Initiation à la théorie de l'échantillonnage

Jean VAILLANT

Octobre 2005

1 Notions de base en échantillonnage

L'étude de propriétés caractéristiques d'un ensemble, quand on ne dispose pas encore de données, nécessite d'examiner, d'observer des éléments de cet ensemble. La manière de recueillir ces données fait l'objet d'une théorie mathématique appelée *théorie des sondages* ou encore *théorie de l'échantillonnage*, en anglais *sampling theory*. Cette théorie concerne l'optimisation de la collecte des données selon divers critères et répond à certaines interrogations sur la façon de procéder à cette collecte en rapport avec l'information disponible et l'effort d'échantillonnage consenti.

1.1 Population, individu statistiques

La **population statistique** est l'ensemble sur lequel des méthodes et techniques de présentation, de description et d'inférence statistique sont appliquées. Il ne s'agit donc pas forcément d'une population au sens biologique du terme.

Les **individus statistiques** ou **unités statistiques** sont les éléments de la population statistique.

Les exemples sont innombrables :

(1) On désire étudier la préférence pour tel ou tel candidat dans une circonscription. La population statistique est l'ensemble des électeurs de la circonscription.

(2) On s'intéresse à l'action d'un parasite sur les pontes de la pyrale de la canne à sucre dans une région. La population statistique est l'ensemble des plantes des parcelles cultivées en canne à sucre de la région étudiée.

(3) On s'intéresse à la répartition d'une maladie sur les arbres d'une forêt. La population statistique est l'ensemble des arbres de cette forêt.

(4) On désire évaluer le budget mensuel moyen des étudiants d'une université. La population statistique est l'ensemble des étudiants de cette université.

1.2 Sondage, échantillonnage

On appelle **sondage** toute observation partielle d'une population statistique c'est-à-dire l'observation d'une partie de cette population. On cherche généralement à extrapoler les résultats observés à la totalité de la population. Une **unité de sondage** (ou **unité d'échantillonnage**) est un regroupement d'unités statistiques.

Une **méthode de sondages** (ou **d'échantillonnage**) décrit la façon dont la population statistique sera observée partiellement à travers un de ses sous-ensembles appelé **échantillon**.

Plan de sondages, plan d'échantillonnage, procédure d'échantillonnage ont des définitions équivalentes à celles de méthode de sondages.

Il est important de ne pas confondre **sondage** et **sondage d'opinion**. Les sondages d'opinion vise à obtenir des informations sur l'état d'esprit d'une population humaine. Il s'agit donc d'une forme particulière de sondage: les individus statistiques sont des personnes interrogées à travers un questionnaire sur leur opinion. Parmi les exemples ci-dessus, seul [1] est un sondage d'opinion ([4] n'en est pas un, bien que l'on y interroge des personnes). Les sondages d'opinion sont très médiatisés, particulièrement pendant les périodes préélectorales.

La **théorie des sondages** est un ensemble d'outils statistiques permettant l'étude d'une population statistique à partir de l'examen d'un échantillon tiré de celle-ci (Tillé, 2001). On parle aussi, de façon équivalente, de la **théorie de l'échantillonnage**. Cette dernière expression est davantage utilisée en sciences agronomiques ou biologiques.

1.3 Population finie, taux de sondage

La **taille de la population statistique** est l'effectif de cette population c'est-à-dire le nombre d'individus statistiques dont elle est constituée. Une **population finie** est une population dont la taille est finie. Une **population infinie** est une population dont la taille est infinie. Dans la pratique, on peut considérer une population finie comme étant infinie si elle est d'effectif très grand.

La **taille de l'échantillon** est l'effectif de cet échantillon c'est-à-dire le nombre d'individus statistiques observés dans la population statistique.

Le **taux de sondage** (ou **d'échantillonnage**), dans le cas de population finie, est le rapport

$$\frac{\text{taille d'échantillon}}{\text{taille de population}}.$$

Le **facteur correctif de population finie** est

$$1 - \frac{\text{taille d'échantillon}}{\text{taille de population}}.$$

Quand la population statistique est observée complètement, c'est-à-dire que l'échantillon est la population statistique toute entière, on parle d'**échantillonnage exhaustif** ou de **recensement**. Le taux de sondage est alors de 100%.

Pour des raisons de coûts financiers ou techniques, il est, dans bien des cas, impossible de faire un recensement. L'utilisation de sondages est alors incontournable.

