Statistiques avec R

SCI 1018

Intervalles de confiance et stratégies d'échantillonnage

Marc J. Mazerolle

Département des sciences du bois et de la forêt, Université Laval

Avec révisions mineures de

Élise Filotas, *Département science et technologie, Université TÉLUQ*, et

Marc-Olivier Martin-Guay, *Département des sciences biologiques, Université du Québec à Montréal*

Table des matières

Introduction

Dans la leçon précédente, nous avons vu des concepts de base importants en statistique, notamment les caractéristiques d'une population et d'un échantillon, les variables aléatoires, les mesures de tendance centrale et de dispersion. Nous avons également présenté le théorème de la limite centrale et la loi des grands nombres. Dans la présente leçon, nous poursuivons avec le concept de l'échantillonnage ainsi que différentes stratégies afin de récolter un échantillon.

Échantillonnage

Il est peu pratique voire souvent impossible d'étudier la population au complet. On utilise plutôt un échantillon pour faire des inférences sur la population. L'échantillon est constitué d'**unités d'échantillonnage** (*sampling unit*). Par unité d'échantillonnage, on entend la plus petite unité indépendante d'une expérience ou d'une étude d'observation sur laquelle on prend une mesure. L'échantillonnage permet de récolter de l'information afin de répondre à des questions du genre :

- Quel est le nombre d'arbres par hectare affligés par l'agrile du frêne à Gatineau ?
- Combien y a-t-il de gisements d'or en Abitibi ?
- Combien d'heures un portable d'une certaine compagnie dure-t-il avant de tomber en panne ?
- Quelle partie de la population canadienne appuie la plus récente décision du gouvernement ?

Exemple 2.1 On veut faire un sondage téléphonique pour déterminer le nombre d'heures passées par les téléspectateurs devant leur téléviseur par jour. Un bon nombre de décisions peuvent influencer le résultat — la région, le choix des répon-

dants, la classe d'âge, la classe socio-économique, l'heure de l'appel. Que faire des biais potentiels associés à l'échantillonnage, comme les personnes sans téléphone, les personnes avec un numéro non affiché, les personnes ne voulant pas participer et les personnes possédant un afficheur ?

Plusieurs stratégies d'échantillonnage sont disponibles afin d'assembler un échantillon. Le choix de la stratégie d'échantillonnage dépend des caractéristiques de la population et des problèmes potentiels associés aux unités d'échantillonnage. Nous verrons plusieurs stratégies d'échantillonnage classiques dans les prochaines sections.

Peu importe la stratégie d'échantillonnage utilisée, l'échantillon doit être représentatif de la population afin de permettre une bonne estimation des paramètres qui nous intéressent. On fait face à un problème de représentativité si un groupe particulier d'unités d'échantillonnage parmi la population d'intérêt n'apparaît pas dans l'échantillon. Le paramètre estimé à partir d'un échantillon qui n'est pas représentatif de la population risque d'être biaisé.

1 Problèmes d'échantillonnage

Lors de l'échantillonnage, on présume souvent que :

- la variable d'intérêt est mesurée sans erreur sur chaque observation, individu ou unité d'échantillonnage ;
- les erreurs dans les estimations proviennent uniquement de la nature de l'échantillon (erreur d'échantillonnage).

L'erreur d'échantillonnage (*sampling error*) se définit comme étant la variabilité aléatoire d'un échantillon qui est tiré d'une population donnée. Par exemple, disons que l'on inscrive l'âge de chaque élève d'une école secondaire sur des bouts de papier à raison d'une valeur par bout de papier. S'il y a 1000 élèves, nous aurons une valeur d'âge pour chaque bout de papier. Si on sélectionne aléatoirement 50 observations à partir des élèves présents à

la cafétéria pendant l'heure du dîner, on obtiendra 50 bouts de papier sur lequel est inscrit l'âge d'un élève. On peut demander à un collègue d'échantillonner 50 observations à partir de la même population lors de la même période de dîner. Ce dernier obtiendra fort probablement un échantillon différent du premier. La variation qui existe entre ces deux échantillons de la même population est communément appelée l'erreur d'échantillonnage – elle reflète uniquement le fait que le processus d'échantillonnage amène une variabilité entre les échantillons. Toutefois, on peut avoir d'autres types d'erreurs plus sournoises qui peuvent biaiser nos estimations.

1.1 Portion de la population non représentée

L'impossibilité de sélectionner certains individus parmi la population amène des problèmes par rapport aux interprétations et aux conclusions que l'on peut tirer à propos de la population. Voici quelques exemples de ce problème :

- Certains sites sélectionnés en haute altitude ne peuvent être échantillonnés à cause de la température inclémente.
- Dans une étude sur la croissance des arbres, on ne peut pas mesurer la taille des grands arbres avec une échelle trop courte.
- Lors d'un sondage téléphonique, il est impossible de rejoindre les individus qui n'ont pas de téléphone et ceux qui ne veulent pas répondre au sondage.

Pour résoudre ce problème, on peut :

- utiliser de l'échantillonnage additionnel afin d'obtenir des observations provenant de la portion manquée de la population ;
- choisir une nouvelle variable qui permet de mesurer toute la population et qui est fortement corrélée à la variable d'intérêt (p. ex., le diamètre du tronc plutôt que la hauteur pour les arbres) ;
- redéfinir la population afin de mieux refléter ce qui a été mesuré au lieu de s'intéresser

à toute la population, on cible une partie particulière de cette population.

1.2 Erreurs de mesure ou de saisie des données

Parfois, des problèmes peuvent se glisser dans la mesure ou dans la saisie des données. Pour éviter ce type de problème, le contrôle de la qualité du travail, tant en ce qui concerne la mesure des données que la saisie et la gestion de celles-ci, s'avère la meilleure solution. Le meilleur moyen pour détecter les erreurs de mesure ou de saisie des données est d'utiliser des méthodes graphiques, lesquelles permettent d'identifier rapidement des valeurs extrêmes, aberrantes ou erronées, c'est-à-dire, des observations qui divergent nettement des autres.

1.3 Erreurs liées à la probabilité de détection

Certains organismes ou éléments d'intérêt sont difficiles à détecter lors d'un inventaire, malgré les efforts déployés pour standardiser le protocole. Par exemple, il est souvent impossible de détecter tous les éléments liés à un phénomène d'intérêt en raison de sa visibilité (p. ex., symptômes peu apparents pour une maladie), des conditions au moment de l'observation (p. ex., beaucoup de vent lors de l'écoute de chants d'oiseaux), de l'expérience de l'observateur (p. ex., jeune médecin analysant une radiographie), ou des conditions à l'endroit de l'observation (p. ex., couches géologiques empêchant la détection de gisements).

