenir l'équivalent d'une plantation
Siaa_cr.xls	Suivi dans un ensemencement ou dans un regarni pour obtenir l'équivalent d'un peuplement récolté

Estimation des volumes de bois affectés par les opérations de récolte (VAOR) ou inventaire de la matière ligneuse non récoltée (MLNU)

Nom du compilateur	Nature de traitement du compilateur
Iml_cub.xls	Programme de cubage des tiges debout, des tronçons et des souches
Iml_con.xls	Programme de calcul de précision de volume de bois récolté

Suivi de l'article 89 du Règlement sur les normes d'intervention (RNI)

Nom du compilateur	Nature de traitement du compilateur
Rni_89.xls	Programme d'évaluation du pourcentage d'occupation des sentiers et du taux de protection de la régénération

Les compilateurs de la DAT sont déposés dans le forum de discussion du MRNF. Or, étant donné que l'accès au forum est réservé uniquement au personnel du ministère, les bénéficiaires de CAAF peuvent obtenir une copie de tous ces programmes en s'adressant soit à la DAT ou aux responsables régionaux de l'assistance technique du MRNF.

6. Conservation des données

(tiré et adapté du Guide de mise en place des métadonnées pour les ressources Web du gouvernement du Canada, 3e édition –Juillet 2004)

6.1 Les métadonnées

Les métadonnées sont des renseignements structurés sur les caractéristiques d'un objet physique ou numérique. Les métadonnées jouent le même rôle qu'une étiquette. Tout comme d'autres étiquettes, les métadonnées fournissent de l'information au sujet du contenu d'un objet.

Par exemple, une boîte en fer-blanc scellée non étiquetée peut contenir de l'huile à moteur, des « fèves au lard » ou de la nourriture pour chats. La seule façon de savoir ce qu'elle contient est de l'ouvrir. Par contre, si elle porte une étiquette décrivant le contenu, vous êtes alors en mesure de décider si vous voulez l'acheter ou l'ouvrir.

Le gouvernement du Canada s'est donné des règles de normalisation pour les pages de ses sites Web. Celles-ci doivent contenir certains renseignements (métadonnées) obligatoires, exprimés d'une certaine façon (selon la norme).

6.2 Pourquoi des métadonnées

Ce type d'information facilite l'exploration des ressources de la même manière que l'affichage dans un supermarché facilite le repérage des produits ou que les catalogues de bibliothèque permettent d'accéder au rayon où se trouve l'information désirée. Les moteurs de recherche utilisent les métadonnées pour établir une meilleure correspondance entre les demandes de renseignements de l'utilisateur et les descriptions des ressources indexées par le moteur.

Des métadonnées adéquates permettent de trouver l'information que l'on recherche. Toutefois, pour y arriver, le contrôle de la qualité et la cohérence sont de rigueur. Dans une base de données, s'il manque des métadonnées essentielles, ou si les métadonnées sont inexactes ou erronées, les résultats obtenus par les moteurs de recherche en seront affectés.

Afin de contribuer à la conservation des données et faciliter le travail de récupération ultérieur des informations recueillies dans les forêts du domaine public au Québec, l'ensemble des intervenants pourrait avoir recours à une telle façon de faire pour la gestion de l'ensemble des bases de données compilées dans toutes les régions par l'ensemble des bénéficiaires de CAAF de la province.

6.3 Une norme commune

L'adoption d'une seule norme de métadonnées permettrait de s'assurer que les informations qui se trouveront sur les bases de données du MRNF et des bénéficiaires de CAAF seraient facilement compréhensibles par tous les utilisateurs. Dans le cas de la gestion des forêts du domaine public du Québec, une seule norme permettrait aux industriels et aux fonctionnaires provinciaux d'effectuer la recherche d'information sur les bases de données de toute la province, sans avoir besoin de savoir quelle région ou quel organisme a produit cette information. Une telle façon de faire est à la base de l'acquisition de connaissances permettant d'étudier la dynamique des peuplements forestiers à long terme.

Le gouvernement du Canada s'est doté d'une norme sur les métadonnées appelée « Dublin Core ». Cette norme rend cinq éléments obligatoires pour la description des ressources Web du gouvernement du Canada. Ces cinq éléments obligatoires sont les suivants : Titre, Créateur, Date, Langue et Sujet. Les cinq éléments obligatoires ne constituent qu'un point de départ pour les organismes fédéraux qui se servent des métadonnées dans leur stratégie de gestion de l'information. On encourage l'expansion de l'ensemble des éléments de métadonnées par l'ajout d'autres éléments. Les cinq éléments obligatoires sont décrits selon des directives détaillées, afin d'en assurer la conformité.

Selon les mêmes principes appliqués par le gouvernement fédéral, les aménagistes forestiers devraient mettre en place une telle procédure et définir une structure universelle d'acquisition des connaissances visant à conserver toutes les données forestières recueillies sur le territoire de la province de Québec. Ceci, dans l'optique que les générations futures puissent en tirer profit, permettant une précieuse acquisition de connaissances concernant les multiples fonctions de caractérisation, d'évolution et de dynamique des peuplements forestiers.

CONCLUSION

L'utilisation des statistiques permet de réaliser que :

- l'on travaille dans un milieu hétérogène;
- les estimations sont empreintes d'une grande variabilité;
- l'on ne connaît jamais la vraie réponse.