1.4 Population cible, base de sondage

Dans certaines études, on souligne la notion de **population cible** puis celle de **base de sondage**. On a les définitions suivantes :

- La **population cible** est l'ensemble pour lequel on veut recueillir des informations et sur lequel doivent porter les conclusions de l'étude. Elle peut être distincte de la population statistique, en particulier quand ses éléments ne peuvent être tous répertoriés ou sont soumis à des contraintes liées à l'étude menée. Dans la table 1, les exemples 2, 4 et 5 correspondent à une situation où la population cible est différente de la population statistique.
- La **base de sondage** est déterminée, après avoir définie la population cible. Idéalement, c'est une liste de tous les individus de la population cible: liste électorale, liste des entreprises, liste d'étudiants, liste des arbres d'une parcelle sylvicole, liste des parcelles d'un domaine expérimental, etc... L'échantillon est alors extrait de cette liste à l'aide d'un algorithme de tirages des individus. Par contre, il arrive fréquemment qu'une telle base de sondage ne soit pas accessible directement. La **base de sondage** est par définition une liste d'individus statistiques identifiés permettant d'avoir accès à la majorité des individus de la population cible. Tous les individus de la population cible ne sont donc pas forcément inclus dans cette base.

En résumé, une base de sondage est donc une liste d'individus de la population statistique à partir de laquelle (liste) on tire l'échantillon avec, pour chaque individu, divers renseignements utiles à la réalisation de l'étude par échantillonnage. Un exemple de telle base est le registre des exploitants de tel département archivé par la Chambre d'Agriculture de ce département.

	Population statistique	Unité statistique	Population cible	Caractère étudié	Paramètre à estimer
1	Les arbres d'une forêt	Arbre de la forêt	Ensemble des arbres de cette forêt	Présence-absence d'une maladie	Proportion d'arbres malades dans la forêt
2	Les parcelles obtenues par quadrillage d'un champ de tournesol	Parcelle	Ensemble des plantes de tournesol du champ	Nombre de plantes crispées (action du puceron du tournesol)	Nombre moyen de plantes crispées dans le champ
3	Les étudiants d'une université	Etudiant	Ensemble des étudiants de cette université	Budget annuel	Budget annuel moyen par étudiant
4	Les entreprises répertoriées en début d'année dans une région	Entreprise de cette région	Ensemble des entreprises de cette région	Taux d'endettement	Taux d'endettement moyen des entreprises de cette région
5	Les plants de canne à sucre d'une région	Plant de canne à sucre	Ensemble des pontes de pyrale dans la région	Nombre de pontes parasitées par un oophage	Nombre moyen de pontes parasitées dans la région (efficacité de l'oophage)

Table 1: Exemples de population statistique, cible et caractère étudié.

Dans certains cas, on a plusieurs choix possibles pour la base de sondage. Ce choix dépendra des objectifs de l'étude, des données disponibles sur la base de sondage, de la qualité de la base de sondage et du budget de l'étude, comme le montrent les exemples suivants :

1) *Enquête auprès d'exploitants agricoles d'un département* : afin d'étudier l'utilisation d'intrants en agriculture en 2004 et analyser les risques pour l'environnement, on a considéré le registre des exploitants de 2003 archivé à la chambre d'Agriculture de ce département. En tenant compte des départs et arrivées enregistrés en 2004 ainsi que du secteur informel, on estime que cette base de sondage permet d'atteindre environ 89% des exploitants de ce département.

2) *Enquête concernant la dengue auprès des ménages d'une commune* : Le comportement vis-à-vis de la dengue et de son vecteur, le moustique Aedes aegypti, veut être étudié afin de mettre en place une campagne efficace d'éradication de ce moustique. L'individu statistique

est le logement. Il est assimilé au **ménage** qui est, par définition, l'ensemble des individus qui habitent le même logement. A la suite de chaque recensement, l'INSEE dispose d'une base de sondage comprenant tous les logements recensés. Cette base contient toutes les constructions achevées lors du recensement le plus récent. Elle est complétée par une base de sondage des logements neufs éditée par la Direction Départementale de l'Equipement. Ces bases de sondage représentent des fichiers énormes au niveau de la France. Pour une commune donnée, on peut extraire une base de sondage à partir de ces fichiers. Selon la commune concernée, le pourcentage des ménages couverts par une telle base de sondage varie à cause des éventuelles lenteurs de mise à jour des fichiers et des constructions illégales.