Le même problème se manifeste lorsqu'on étudie des populations humaines, par exemple lorsqu'on désire estimer la population de sans-abris au Canada. Le problème est similaire lorsqu'on s'intéresse à la présence d'un élément d'intérêt dans une série de sites. Ainsi, même si on ne détecte pas l'élément dans un site, cela ne signifie pas pour autant qu'il y est absente.

L'erreur associée à la détection imparfaite ne peut pas être contrôlée en standardisant le protocole. La meilleure méthode pour pallier le problème consiste à estimer explicitement la probabilité de détection avec des méthodes appropriées telles que les modèles de capture-marquage-recapture (CMR), les modèles d'échantillonnage de la distance (*distance sampling*) et les modèles d'occupation de site (*site occupancy analyses*).

2 Échantillonnage complètement aléatoire

L'échantillonnage complètement aléatoire est la stratégie d'échantillonnage la plus simple et celle qui offre le plus de flexibilité à l'étape des analyses. Avec cette approche, chaque unité d'échantillonnage a une chance égale d'être incluse dans l'échantillon et la sélection de chaque unité est le résultat d'un processus aléatoire. La sélection aléatoire peut être réalisée :

1. en écrivant les numéros de chaque unité d'échantillonnage sur des bouts de papier, de les mettre dans une urne, de bien mélanger, et de tirer un nombre donné de bouts papier au hasard ;
2. en utilisant une table de nombres aléatoires dans un livre de statistiques ;
3. en utilisant un générateur de nombres aléatoires dans `R`.

Pour générer des nombres aléatoires avec `R`, on peut utiliser une distribution uniforme (continue). Cette distribution se définit par

$$f(x|\text{min}, \text{max}) = \begin{cases} \frac{1}{(\text{max}-\text{min})} & \text{si min} \leq x \leq \text{max}, \\ 0 & \text{autrement .} \end{cases}$$

On remarque deux paramètres, soit le minimum de l'intervalle (min) et le maximum de l'intervalle (max). Cette distribution décrit des événements qui ont exactement la même probabilité de se produire pour un intervalle donnée (max − min, fig. 1).

Exemple 2.2 On veut déterminer la densité de probabilité associée à la valeur 3.46 dans un intervalle de 0 à 10 d'une distribution uniforme.

$$f(3.46|0, 10) = \frac{1}{(\text{max} - \text{min})} = \frac{1}{(10 - 0)} = 0.1$$

On peut aussi déterminer la densité associée aux valeurs 8, 6.15, ou 10 dans la

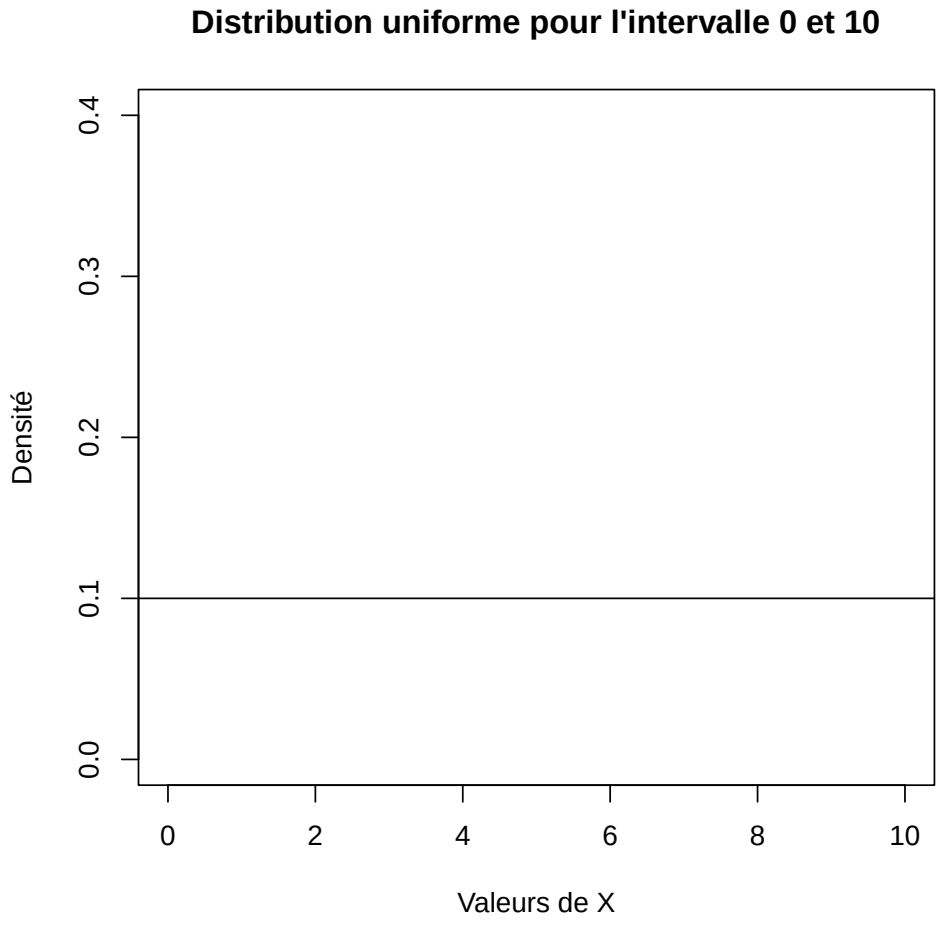

FIGURE 1 – Distribution uniforme pour l'intervalle 0 et 10.

même distribution :

$$f(8|0,10) = \frac{1}{(\max - \min)} = \frac{1}{(10-0)} = 0.1$$
$$f(6.15|0,10) = \frac{1}{(\max - \min)} = \frac{1}{(10-0)} = 0.1$$
$$f(10|0,10) = \frac{1}{(\max - \min)} = \frac{1}{(10-0)} = 0.1$$

On remarque que, peu importe la valeur, la densité de probabilité est de 0.1 tant et aussi longtemps que la valeur est dans l'intervalle 0 et 10. Ainsi, dès que les valeurs sont à l'extérieur de l'intervalle, la densité est de 0 :

$$f(100.2|0,10) = 0$$
$$f(-1.4|0,10) = 0$$
$$f(10.1|0,10) = 0$$

Puisqu'elle assigne à tous les éléments une probabilité égale d'être sélectionnés, la distribution uniforme est souvent utilisée pour exécuter des sélections aléatoires. Si nous avons une population de 400 éléments (par exemple des arbres), et que nous voulons choisir un échantillon de 40 unités, nous pourrions procéder en assignant un numéro à chaque élément et en utilisant la distribution uniforme pour sélectionner aléatoirement les observations :

```
> ##on "numérote" 400 éléments
> population <- 1:400
> ##on utilise la distribution uniforme
> round(runif(n = 60, min = 0, max = 400), digits = 0)
 [1] 332 265 301 309 361 271 180 364 154 316 227 157 252 397
[15] 388 274 224 388 229 232 347  90 163  37 295 221 385 374
```

```
[29]  88  51 332  90 157 266 204  54 213 365 296 197 229 170
[43] 115 161 280 116 101  85 215  58 246 371 309 128 297 180
[57] 343 281 122  49
```

On note ici que nous avons demandé 60 valeurs aléatoires d'une distribution uniforme continue alors que nous en voulons 40, au cas où il y aurait des doublons dans les 40 premières valeurs. De plus, l'utilisation de la fonction `round( )` permet d'arrondir les valeurs à l'entier le plus près.

