PratiqueR

Techniques de sondage

De la théorie à la pratique avec R

Tome 1

Camelia Goga
Hélène Juillard
Alina Matei
Anne M. Ruiz

edp sciences

Techniques de sondage

De la théorie à la pratique avec R – Tome 1

Camelia Goga – Hélène Juillard – Alina Matei – Anne M. Ruiz

Performant, évolutif, libre, gratuit, le monde du logiciel libre s'est imposé depuis une vingtaine d'années comme la base des outils de calculs et d'intelligence artificielle déportés dans le cloud ou en local. Parmi les langages de cet écosystème gratuit et multiplateformes R, python ou julia sont devenus des outils incontournables en machine learning, IA, optimisation ou statistiques tant dans les milieux académiques qu'industriels.

La collection « PratiqueR » répond à cette évolution récente et propose d'intégrer pleinement l'utilisation d'un langage dans des ouvrages couvrant les aspects théoriques et pratiques de diverses méthodes statistiques appliquées à des domaines aussi variés que l'analyse des données, la gestion des risques, les sciences médicales, l'économie, etc.

Elle s'adresse aux étudiants, enseignants, ingénieurs, praticiens et chercheurs de ces différents domaines qui utilisent quotidiennement des données dans leur travail et qui apprécient ces langages pour leur fiabilité, leur confort d'utilisation et leur extensibilité via des modules ou des packages.

La collection **PratiqueR** est dirigée par Pierre-André Cornillon et Eric Matzner-Løber

Dans ce premier tome de l'ouvrage *Techniques de sondage : de la théorie à la pratique avec **R***, les principales méthodes de tirage d'échantillon et d'estimation de totaux et de moyennes sont présentées d'un point de vue théorique, puis mises en œuvre avec **R**, à partir d'exemples inspirés de situations réelles classiques. Le cadre étudié est celui d'une population finie où l'inférence statistique repose sur le tirage aléatoire des unités et sur des méthodes d'estimation adaptées à ce cadre.

Ce premier tome étudie les stratégies d'échantillonnage probabilistes élémentaires, à probabilités égales ou inégales, avec ou sans remise, utilisant les estimateurs classiques de Horvitz-Thompson et de Hansen-Hurwitz. Les plans de sondage stratifiés, en grappes et à deux degrés y sont également décrits de manière détaillée. De nombreux exemples numériques et applications à des enquêtes réelles facilitent la compréhension des résultats théoriques. Des études par simulation Monte-Carlo complètent l'analyse en étudiant le biais et la variance des estimateurs sous différents plans de sondage, soulignant les avantages, les limites et les pistes d'amélioration possibles des stratégies étudiées.

Alliant rigueur théorique et mise en œuvre pratique avec **R**, l'ouvrage s'adresse à toutes celles et ceux qui souhaitent maîtriser les principes et l'application des techniques de sondage. Les autrices, qui enseignent les sondages et la statistique depuis de nombreuses années et ont acquis une solide expérience pratique grâce à leurs collaborations avec l'Ined, La Poste et EDF, mettent ici à profit leur expérience pédagogique et leur connaissance approfondie du domaine. Le tome 1 constitue une ressource précieuse pour un cours de sondages de niveau débutant ou intermédiaire, tandis que le tome 2 est consacré aux méthodes avancées : échantillonnage spatial, estimation de paramètres non-linéaires, estimation avec prise en compte d'information auxiliaire et théorie asymptotique.

Les données, codes, corrections des exercices et compléments sont accessibles sur le site :

https://techniques-sondage-r.github.io/Tome-1/

9 782759 839797

978-2-7598-3979-7

www.edpsciences.org

edp sciences

Techniques de sondage

De la théorie à la pratique avec R

Tome 1

Camelia Goga, Hélène Juillard,
Alina Matei, Anne M. Ruiz

Techniques de sondage

De la théorie à la pratique avec R

Tome 1

edp sciences

ISBN (papier) : 978-2-7598-3979-7 — ISBN (ebook) : 978-2-7598-3980-3

Imprimé en France

Collection Pratique R

dirigée par Pierre-André Cornillon et Eric Matzner-Løber

Université Rennes-2
et ENSAE formation continue Le Cepe, France

Comité éditorial

Déjà paru dans la même collection :

Régression avec Python – Pierre-André Cornillon, Nicolas Hengartner, Eric Matzner-Løber, Laurent Rouvière, 2025
ISBN : 978-2-7598-1779-5 – EDP Sciences

Régression avec R, 3e édition – Pierre-André Cornillon, Nicolas Hengartner, Eric Matzner-Løber, Laurent Rouvière, 2023
ISBN : 978-2-7598-1779-5 – EDP Sciences

Calcul parallèle avec R – Vincent Miele, Violaine Louvet, 2016
ISBN : 978-2-7598-2060-3 – EDP Sciences

Séries temporelles avec R – Yves Aragon, 2016
ISBN : 978-2-7598-1779-5 – EDP Sciences

Psychologie statistique avec R – Yvonnick Noël, 2015
ISBN : 978-2-7598-1736-8 – EDP Sciences

Réseaux bayésiens avec R – Jean-Baptiste Denis, Marco Scutati, 2014
ISBN : 978-2-7598-1198-4 – EDP Sciences

Analyse factorielle multiple avec R – Jérôme Pagès, 2013
ISBN : 978-2-7598-0963-9 – EDP Sciences

Méthodes de Monte-Carlo avec R – Christian P. Robert, George Casella, 2011
ISBN : 978-2-8178-0181-0 – Springer

REMERCIEMENTS

Les autrices souhaitent d'abord rendre hommage à Yves Aragon, collègue décédé dont le cours de sondage rédigé en collaboration avec certaines d'entre elles, a constitué le point de départ de cet ouvrage.