3) *Etude de la pression anthropique sur le crabe de terre en zone littorale* : Pour analyser l'impact anthropique sur la dynamique du crabe de terre, les différents terriers de la zone sont répertoriés. Ils constituent la base de sondage. La population cible est l'ensemble des crabes de terre.

L'existence d'une base adéquate de sélection des individus statistiques est donc un aspect important de la faisabilité de la plupart des plans d'échantillonnage.

1.5 Population fixe, superpopulation

Considérons une population notée $\mathcal{P}$ de taille N dont les individus sont répertoriés.

On peut donc affecter un numéro allant de 1 à N à chaque individu et assimiler notre population à l'ensemble des N premiers entiers naturels non nuls :

$$\mathcal{P} = \{1, \cdots, N\}.$$

Sur cette population statistique, on étudie un caractère $\mathcal{Y}$ prenant la valeur y_i sur l'individu i. **Avant l'exécution de l'échantillonnage, on a N valeurs inconnues** $y_1, y_2, \cdots, y_N$. Le fait d'échantillonner dans la population **permet d'accéder à certaines de ces valeurs**.

Le but de l'échantillonnage est souvent d'estimer une quantité $h(y_1, y_2, \cdots, y_N)$ dépendant donc par conséquence des $y_1, y_2, \cdots, y_N$. Cette quantité $h(y_1, y_2, \cdots, y_N)$ est appelé **fonction d'intérêt**. L'exemple le plus courant de fonction d'intérêt est **la moyenne de la population** :

$$\bar{y} = \frac{1}{N}(y_1 + y_2 + \cdots + y_N).$$

Remarque : la proportion est une forme particulière de moyenne : il s'agit de la moyenne d'un caractère ne pouvant prendre que les valeurs 0 ou 1 (**caractère binaire**).

On peut s'intéresser également à la fonction d'intérêt **variance de la population** :

$$V = \frac{1}{N}((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2).$$

Il est important de noter que, dans certaines situations, des suppositions de nature probabiliste sont faites sur les valeurs inconnues $y_1, y_2, \cdots, y_N$. Par exemple, on peut supposer que y_i est la réalisation d'une variable aléatoire suivant la loi gaussienne. On parle alors de **modèle de superpopulation**, par opposition au cas où aucune supposition distributionnelle n'est faite sur les $y_1, y_2, \cdots, y_N$ auquel cas on parle de **population fixe**. On parle donc de population fixe lorsque les y_i sont considérées fixes, ou en d'autres termes, quand on ne fait aucune supposition probabiliste sur les valeurs possibles du caractère $\mathcal{Y}$ sur les différents individus de la population.

1.6 Plan stochastique, plan empirique

Il existe deux grandes catégories de plans d'échantillonnage :

- **les plans probabilistes**, dits aussi **plans stochastiques**. Ces plans se caractérisent par le fait que les individus statistiques devant faire partie de l'échantillon sont sélectionnés par tirages probabilistes. Chaque individu de la population statistique a une probabilité connue d'être inclus dans l'échantillon (cette probabilité est appelée **probabilité d'inclusion d'ordre un** de l'individu pour le plan d'échantillonnage considéré). Avec de tels plans, il est possible d'utiliser la théorie des probabilités: les observations sur l'échantillon sont des variables aléatoires. On peut utiliser des outils d'inférence statistique pour estimer des paramètres de la population et également évaluer les précisions d'estimation.
- **les plans non probabilistes**, dits aussi **plans empiriques** ou **plans par choix raisonné**. L'échantillon est construit par des procédés comportant une part d'arbitraire et ne permettant pas l'évaluation de la précision d'estimation.

Les plans probabilistes classiques sont les suivants : plan aléatoire simple, plan aléatoire systématique, plan aléatoire stratifié, plan aléatoire en groupes.

Les plans non probabilistes sont utilisés dans les études qualitatives où il n'est pas envisagé une extrapolation à la population statistique dans son entier. Quelques exemples:

- Plan par commodité : on choisit des individus statistiques qui sont d'accès facile
- Enquête boule de neige : on choisit quelques individus (au sein d'une population humaine) qui sont pertinents pour l'étude, et ensuite on leur demande de proposer d'autres individus pour l'enquête.
- Plan par quotas : on construit un échantillon qui respecte les proportions connues pour certaines catégories de la population.