Encore plus simple, nous pourrions utiliser la fonction `sample( )` qui permet de sélectionner aléatoirement avec ou sans remise. La sélection sans remise consiste à retirer les observations qui ont déja été tirées afin de ne pouvoir les sélectionner à nouveau. La sélection avec remise implique de pouvoir sélectionner la même observation à plusieurs reprises pour constituer l'échantillon. Autrement dit, la même observation peut apparaître plusieurs fois dans l'échantillon.

```
> ##on crée une série de 10 observations
> serie10 <- 1:10
> serie10
 [1]  1  2  3  4  5  6  7  8  9 10
> ##on sélectionne aléatoirement sans remise
> sample(x = serie10, size = 10, replace = FALSE)
 [1]  8  3  1  5  2  6 10  4  7  9

> ##on sélectionne aléatoirement avec remise
> sample(x = serie10, size = 10, replace = TRUE)
 [1]  5 10  7  4 10  8  8  4 10  7
```

En observant le code attentivement, on remarque que la sélection avec remise (`replace = TRUE`) peut donner plusieurs fois la même valeur. Typiquement, pour sélectionner des observations et construire notre échantillon, nous utiliserons la sélection sans remise.

À noter qu'on peut également utiliser la **randomisation** afin d'attribuer des **traitements** particuliers à certaines unités ou encore randomiser la séquence des observations à effectuer [1]. Par randomisation, on entend l'application d'un processus aléatoire sur l'attribution d'une condition aux unités d'échantillonnage. Par exemple, dans une expérience étudiant l'effet d'un nouveau médicament sur des patients, on pourrait sélectionner aléatoirement 50 patients afin de constituer deux groupes : un groupe de patients qui prendront le médicament et un autre groupe de patients qui ne le prendront pas. Ici, on pourrait randomiser le traitement, c'est-à-dire, déterminer aléatoirement le traitement que recevra chaque patient.

En contrepartie, la sélection avec remise est plutôt utilisée dans le contexte du **bootstrap**, qui est une approche permettant d'obtenir des erreurs-types d'une statistique d'intérêt ou de construire des intervalles de confiance.

3 Intervalle de confiance

Le concept **d'intervalle de confiance** (*IC, confidence interval*) est une des notions les moins bien comprises dans les cours de statistiques. Proprement dit, c'est l'intervalle à l'intérieur duquel se trouvera le paramètre de la population (p. ex., μ) si l'on répète l'échantillonnage avec la même taille d'échantillon un grand nombre de fois. Avant de construire un intervalle de confiance, il faut choisir une probabilité d'erreur ou seuil (α). Par convention, α prend la valeur 0.01, 0.05 ou 0.1. Le niveau de confiance (coefficient de confiance) est $1 - \alpha$. Par conséquent, si on utilise un $\alpha = 0.05$, le niveau de confiance est 1 - 0.05 = 0.95 = 95 %. L'intervalle est un intervalle de confiance à $100(1 - \alpha)\% = 95\ \%$.

À partir du théorème de la limite centrale vu à la leçon précédente (*Statistiques descriptives*), on sait que la distribution des moyennes des échantillons approxime une distribution

1. Plusieurs fonctions de `R` utilisent des générateurs de nombres aléatoires dans `R`, notamment `rnorm( )` et `sample( )`. Les algorithmes aléatoires nécessitent une valeur initiale appelée *seed* afin de démarrer le processus aléatoire. Par défaut, ces valeurs proviennent de l'horloge interne de l'ordinateur (année, mois, jour, heure, minute, seconde, microseconde). Ainsi, si on effectue la fonction `rnorm( )` ou `sample(, replace = FALSE)` plusieurs fois, on obtiendra un résultat différent à chaque fois. On peut aussi procurer une valeur initiale (*seed*) afin d'obtenir le même résultat à chaque fois à l'aide de la fonction `set.seed( )`. L'argument `seed` de cette fonction prend un entier comme valeur.

normale. Nous connaissons aussi certaines propriétées de la distribution normale centrée réduite, notamment, que 95 % des observations se trouvent à 1.96 σ de la moyenne. Ainsi, on peut construire un intervalle de confiance autour de μ :

$$\text{Si } n \geq 50,$$
$$\bar{x} \pm z_{\alpha/2} \cdot SE$$

où la première ligne signifie que cet intervalle est seulement valide pour des échantillons d'au moins 50 observations (voir l'explication plus bas). Dans la deuxième ligne, SE correspond à l'erreur-type de la moyenne et $z_{\alpha/2}$ correspond à l'écart normal qui définit les deux bornes (négative et positive) d'un intervalle qui exclue une proportion $\alpha/2$ des valeurs à gauche et une proportion $\alpha/2$ des valeurs à droite de la distribution normale centrée réduite, c'est-à-dire qu'inversement, cet intervalle inclue $100(1-\alpha)\%$ des valeurs de cette distribution. Nous divisons donc par 2 le seuil d'erreur α puisque l'intervalle de confiance a une borne à gauche et une autre à droite. Donc, si $\alpha = 0.05$, on a $z_{\alpha/2} = z_{0.025} = 1.96$, et il y a 2.5 % de chance d'avoir une valeur à gauche de la borne $1.96 \cdot SE$ (aire sous la courbe de 0.025) et 2.5 % de chance d'avoir une valeur à droite de la borne $1.96 \cdot SE$ (aire sous la courbe de 0.025). Nous pouvons écrire :

IC à 95 % :

$$P(\bar{x} - 1.96 \cdot SE \leq \mu \leq \bar{x} + 1.96 \cdot SE) = 0.95$$

IC à 90 % :

$$P(\bar{x} - 1.64 \cdot SE \leq \mu \leq \bar{x} + 1.64 \cdot SE) = 0.90$$

On peut transposer les notions de la distribution normale dans `R`. La fonction `dnorm( )` correspond à la fonction de densité de probabilité de la distribution normale pour une moyenne et un écart-type donnés. Ainsi, pour obtenir la densité de probabilité associée à $X = 3.5$

dans une distribution normale avec une moyenne (μ) de 4.2 et un écart-type (σ) de 10, nous procédons ainsi :

```
> dnorm(x = 3.5, mean = 4.2, sd = 10)
[1] 0.03979661
```