Les statistiques permettent de distinguer le niveau de confiance que l'on peut accorder à l'estimation d'une variable. Avant de prendre des décisions d'importance, il serait prudent :

- que le jugement soit principalement basé sur les variables estimées avec la plus grande précision;
- que les estimations des variables contenant des intervalles de confiance moyennement grands ne soient utilisées qu'à titre indicatif;
- que les estimations des variables contenant de grands intervalles de confiance ne soient pas utilisées.
- que la comparaison des résultats d'un intervenant se fasse sur l'ensemble de travaux réalisés au cours d'une même saison;

Enfin, il importe de rappeler que la statistique est un outil d'aide à la décision et que le jugement professionnel a toujours sa place dans un processus de prise de décision.

REMERCIEMENTS

Nos remerciements s'adressent à M. Jacques Bélanger, professeur du département des sciences du bois et de la forêt de l'Université Laval et à M. Gilles Desaulniers, statisticien de la direction de l'assistance technique du ministère des Ressources naturelles et de la Faune du Québec pour la révision du contenu de cette formation. Des remerciements particuliers s'adressent à Mme Michèle Bernier-Cardou, statisticienne principale au Centre de foresterie des Laurentides du ministère des Ressources naturelles du Canada pour ses commentaires et ses précieux conseils, particulièrement sur l'application du théorème central-limite.

REFERENCES

BAILLARGEON, G. et J. RAINVILLE, 1976. Introduction à la statistique appliquée. Une approche multidisciplinaire. 3ième édition. Les Éditions SMG. Trois-Rivières. 490 pages + tables.

BÉDARD, S. et A. CARLE, 1996. Guide d'inventaire d'intervention en forêt feuillue. Institut québécois d'aménagement de la forêt feuillue (IQAFF). ISBN 0-662-80714-6, 62 p.

BLOUIN, D., A. PATRY et B. MÉNARD, 2001. Considérations sur les méthodes d'inventaire et leur précision. Rapport du CERFO, 2001-04. 20 p.

COCHRAN, W.G., 1977. Sampling Techniques (third edition). John Wiley & Sons inc. US, ISBN 0-471-16240-X, 428 p.

COLLIN, J., 2003. Dispositifs expérimentaux, notes de cours : BVG-60678. Université Laval. 16 chapitres + annexes.

CÔTÉ, S. et L. BÉLANGER, 1988. Étude de la régénération après épidémie de tordeuse des bourgeons de l'épinette. Rapport d'établissement du dispositif expérimental. Service des traitements sylvicoles, Direction de la sylviculture, MER. Québec.

DESAULNIERS, G., 2004. Comparaison de cinq types de placettes lors de l'échantillonnage dans le jardinage forestier en Outaouais. Forêt Québec, Direction de l'assistance technique. 47 p.

GOOD, P. I. et J. W. HARDIN, 2003. Common errors in statistics (and how to avoid them). John Wiley & sons, inc. US, ISBN 0-471-46068-0, 221 p.

McCALLOUGH, B. D. et B. WILSON, 1999. On the accuracy of statistical procedures in Microsoft Excel 97. Computational statistics and data analysis, vol. 31 : 27-37.

MENDENHALL, W., L. OTT ET R. L. SCHEAFFER, 1971. Elementary survey sampling. Belmont, California, Duxbury Press, 247 p.

MINISTÈRE DES RESSOURCES NATURELLES DU QUÉBEC, 2004a. Instructions relatives à l'application du règlement sur la valeur des traitements sylvicoles admissibles en paiement des droits – Exercice 2004-2005 (Version finale – Mai 2004). Forêt Québec, Dir. de l'assistance technique, Div. des traitements sylvicoles. ISBN 2-550-42610-X, 134 p.

MINISTÈRE DES RESSOURCES NATURELLES DU QUÉBEC, 2004*b*. Méthodes d'échantillonnage pour les inventaires d'intervention (inventaire avant traitement) et pour les suivis des interventions forestières (après martelage, après coupe et années antérieures) - Exercice 2004-2005 (Version finale – Mai 2004). Forêt Québec, Dir. de l'assistance technique, Div. des traitements sylvicoles. ISBN 2-550-42611-8, 389 p.

MRN, 2002. TIGE Manuel de l'usager v.3.04. Ministère des Ressources naturelles du Québec. 72 p. + 3 annexes.

ORDRE DES INGÉNIEURS FORESTIERS DU QUÉBEC (OIFQ), 2001. Guide de pratique professionnelle. 38 p.

ORDRE DES INGÉNIEURS FORESTIERS DU QUÉBEC (OIFQ), 1996. Manuel de foresterie. Les Presses de l'Université Laval. ISBN 2-7637-7479-2, 1428 p.

RONDEUX, J., 1993. La mesure des arbres et des peuplements forestiers. Les Presses agronomiques de Gembloux. ISBN 2-87016-041-0, 521 p.

STEEL, R.G.D., et J.H. TORRIE, 1980. Principles and procedures of statistics – A Biometrical Approach (Second edition). McGraw-Hill inc. US, ISBN 0-07-060926-8, 633 p.

SOKAL, R. A. et ROHLF, F. J., 1969. Biometry. W.H. Freeman, San Francisco. 776 pages.