1.7 Représentativité d'un échantillon

La définition d'**échantillon représentatif** diffère selon que le plan d'échantillonnage est probabiliste ou non probabiliste :

- un plan probabiliste fournit un échantillon représentatif dès lors que chaque individu de la population a une *probabilité connue et non nulle d'être inclus dans l'échantillon.*
- un plan non probabiliste fournit un échantillon représentatif si la *structure de l'échantillon pour certaines variables clés est similaire à celle de la population cible.* Par exemple, on peut vouloir construire un échantillon pour lequel les proportions de catégories d'individus soient similaires dans l'échantillon à celles de la population cible (c'est le principe de la méthode dite des quotas).

En population fixe, un échantillon n'est représentatif que de la population au sein de laquelle il a été sélectionné.

1.8 Précision statistique, distribution d'échantillonnage

Une **statistique** est une valeur calculée à partir d'observations effectuées sur les individus de l'échantillon. Comme l'objectif de l'échantillonnage est généralement d'**inférer** sur la population statistique (c'est-à-dire tirer des conclusions concernant cette population), le calcul de cette statistique correspond souvent à **l'estimation d'un paramètre** (c'est-à-dire l'évaluation numérique de ce paramètre), paramètre de la population sur laquelle est prélevé l'échantillon. Les deux exemples les plus courants sont indiqués dans la table 2.

L'ensemble des valeurs possibles d'une statistique, affectées de leur probabilité de réalisation s'appelle la **distribution d'échantillonnage** de cette statistique. Les figures 1 et 2 nous montrent deux exemples de distributions d'échantillonnage.

La distribution d'échantillonnage d'une statistique peut correspondre à une loi de probabilité usuelle. Ainsi, le nombre d'individus malades observé dans l'échantillon suit :

- une loi binomiale si les tirages sont effectués avec remise,
- une loi hypergéométrique si les tirages sont effectués sans remise,
- une loi approximativement gaussienne si les tirages sont suffisamment nombreux.

La **précision statistique** d'une méthode d'estimation d'un paramètre de la population est définie comme *une mesure de l'écart entre l'estimation obtenue à partir de l'échantillon et la vraie valeur du paramètre*. Cet écart est attribuable à deux types d'erreur :

- **l'erreur d'échantillonnage** : c'est l'erreur liée à l'aléa de tirage de l'échantillon car, à partir d'un échantillon, quand on calcule une statistique, on obtient une valeur parmi toutes les valeurs possibles de la distribution d'échantillonnage de cette statistique. L'erreur d'échantillonnage diminue généralement avec l'accroissement de la taille d'échantillon.
- **l'erreur d'observation** : elle est la conséquence d'erreur de mesure, de notation lors de la cueillette de l'information, mais aussi, dans le cas d'enquêtes auprès de personnes, des réponses erronées, des refus de réponse. Ce type d'erreur peut être minimisé par une formation approfondie des observateurs ou des enquêteurs et par le contrôle de la qualité du travail effectué aux différentes étapes du plan d'échantillonnage.

Un **estimateur d'une fonction d'intérêt** $\theta(y_1, y_2, \cdots, y_N)$ (ou plus simplement d'un paramètre θ) est une statistique T qui fournit une évaluation pertinente de θ.

L'**espérance mathématique** $E(T)$ d'une statistique T est une moyenne théorique qui est la somme des valeurs possibles de T pondérées par leurs probabilités de réalisation. On l'appelle aussi **valeur espérée** de T.

Statistique	**Paramètre**
Moyenne de l'échantillon	Moyenne de la population
Proportion dans l'échantillon	Proportion dans la population

Table 2: Exemples de statistique associée à un paramètre

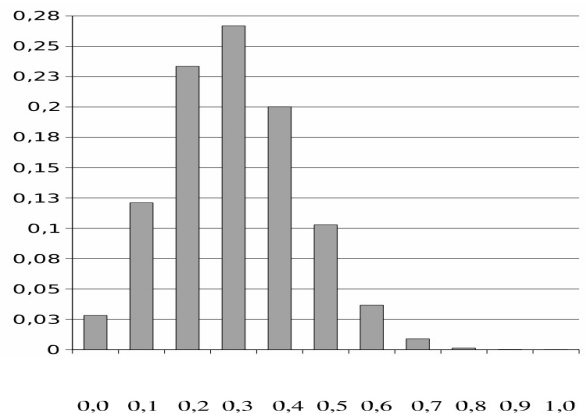

Figure 1: Exemple de distribution d'échantillonnage de la proportion observée pour une taille d'échantillon $n = 10$ obtenu par **tirages avec remise** dans une population de taille 50. La proportion dans la population (inconnue et à estimer dans la pratique) est $p = 0,3$.