La fonction `qnorm( )` permet de trouver le quantile associé à une probabilité cumulative d'une distribution normale donnée pour un μ et σ donnés. Pour déterminer l'écart-normal $z_{0.05}$, on peut faire :

```
> ##z pour 0.05
> qnorm(p = 0.05, mean = 0, sd = 1)
[1] -1.644854
> ##la distribution est symétrique et on
> ##peut obtenir le quantile du côté droit
> qnorm(p = 0.95, mean = 0, sd = 1)
[1] 1.644854
> ##z pour 0.025
> qnorm(p = 0.025, mean = 0, sd = 1)
[1] -1.959964
```

La probabilité cumulative $P(X < x)$ associée à un quantile, μ et σ s'obtient avec `pnorm( )` :

```
> ##probabilité cumulative pour z = 0.025
> pnorm(q = -1.96, mean = 0, sd = 1)
[1] 0.0249979
> ##probabilité cumulative pour z = 0.05
> pnorm(q = -1.64, mean = 0, sd = 1)
```

```
[1] 0.05050258
```

On connaît rarement la SE réelle de la population ($\sigma_{\bar{x}}$) et on doit l'estimer à partir de l'échantillon. L'estimation de ce paramètre est bonne lorsque $n > 50$. Toutefois, l'estimation de la SE est moins bonne pour des échantillons avec moins de 50 observations. En réalité, avec de petits échantillons, l'intervalle de confiance basé sur le z est trop étroit – il indique une précision plus grande qu'elle ne l'est en réalité. On doit donc corriger notre intervalle de confiance en utilisant la distribution du t de Student.

3.1 Distribution du t de Student

La **distribution du t de Student** vient du statisticien et maître brasseur de Guinness, William Sealy Gosset. Au début du siècle dernier, la brasserie Guiness a investi beaucoup de ressources afin de sélectionner les variétés de houblon et d'orge qui possédaient les meilleures caractéristiques. Pour ce faire, ils embauchèrent des chimistes et des statisticiens pour arriver à leurs fins en leur donnant le statut de brasseurs et en mettant à leur disposition des laboratoires et des champs pour effectuer des expériences. Étant lié à Guinness et pour ne pas divulguer des secrets de production, Gosset publia sa découverte de la distribution sous le nom de *Student*. Cette distribution ne possède qu'un paramètre, soit les degrés de liberté $(n - 1)$.

Avec `R`, nous pouvons déterminer la densité de probabilité, le quantile et la probabilité cumulative pour un degré de liberté donné avec les fonctions `dt( )`, `qt( )` et `pt( )`, respectivement. Bien que de forme très semblable à la distribution normale, la distribution du t de Student a un plus grand nombre d'observations dans les extrémités que la distribution normale (fig. 2). Pour un n et une probabilité cumulative donnés, les quantiles de la distribution du t de Student sont plus grands que celui de la distribution normale centrée-réduite (fig. 3). Par conséquent, l'intervalle de confiance obtenu avec le t est plus large qu'avec z lorsque $n < 50$.

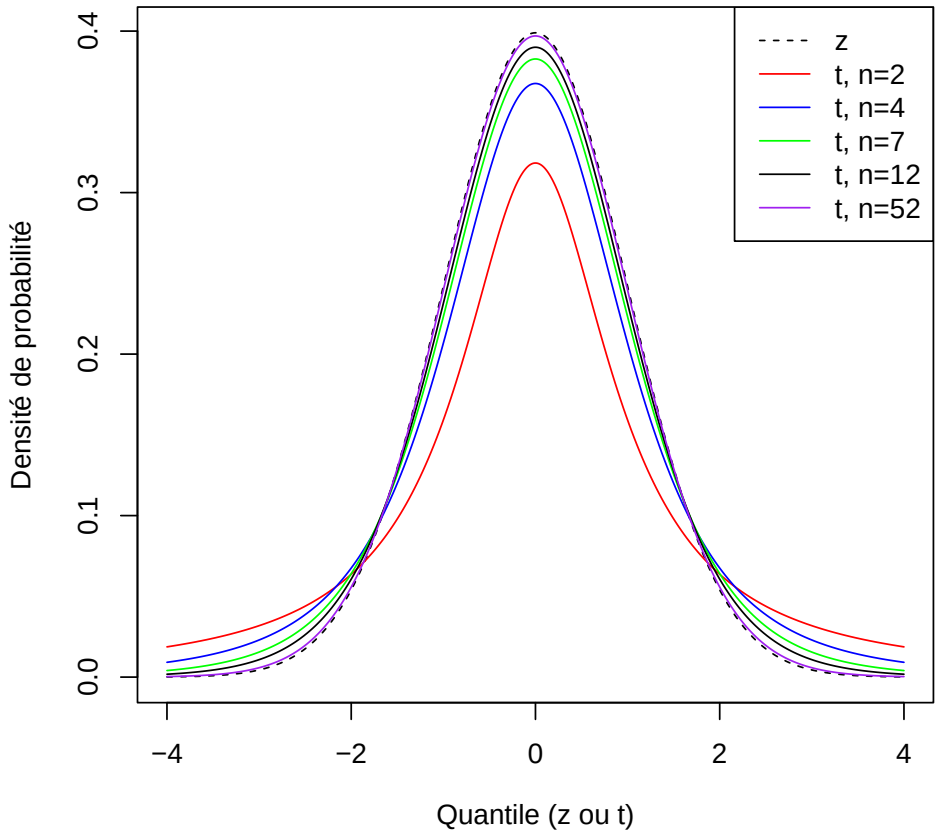

FIGURE 2 – Comparaison de la distribution normale centrée réduite et de la distribution du t de Student. On note qu'il y a un plus grand nombre de valeurs dans les extrémités de la distribution du t que dans la distribution normale centrée réduite dès que $n < 50$.