Le **biais** $B(T)$ de la statistique T pour le paramètre θ est l'écart entre la valeur espérée de T et θ :

$$B_\theta(T) = E(T) - \theta.$$

La **variance d'échantillonnage** $V(T)$ de la statistique T est par définition

$$V(T) = E((T - E(T))^2).$$

L'**écart quadratique moyen** $EQM(T)$ d'estimation de θ par T est défini de la façon suivante

$$EQM_\theta(T) = E((T - \theta)^2).$$

Le lien entre écart quadratique, biais et variance d'échantillonnage est :

$$EQM_\theta(T) = V(T) + (B_\theta(T))^2.$$

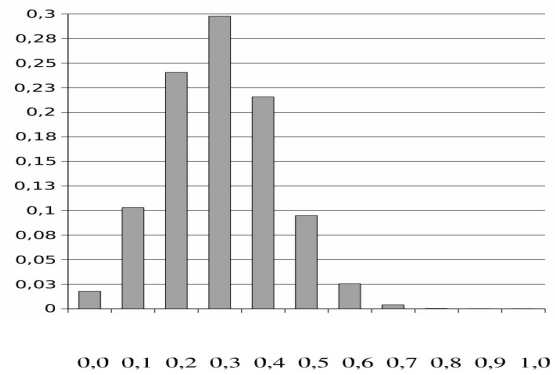

Figure 2: Exemple de distribution d'échantillonnage de la proportion observée pour une taille d'échantillon $n = 10$ obtenu par **tirages sans remise** dans une population de taille 50. La proportion dans la population (inconnue et à estimer dans la pratique) est $p = 0,3$.

L'estimateur T est dit **sans biais** pour θ si $E(T) = \theta$, ce qui est équivalent à $B_\theta(T) = 0$. La figure 3 illustre les notions de distribution, de biais et de variabilié d'échantillonnage. Elle fait le parallèle suivant entre qualités d'un estimateur et d'un tireur sur cible. Un estimateur fournit, pour un échantillon donné, une évaluation numérique du paramètre considéré; un tireur obtient, pour un tir effectué, un impact sur la cible alors qu'il cherche à atteindre le centre de la cible. Les estimations varient d'un échantillon à l'autre; les points d'impact varient d'un tir à l'autre. On espère qu'avec un bon estimateur on a des valeurs estimées qui ne sont pas trop éloignées du paramètre; et des points d'impact pas trop éloignés du centre de la cible pour un bon tireur.

Généralement, on recherche un estimateur qui soit de moindre écart quadratique moyen. La situation la plus intéressante, dans la pratique, est de disposer d'un estimateur sans biais et de moindre variance. La **convergence** est une autre propriété très

Figure 3: Illustration du biais et de la variabilité d'un estimateur

recherchée : elle signifie que plus on a des données, plus notre estimation se rapproche de la vraie valeur du paramètre inconnu.

Les trois méthodes d'estimation les plus répandues sont celles du maximum de vraisemblance, des moments et des moindres carrés.

Estimation par maximum de vraisemblance : La vraisemblance est une fonction du paramètre θ conditionnelle aux observations. Elle traduit la probabilité d'observer l'échantillon obtenu pour la valeur θ du paramètre.

Par exemple, on choisit n individus par tirages indépendants dans une population dont une proportion p est malade. Notons X le nombre d'individus malades dans l'échantillon. La vraisemblance est par conséquent la fonction :

$$p \longmapsto C_n^X p^X (1-p)^{n-X}$$

car X suit une loi binomiale de paramètres n et p en tant que nombre de succès au bout de n épreuves avec pour chaque épreuve une probabilité de succès égale à p.

La valeur qui maximise cette fonction n'est autre que X/n c'est-à-dire la proportion de malades observée dans l'échantillon.

Estimation des moments (ou M-estimation) : Soit k un entier supérieur à 1. Le moment d'ordre k d'une variable aléatoire X, noté $m_k(X)$, est l'espérance de X^k : $m_k(X) = E(X^k)$. La méthode des moments est utilisée quand on sait expliciter les moments d'une statistique en fonction des paramètres inconnus que l'on veut estimer. Pour

un nombre r de paramètres à estimer, on utilise les moments d'ordre 1 à r pour construire un système de r équations à r inconnues. Dans ce système, on introduit les moments dit empiriques puis on calcule les r solutions qui sont appelées estimateurs des moments.

Prenons l'exemple de comptages $x_1, x_2, \cdots x_n$ d'individus d'une espèce dans n unités expérimentales de même taille. Pour des individus à comportement indépendant mais avec des affinités communes, on admet que la loi de probabilité du nombre d'individus X dans une unité expérimentale suit une loi binomiale négative d'espérance μ et de paramètre d'agrégation γ. Pour estimer μ et γ par la méthode des moments, on utilise les expressions des moments d'ordre 1 et 2 de X :

$$E(X) = \mu \quad \text{et} \quad E(X^2) = \mu + (1 + \frac{1}{\gamma})\mu^2.$$

En remplaçant dans ces équations, les termes de gauche par les moments empiriques $\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\widehat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, on obtient :

$$\widehat{m}_1 = \mu \quad \text{et} \quad \widehat{m}_2 = \mu + (1 + \frac{1}{\gamma})\mu^2.$$

La résolution en μ et γ de ce sytème d'équations nous donne les solutions :

$$\widehat{\mu} = \widehat{m}_1 \quad \text{et} \quad \widehat{\gamma} = \frac{\widehat{m}_1^2}{\widehat{m}_2 - \widehat{m}_1 - \widehat{m}_1^2}$$

qui sont donc les M-estimateurs de μ et de γ.

Estimation des moindres carrés : Le principe est de minimiser la somme des carrés d'écart entre valeurs observées et valeurs espérées sous un modèle.

Une application classique de cette méthode est l'estimation des paramètres du modèle de régression linéaire simple. Ce modèle associe une variable quantitative (dite *réponse*) à une variable explicative de la faççon suivante : $y_i = ax_i + b + \epsilon_i$, où y_i et x_i sont les valeurs prises par la réponse et la variable explicative sur l'individu statistique i. La pente a et l'interception b sont appelées paramètre de régression. ϵ_i est l'erreur expérimentale pour l'individu i.

La méthode des moindres carrés vise ici à minimiser la quantité $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ par

rapport à a et b. Les valeurs qui réalisent ce minimum sont :

$$\widehat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{et} \quad \widehat{b} = \bar{y} - \widehat{a}\bar{x}$$

où $\bar{x}$ est la moyenne des x_i et $\bar{y}$ celle des y_i.

2 Quelques plans d'échantillonnage classiques

Nous considérons le cas d'une population statistique de taille N.

2.1 Plan aleatoire simple

Le plan aléatoire simple (PAS) de taille n consiste à effectuer n tirages équiprobables dans la population statistique. Les tirages peuvent être avec ou sans remise.

Pour un plan aléatoire simple sans remise (PASSR), on a C_N^n échantillons possibles, ayant tous la même probabilité de réalisation.

Pour un plan aléatoire simple avec remise (PASAR), on a N^n échantillons possibles, ayant tous la même probabilité de réalisation.

Le plan aléatoire simple est utilisé en phase exploratoire quand on désire estimer un paramètre de la population et qu'il n'y a pas de structure spatiale à étudier.

La moyenne d'échantillon pour un PAS est sans biais pour la moyenne de la population.

2.2 Plan aléatoire stratifié

La population est divisée en H strates de taille $N_1, \cdots, N_H$. La moyenne d'échantillon dans la strate h est notée $\overline{y}_h$. La procédure d'échantillonnage consiste à exécuter un PASSR de taille n_h dans la strate h, indépendamment des autres strates.

Le nombre d'échantillons possibles est $\prod_{h=1}^{H} C_{N_h}^{n_h}$.

La moyenne d'échantillon global n'est pas forcément sans biais pour $\overline{Y}$ pour ce type d'échantillonnage. On utilise donc la moyenne dite stratifiée qui, elle, est sans biais :

$$\overline{y}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h$$

et de variance $Var(\overline{y}_{st}) = \sum_{h=1}^{H} \Big(\frac{N_h}{N}\Big)^2 (1 - \frac{n_h}{N_h}) \frac{V_h}{n_h} \frac{N_h}{N_h - 1}$ où V_h est la variance de la strate h.

Le plan aléatoire stratifié est intéressant quand la variabilité intra-strate est faible.

2.3 Plan aléatoire en grappes

La population est divisée en G grappes, pas forcément de même taille. L'échantillonnage consiste à choisir g grappes selon un plan aléatoire simple sans remise.

Le nombre d'échantillons possibles est C_G^g.

Un estimateur sans biais de la moyenne de la population est donnée par

$\bar{y}_{grappes} = \dfrac{G}{g \times N} \times$ [somme des valeurs observées sur les grappes échantillonnées].

Le plan aléatoire en grappes est intéressant quand la variabilité inter-grappe est faible.