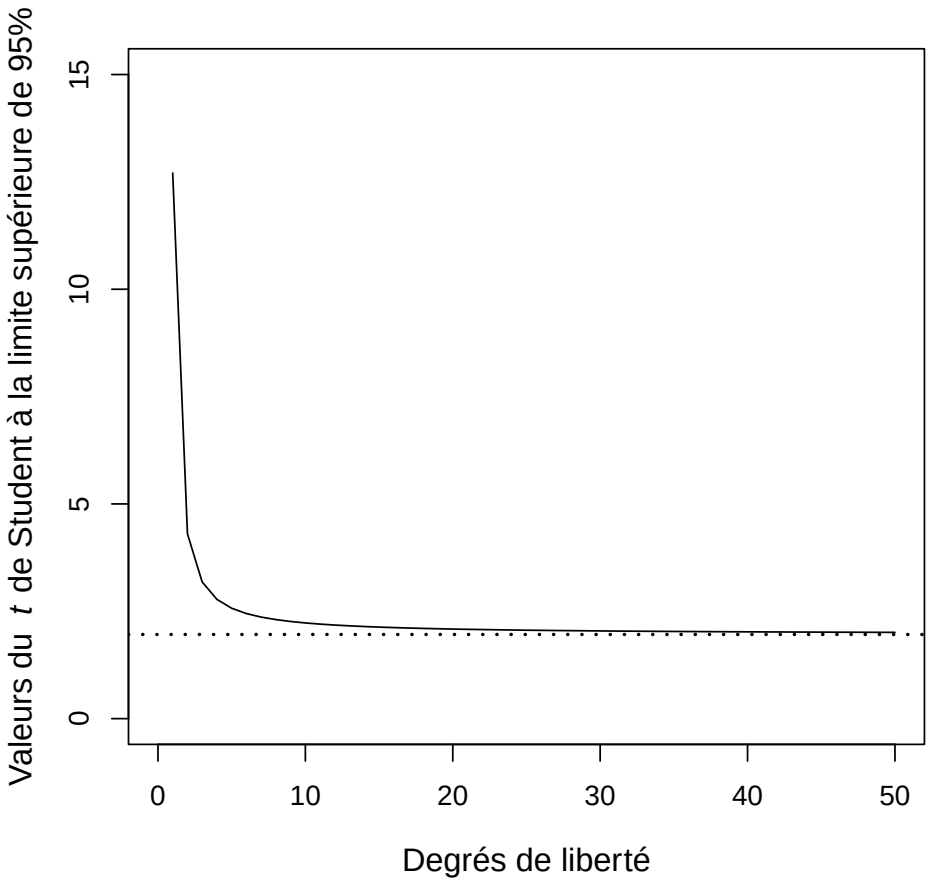

FIGURE 3 – Valeurs des quantiles de la distribution du t de Student en fonction de la taille d'échantillon ($df = n - 1$) pour une probabilité cumulative de 0.95. La ligne horizontale pointillée correspond à l'écart-normal de la distribution normale centrée réduite pour la même probabilité cumulative.

De façon plus générale, on peut construire un intervalle à $(1-\alpha)$:

$$P(\bar{x} - t_{\alpha/2,(n-1)} \cdot SE \leq \mu \leq \bar{x} + t_{\alpha/2,(n-1)} \cdot SE) = (1-\alpha)$$

Pour un intervalle de confiance à 95 %, $(1-\alpha) = 0.95$,

$$P(\bar{x} - t_{0.05/2,(n-1)} \cdot SE \leq \mu \leq \bar{x} + t_{0.05/2,(n-1)} \cdot SE) = (1-\alpha)\,.$$

Exemple 2.3 Une inspectrice municipale mesure l'épaisseur de l'asphalte à 32 endroits où une compagnie privée a fait la chaussée dans Montréal. L'épaisseur moyenne de l'asphalte est de 20.1 cm et l'écart-type estimé à partir de l'échantillon est de 3.39 cm. Quel est l'intervalle de confiance à 95 % autour de la moyenne ? On peut obtenir les quantiles nécessaires à la solution du problème comme suit :

```
> ##détermination du quantile du t à alpha/2
> t.stat <- qt(p = 0.025, df = 31)
> t.stat
[1] -2.039513
> ##convertir le quantile en valeur positive
> ##pour procéder au calcul des bornes
> t.stat <- t.stat *(-1)
> t.stat
[1] 2.039513
> ##calcul de la SE
> SE <- 3.39/sqrt(32)
> SE
```

```
[1] 0.599273
> ##calcul de la borne inférieure
> 20.1 - t.stat * SE
[1] 18.87777
> ##calcul de la borne supérieure
> 20.1 + t.stat * SE
[1] 21.32223
```

Nous pouvons donc écrire :

$$\bar{x} = 20.1\,\text{cm}$$

$$s = 3.39\,\text{cm}$$

$$n = 32$$

$$SE = s_{\bar{x}} = \frac{3.39}{\sqrt{32}} = 0.60\,\text{cm}$$

IC à 95 % :

$$P(\bar{x} - t_{0.05/2,(n-1)} \cdot SE \leq \mu \leq \bar{x} + t_{0.05/2,(n-1)} \cdot SE) = 0.95$$

$$P(\bar{x} - t_{0.025,(32-1)} \cdot 0.60 \leq \mu \leq \bar{x} + t_{0.025,(32-1)} \cdot 0.60) = 0.95$$

$$P(18.88 \leq \mu \leq 21.32) = 0.95$$

IC à 95% : $(18.88, 21.32)$.

On peut comparer les intervalles de confiance à 90 % et 99 % à partir du même

échantillon :

IC à 90% :

$$P(\bar{x} - t_{0.10/2,(32-1)} \cdot SE \leq \mu \leq \bar{x} + t_{0.10/2,(32-1)} \cdot SE) = 0.90$$

IC à 99% :

$$P(\bar{x} - t_{0.01/2,(32-1)} \cdot SE \leq \mu \leq \bar{x} + t_{0.01/2,(32-1)} \cdot SE) = 0.99$$

IC à 90% : $(19.08, 21.12)$

IC à 95% : $(18.88, 21.32)$

IC à 99% : $(18.45, 21.75)$.

On remarque rapidement que la largeur de l'intervalle de confiance augmente avec le seuil de confiance. Dans notre exemple, il est le plus étroit à 90 % et le plus large à 99 %. Comme nous l'avons mentionné plus haut, la moyenne de population (μ) sera incluse dans l'intervalle de confiance $(1-\alpha)\%$ du temps lorsqu'on répète l'expérience un grand nombre de fois. L'intervalle de confiance implique le concept de rééchantillonnage (*resampling*), c'est-à-dire, la construction de plusieurs échantillons à partir de la même population, et c'est ce que nous verrons dans la prochaine section.

3.2 Rééchantillonnage

Les statistiques classiques impliquent souvent un échantillonnage répété dans leur interprétation. Tout comme l'intervalle de confiance, la probabilité d'un test statistique représente ce que l'on devrait observer en moyenne lorsqu'il n'y a pas d'effet dans la population. Ici, « en moyenne » indique ce qu'on obtiendrait si on répétait l'expérience un grand nombre de fois avec une série d'échantillons de taille égale à celle de notre échantillon original, tous provenant de la même population. L'exemple suivant illustre ce principe à l'aide d'une population