2.4 Plan aléatoire systématique

Dans le cas d'une population ordonnée, de taille $N = kn$, le plan consiste à choisir un individu dans $\{1, \cdots, k\}$, soit i, et à constituer l'échantillon $\{i, i + k, \cdots, i + (n - 1)k\}$.

On a seulement N/n échantillons possibles. La moyenne d'échantillon est sans biais pour la moyenne de la population.

Ce plan est utilisé quand une exploration spatiale a un intérêt. Il est décommandé en cas de périodicité supposée de la variable étudiée si l'on veut estimer la moyenne de la population.

3 Recherche d'une procédure d'échantillonnage

La figure 4 schématise le processus décisionnel permettant de mettre au point un plan d'échantillonnage. Il s'agir de trouver un juste milieu entre l'effort d'échantillonnage qui sera consenti et la fiabilité du plan en termes de précision ou de minimisation des risques d'erreur.

Le processus décisionnel décrit en figure 3 conduit souvent à la succession d'étapes méthodologiques que voilà :

1. *Etude bibliographique.* Il s'agit de mettre à profit des études antérieures pour construire un plan de sondage performant.
2. *Définition claire des objectifs de l'échantillonnage (ou sondage).* Cette étape doit déboucher sur la définition des variables à prendre en compte et la confection d'une feuille de saisie (ou d'un questionnaire quand il s'agit de sondages d'opinion).
3. *Définition de la population à étudier.* Elle doit être définie sans ambiguité. On définit d'abord la population cible puis on détermine la liste des unités statistiques sélectionnables, autrement dit la base de sondage.
4. *Construction du plan de sondage.* Il s'agit de déterminer la façon dont les individus doivent être sélectionnés, d'organiser l'observation en fonction des contraintes naturelles et techniques. Si les individus sont sélectionnés selon une procédure aléatoire, on parle de plan probabiliste. Sinon, on parle de plan empirique.
5. *Collecte des informations.* L'exécution du plan doit respecter les règles établies à l'étape précédente. La collecte des informations peut se faire à partir d'une base de données informatique, mais aussi à l'aide d'observations sur le terrain, en plein champ. Il peut s'agir également d'enquêtes faites par un enquêteur par entretien, par courrier, par téléphone, par examen. Quelque soit la procédure, il est nécessaire que la qualité, la fiabilité des données soient garanties.
6. *Encodage et archivage des données.* Il s'agit de choisir les logiciels ou programmes informatiques les plus appropriés.
7. *Traitement statistique des données.* Les méthodes doivent tenir compte des caractéristiques du plan d'échantillonnage.

Une partie de la théorie des sondages consiste en l'étude des propriétés de la distribution d'échantillonnage de la moyenne d'échantillon pour différents plans d'échantillonnage