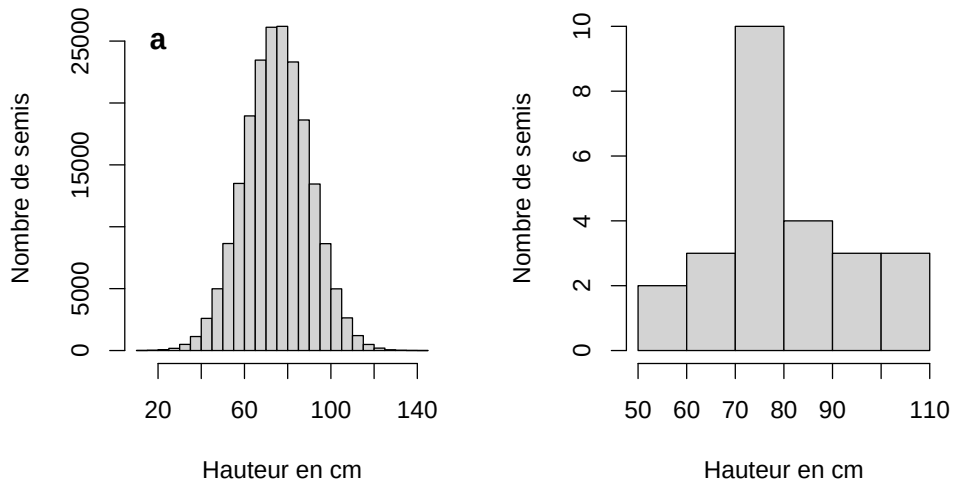

FIGURE 4 – Distribution de la population complète de semis dans une plantation où $\mu = 74.99$ et $\sigma = 14.99$ (a) et d'un échantillon aléatoire de 25 observations tiré de cette population (b).

de 200 000 semis dans une plantation.

Exemple 2.4 Imaginons qu'une compagnie d'exploitation forestière coupe la forêt sur un site de 20 hectares (ha). La compagnie a replanté le site à raison d'un semis d'épinette noire par m^2, ce qui équivaut à un total de 200 000 semis. Deux ans après avoir planté les semis, on veut connaître la croissance moyenne des semis sur le site. Le total de 200 000 semis du site de 20 ha constitue la population d'intérêt.

Chacun des semis de la population croît à sa propre vitesse (fig. 4a), mais on ne peut connaître sa croissance sans la mesurer. On embauche un technicien pour sélectionner aléatoirement 25 semis à mesurer afin d'estimer la hauteur moyenne des semis dans cette même population (fig. 4b). On peut calculer l'intervalle de

confiance à 95 % à partir de notre échantillon, ce qui donne :

$$\bar{x} = 78.45 \text{ cm}$$

$$s = 14.4 \text{ cm}$$

$$n = 25$$

$$SE = \frac{14.4}{\sqrt{25}} = 2.88 \text{ cm}$$

IC à 95 % :

$$P(\bar{x} - t_{0.05/2,(25-1)} \cdot SE \leq \mu \leq \bar{x} + t_{0.05/2,(25-1)} \cdot SE) = 0.95$$

$$P(78.45 - 2.06 \cdot 2.88) \leq \mu \leq 78.45 + 2.06 \cdot 2.88) = 0.95$$

IC à 95% : $(72.5, 84.39)$

L'intervalle de confiance obtenu à partir de l'échantillon de 25 observations est $(72.5, 84.39)$. À noter que `qt( )` permet d'obtenir le quantile du t associé à la probabilité cumulative correspondant à $\alpha/2$. On constate que la vraie moyenne de la population ($\mu = 74.99$) est contenue dans l'intervalle que l'on a construit. En d'autres mots, l'intervalle de confiance inclut ou n'inclut pas la vraie moyenne.

On pourrait demander à 19 autres techniciens de sélectionner aléatoirement 25 semis, de les mesurer, puis de construire un intervalle de confiance à 95 % à partir de leur échantillon. On devrait s'attendre à ce que 19 des 20 intervalles de confiance ($19/20 = 0.95$) ainsi construits incluent réellement la moyenne de la population (μ).

On peut aussi utiliser le concept de rééchantillonnage pour illustrer l'erreur-type de la moyenne. Dans la leçon 1, nous avons indiqué que la moyenne d'un échantillon provenait

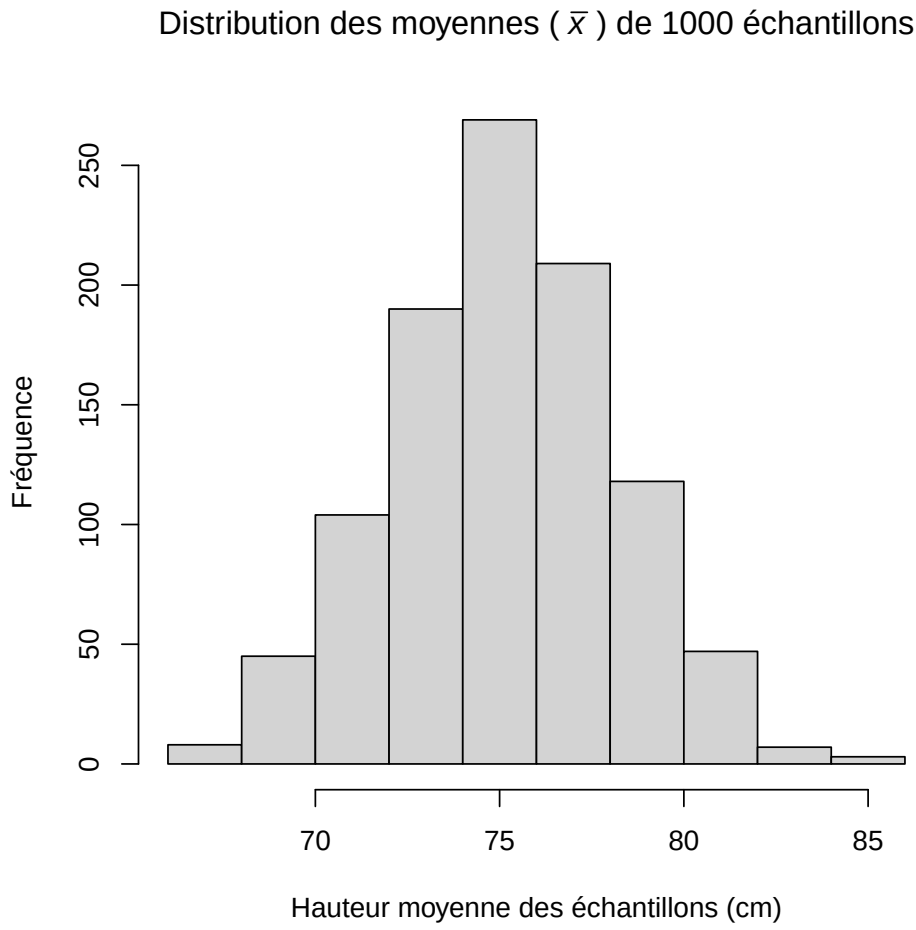

FIGURE 5 – Distribution des 1000 moyennes d'échantillons de 25 semis tirés de la population de 200 000 semis.

d'une distribution de moyennes d'échantillons. On pourrait demander à 1000 techniciens d'obtenir un échantillon aléatoire de 25 semis et de mesurer la hauteur de chaque arbre. Si chaque technicien calcule la moyenne de son échantillon, nous obtiendrons une distribution de moyennes d'échantillons (fig. 5). Rappelons-nous que l'erreur-type de la moyenne est une estimation de la variation entre les moyennes d'échantillons de même taille provenant de la même population. L'écart-type calculé à partir des valeurs des moyennes nous donnera l'erreur-type de la moyenne.

4 Autres stratégies d'échantillonnage

4.1 Échantillonnage stratifié

En plus de l'échantillonnage complètement aléatoire, d'autres approches permettent également de construire un échantillon à partir d'une population. **L'échantillonnage stratifié** (*stratified sampling*) s'applique lorsque la population d'intérêt est divisée en régions ou **strates**. On entend par « strate » une tranche ou une partie de la population. Par exemple, si on s'intéresse à la masse des ours noirs dans l'est du Québec, on pourrait échantillonner des ours dans différents types d'habitats de cette région : urbain, aire de conservation, aire de coupe forestière. Si la masse des ours est sujette à différer d'un type d'habitat à l'autre, il serait inapproprié de procéder à un échantillonnage complètement aléatoire. Effectivement, certains types d'habitats risqueraient d'être sous-représentés et on supposerait erronément que le type d'habitat n'est pas une source de variation de la masse d'ours noir.

L'échantillonnage stratifié consiste à partitionner la population afin que les unités à l'intérieur d'une même strate soient le plus similaires possible. Même si les strates diffèrent les unes des autres, l'échantillon stratifié sera représentatif de la population. On peut stratifier une région par une variable connue (habitat, élévation, type de sol) ou par sous-unité géographique. On peut également stratifier l'échantillonnage par espèce, sexe ou classe d'âge. La sélection des unités d'échantillonnage dans chacune des strates doit être indépendante de la sélection dans les autres strates. L'échantillonnage aléatoire est encore important ici. Lorsque chaque unité est sélectionnée aléatoirement à l'intérieur de chaque strate, on parle d'échantillonnage stratifié aléatoire. L'estimation de la moyenne sera la plus précise si les unités sont le plus semblables possible à l'intérieur de chaque strate.

L'échantillonnage coûte cher en ressources (temps, argent, personnel). C'est pourquoi il est préférable de penser à la stratégie d'échantillonnage avant de débuter la collecte des données. Considérez la notation suivante :

- L : nombre de strates ;

- n : taille de l'échantillon ;
- n_h : nombre d'échantillons dans la strate h ;
- N_h : taille de la strate h ;
- N : taille de population.

Afin d'optimiser l'effort et les coûts, on peut répartir l'effort d'échantillonnage en fonction de la taille des strates.

- Si les L strates sont de taille égale, on peut attribuer n_h égal partout ($n_h = n/L$). Par exemple, si $n = 30$ et $L = 3$, nous aurons donc $n_h = 30/3 = 10$.
- Si les strates sont de taille inégale, on peut attribuer un effort proportionnel à la taille de chaque strate (N_h) par rapport à la somme de l'ensemble des strates (N) afin d'obtenir une fraction constante partout qui équivaut à (n/N). Si la strate h contient N_h unités, l'effort sera $n_h = \frac{n \cdot N_h}{N}$. Par exemple, si $n = 30, N_h = 200$ et $N = 1000$, nous obtiendrons $n_h = \frac{n \cdot N_h}{N} = \frac{30 \cdot 200}{1000} = 6$.
- On peut aussi opter de répartir l'effort d'échantillonnage pour obtenir la moyenne avec la plus faible variance pour une taille d'échantillon n donnée – c'est ce qu'on appelle l'allocation optimale. Cette dernière option requiert une estimation de la variance à partir de données préléminaires ou d'anciens échantillons. Elle permet également d'y incorporer le coût en ressources (e.g., argent, temps). On attribue de plus gros échantillons aux plus grandes strates ou aux strates les plus variables et des plus petits échantillons aux strates plus petites ou qui sont plus coûteuses à échantillonner. Ceci permet d'optimiser l'allocation des ressources dévouées à l'échantillonnage.

4.2 Échantillonnage par grappes et échantillonnage systématique

Dans **l'échantillonnage par grappes**, la population est divisée en groupes (grappes) d'unités secondaires près les unes des autres et les grappes sont sélectionnées aléatoirement. Par grappes, on entend un groupe composés d'unités spatiales adjacentes, comme des qua-

drats ou des transects contigus. La figure (6) montre cinq grappes de quatre quadrat. Ces grappes sont sélectionnées aléatoirement dans l'aire d'étude, mais les unités qui composent la grappe ne le sont pas. Ce genre de dispositif ne peut pas être analysé comme si les données étaient complètement indépendantes les unes des autres. Par conséquent, des analyses moins conventionnelles sont nécessaires pour estimer les paramètres d'intérêt.

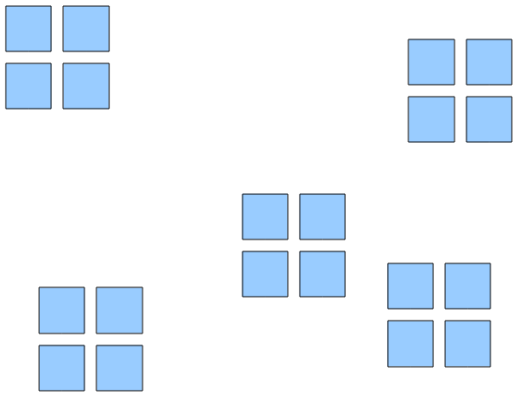

FIGURE 6 – Exemple de grappes de quadrats pour l'échantillonnage par grappes.

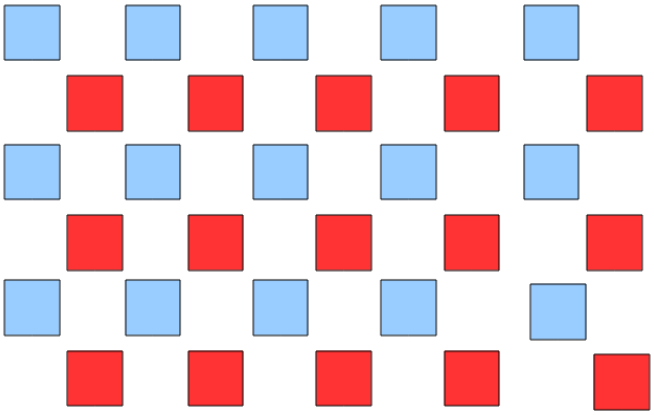

FIGURE 7 – Exemple d'échantillonnage systématique avec deux grappes (grilles) de quadrats.

L'échantillonnage systématique est une variation de l'échantillonnage par grappe où on dispose systématiquement les unités d'échantillonnage parmi la population. Le point de départ de la grille est sélectionné aléatoirement, mais les unités d'échantillonnage ne le sont

pas (fig. 7).