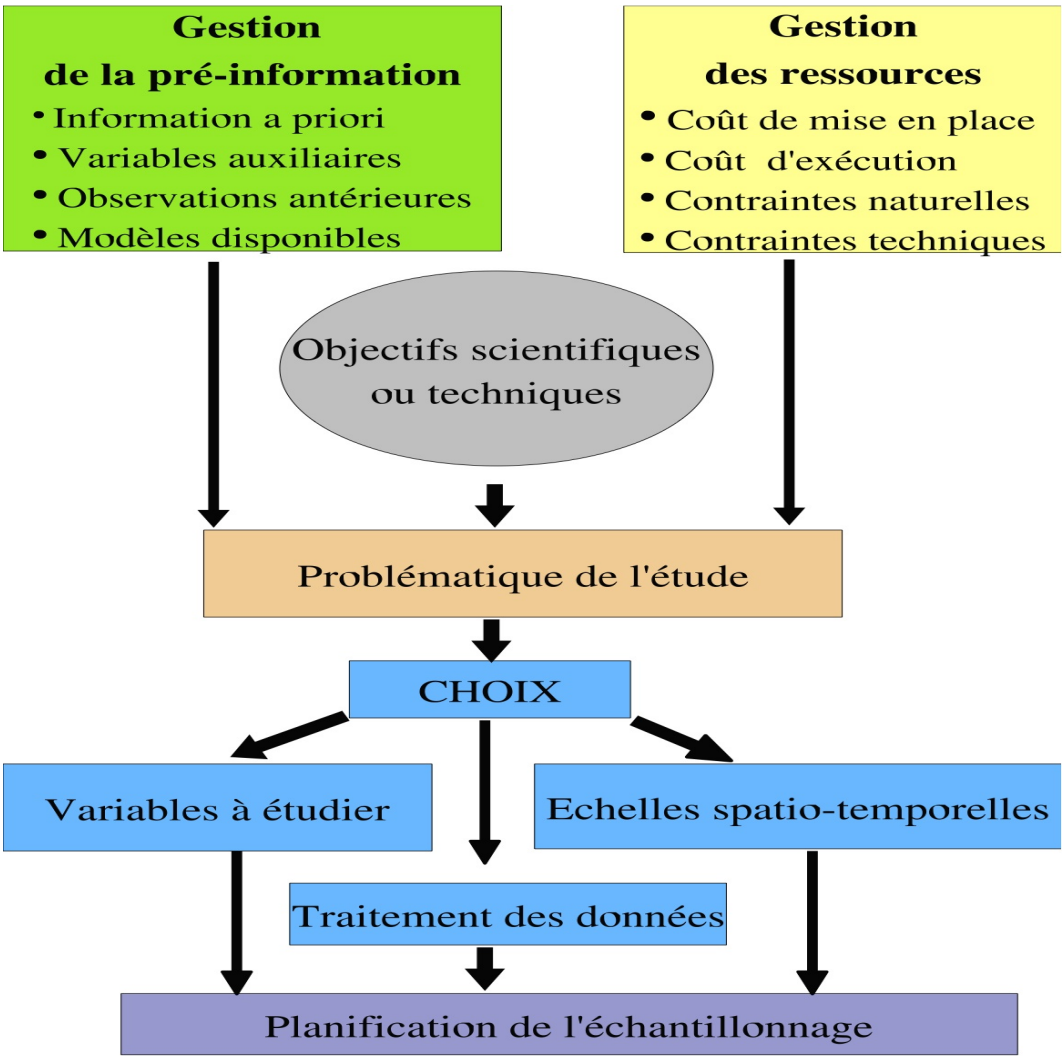

Figure 4: Schéma du processus décisionnel pour le choix d'un plan d'échantillonnage

dans le cadre d'une population finie fixe. En d'autres termes, on aimerait connaître les propriétés de la série statistique que l'on obtiendrait si, pour la population statistique considérée, l'on pouvait :

1. réaliser tous les échantillons possibles avec ce plan d'échantillonnage,
2. calculer la moyenne de chacun de ces échantillons,
3. constituer une série statistique avec ces moyennes d'échantillon.

Illustration numérique : On a une population $\mathcal{P} = \{1, 2, 3, 4\}$ sur laquelle un caractère $\mathcal{Y}$ prend les valeurs $y_1 = 17, y_2 = 8, y_3 = 8$ et $y_4 = 23$. On échantillonne en effectuant deux tirages sans remise dans $\mathcal{P}$. Les trois étapes décrites ci-dessus deviennent :

1. Les échantillons possibles sont $\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}$ et $\{3, 4\}$
2. Les moyennes de ces échantillons sont respectivement 12,5; 12,5; 20; 8; 15,5 et 15,5.
3. La série statistique est donc 12,5 12,5 20 8 15,5 15,5 La distribution d'échantillonnage de la moyenne d'échantillon est donc P(8)=1/6, P(12,5)=1/3, P(15,5)=1/3, P(20)=1/6. Son espérance mathématique est 14 et sa variance 13,5.

Bien sûr, dans la pratique, la taille de population N est bien plus grande et il est utile de rappeler que l'on ne connaîtra les valeurs y_i que pour les individus i inclus dans l'échantillon. On verra qu'il est possible, pour beaucoup de plans d'échantillonnage, d'exprimer l'espérance m et la variance V de la distribution d'échantillonnage de la moyenne d'échantillon puis d'estimer ces paramètres m et V à l'aide des données de l'échantillon observé. On tâchera de répondre plus particulièrement aux questions essentielles suivantes :

Comment échantillonner (quel plan d'échantillonnage appliquer)?

Quelle taille d'échantillon adopter (quel taux de sondage appliquer)?

Comment estimer une fonction d'intérêt avec une bonne précision?

Ouvrages conseillés :

GRAIS Bernard. (1992): Méthodes statistiques. Dunod, 3ième édition.

GROSBRAS Jean-Marie. (1987) : Méthodes statistiques des sondages, Economica.

MORIN Hervé. (1993) Théorie de l'échantillonnage, Les presses de l'Université Laval.

TILLE Yves. (2001) Théorie des sondages, Dunod.