Avec l'échantillonnage par grappes ou l'échantillonnage systématique, on peut obtenir des estimés avec une faible variance si les grappes sont semblables entre elles. Idéalement, toute la variabilité de la population devrait être observée dans chaque grappe afin d'obtenir un échantillon représentatif de la population. Ce dernier point contraste avec l'échantillonnage stratifié qui minimisait la variance à l'intérieur des strates. Il y a toutefois quelques mises en garde :

- l'échantillonnage systématique avec une seule grappe permet d'estimer une moyenne, mais la variance ne se calcule pas comme pour un échantillon aléatoire ;
- la forme et la taille des grappes influencent l'efficacité ;
- en présence de forte hétérogénéité spatiale caractérisée par des patrons périodiques, l'échantillonnage systématique est déconseillé. Par exemple, échantillonner dans un milieu vallonné, où les unités d'échantillonnage tombent toujours dans une vallée sur toute l'aire d'étude ne capture que les caractéristiques des observations sur l'ensemble du milieu vallonné.

4.3 Échantillonnage multistade

Il existe d'autres stratégies d'échantillonnage, telles que **l'échantillonnage multistade** (*multistage sampling*) qui consiste à sélectionner aléatoirement des unités secondaires à l'intérieur d'unités primaires. Pour illustrer cette stratégie, considérons une étude qui vise à caractériser des planches de bois destinées à la construction. Pour ce faire, on pourrait sélectionner 10 camions parmi ceux qui arrivent à l'entrepôt d'un centre de rénovation et choisir aléatoirement 15 planches dans chaque camion. Cette stratégie se distingue de l'échantillonnage par grappes par une sélection aléatoire des unités secondaires dans les grappes.

Tout comme l'échantillonnage par grappes, l'échantillonnage multistade est logistiquement plus facile à réaliser que l'échantillonnage complètement aléatoire. Toutefois, les calculs des moyennes et des variances sont plus complexes que pour l'échantillonnage complètement

aléatoire. Il est possible d'ajouter d'autres niveaux d'échantillonnage. Par exemple, dans une étude sur les insectes en forêt tropicale, on pourrait sélectionner aléatoirement des parcelles de 20 x 20 m dans l'aire d'étude dans lesquelles on sélectionnerait 5 quadrats de 1 x 1 m, pour ensuite établir 1 microquadrat de 0.25 x 0.25 m dans chaque quadrat afin de compter le nombre d'espèces de mousses.

4.4 Échantillonnage adaptatif

L'échantillonnage adaptatif (*adaptive sampling*) est une classe de stratégies d'échantillonnage qui regroupe plusieurs méthodes où la sélection des unités est modifée au cours de l'échantillonnage en fonction des observations qui s'accumulent. On peut l'utiliser pour des espèces rares ou difficiles à observer où les individus sont regroupés spatialement (bancs de poissons, colonie d'insectes), ou encore pour échantillonner des dépôts géologiques (gisements d'or dont on ne connait pas initialement la localisation exacte). Toutefois, les méthodes de calcul des moyennes et de variance sont particulièrement complexes avec cette méthode d'échantillonnage, ce qui fait qu'elle est encore peu utilisée. En effet, peu de techniques d'analyses dans le moment permettent de traiter des données provenant de cette stratégie d'échantillonnage.

4.5 Estimation de la probabilité de détection

Certaines stratégies d'échantillonnage permettent d'estimer la probabilité de détection d'organismes ou d'éléments étudiés qui sont difficiles à détecter lorsque présents sur un site lors d'un inventaire. Les facteurs tels que la coloration, le comportement, le type d'habitat, la méthode d'échantillonnage, les conditions météorologiques, le temps de la journée, l'abondance et la période de l'année peuvent tous influencer la détection. Dans le meilleur des cas, ce genre de problème amène des sous-estimation de l'abondance ou des patrons d'occurrence. Dans le pire des cas, les problèmes de détection entraînent des conclusions erronées lorsqu'on compare différents traitements entre eux. Ces approches sont particulièrement appliquées en

écologie animale et constituent une branche importante des statistiques. La littérature à ce sujet croît à une vitesse fulgurante.

Parmi les stratégies qui estiment la probabilité de détection, on en compte trois principales, notamment l'échantillonnage de la distance, les méthodes de capture-marquage-recapture et les analyses d'occupation de sites.

4.5.1 Échantillonnage de la distance

Pour estimer l'abondance, on peut utiliser **l'échantillonnage de la distance** (*distance sampling*), lequel consiste à établir des parcelles d'échantillonnage circulaires [2] ou des transects [3] disposés aléatoirement sur l'aire d'étude et dans lesquels on mesure la distance entre l'observateur et les individus détectés. Ceci permet de construire une fonction de détection et d'estimer l'abondance.

4.5.2 Méthodes de capture-marquage-recapture

Les méthodes de **capture-marquage-recapture** (CMR) consistent à capturer, marquer et recapturer des individus sur un ou plusieurs sites pendant plusieurs visites successives. On peut ainsi estimer la probabilité de capture et l'abondance des individus sur le site, laquelle est corrigée pour la détection imparfaite. Les méthodes de CMR sont également utilisées pour estimer des paramètres vitaux (*vital rates*), tels que la probabilité de survie, d'immigration et de transition d'un état à un autre, par exemple d'un état non reproducteur à reproducteur.

4.5.3 Analyses d'occupation de sites

Les **analyses d'occupation de sites** (*site occupancy analyses*) constituent une famille de méthodes généralement utilisée dans le domaine de la conservation de la biodiversité et permettant d'estimer la détection d'une espèce animale ou végétale sur une série de sites

2. Une parcelle d'échantillonnage est tout simplement un cercle dans lequel on échantillonne les éléments d'intérêt sur le terrain comme des oiseaux ou des grenouilles.

3. Un transect est un corridor de largeur et longueur définie qui est disposé dans un site afin d'échantillonner les éléments d'intérêt comme des herbivores.

au cours de plusieurs visites successives, et ainsi d'obtenir une meilleure estimation de la probabilité de présence de l'espèce d'intérêt. Des modifications du modèle de base permettent d'estimer les probabilités d'extinction et de colonisation ou encore de co-occurrence entre deux espèces.

Parmi toutes les stratégies d'échantillonnage présentées, nous allons utiliser l'échantillonnage complètement aléatoire et l'échantillonnage stratifié aléatoire. À noter que les exemples vus dans le cours supposent que la détection des individus est parfaite.

5 Conclusion

Dans ce texte, nous avons introduit le concept d'intervalle de confiance en nous appuyant sur le rééchantillonnage. Selon la taille d'échantillon, nous avons utilisé la distribution normale ou la distribution du t de Student pour construire les intervalles de confiance. Parmi les différentes stratégies d'échantillonnage présentées, nous verrons le plus souvent l'échantillonnage complètement aléatoire et l'échantillonnage stratifié aléatoire.

Index