Nos collègues Hervé Cardot, Eve Leconte, Jean-Louis Marchand et Louis-Paul Rivest ont consacré du temps à une lecture attentive du manuscrit ; leurs remarques et suggestions ont contribué de manière déterminante à son amélioration. Thibault Laurent nous a apporté une aide précieuse sur les fichiers de données.

Mehdi Dagdoug et Estelle Médous, anciens doctorants ayant enseigné le cours de sondages à Besançon et Toulouse, ont enrichi ce travail par leurs retours issus de cette expérience pédagogique.

Nous remercions aussi les nombreuses promotions d'étudiants auxquelles nous avons enseigné, notamment celles des Masters « Modélisation Statistique » de Besançon et « Économétrie, Statistiques » de Toulouse, en formation initiale, continue et à distance. Ces années d'enseignement ont affiné notre approche pédagogique. Nous adressons une mention particulière à Alexandra Amani et José Bancole pour leurs contributions.

Nous sommes aussi reconnaissantes à nos collègues spécialistes du sondage à travers le monde, et particulièrement aux participants des colloques internationaux francophones sur les sondages, pour leurs échanges stimulants. Nous pensons particulièrement à Jean-Claude Deville dont l'enseignement, la recherche et les discussions nous ont inspirées dans notre apprentissage des sondages. Nous remercions également Yves Tillé pour son rôle pionnier dans l'utilisation de R en théorie des sondages.

Nous remercions chaleureusement nos éditeurs, Pierre-André Cornillon et Eric Matzner-Løber, pour leur patience et leur soutien constants tout au long de ce projet.

Enfin, nous adressons un grand merci à nos proches, qui nous ont soutenues et encouragées tout au long de la phase d'écriture, ainsi que pour leurs conseils ayant largement contribué à l'amélioration de ce document.

AVANT-PROPOS

Les techniques de sondage occupent une place centrale en statistique publique et constituent un outil fondamental de production d'indicateurs socio-économiques. Leur utilisation dépasse toutefois ce cadre et elles sont largement mobilisées dans de nombreux domaines scientifiques et industriels.

Ce livre prolonge le cours de sondage enseigné dans les universités de Besançon, Dijon et Toulouse Capitole (voir aussi Goga, 2018). Progressivement enrichi et approfondi au fil des années d'enseignement et d'applications pratiques, l'ouvrage a mûri, s'est densifié et s'est finalement structuré en deux tomes complémentaires.

Ce livre s'adresse à toutes celles et ceux qui souhaitent mettre en œuvre des techniques de sondage avec le logiciel R (R Core Team, 2025) : étudiants et enseignants pour des cours de niveau débutant à avancé, professionnels amenés à travailler sur des enquêtes, et toute personne désirant maîtriser R dans ce domaine. Il existe de nombreux packages R permettant la mise en œuvre de techniques de sondage, comme recensés sur le site du CRAN [1]. Cet ouvrage s'appuie essentiellement sur deux packages R : **sampling** [2] (Tillé & Matei, 2025), pour le tirage d'échantillons selon divers plans et l'estimation de paramètres, et **survey** (Lumley *et al.*, 2025) [3], pour l'estimation de paramètres et l'estimation de la variance des estimateurs sous de nombreux plans de sondage. Le package **stratification** (Rivest & Baillargeon, 2022) complète ces outils pour la construction de strates optimales. Les codes de ce tome ont été exécutés avec la version 4.4.3 du logiciel R et le générateur de nombres aléatoires par défaut [4]. Les codes, corrections des exercices et compléments sont disponibles sur le site github du livre : `https://techniques-sondage-r.github.io/Tome-1/`.

Notre ouvrage poursuit trois objectifs complémentaires. D'abord, exposer de manière rigoureuse les définitions et propriétés des principales méthodes d'échantillonnage et d'estimation en populations finies, avec l'inférence basée sur le tirage aléatoire des individus (en s'inspirant du livre de Särndal *et al.* (1992) qui est l'ouvrage de référence en la matière). Ensuite, fournir des explications détaillées pour leur mise en œuvre sous R à partir d'exemples simples mais représentatifs de situations réelles classiques. Enfin, illustrer les propriétés théoriques et comparer les méthodes grâce à des études par simulation Monte-Carlo. Si les livres classiques Särndal *et al.* (1992) et Lohr (2010, 2021) en anglais et Ardilly (2006), Tillé (2001, 2019) et Ardilly & Tillé (2003) en français constituent des références incontournables, notre ouvrage se distingue par l'intégration systématique des développements théoriques et de leur implémentation avec R. Les études de Monte-Carlo, absentes habituellement des manuels, ne constituent pas un simple complément

1. `https://cran.r-project.org/web/views/OfficialStatistics.html`
2. `https://cran.r-project.org/web/packages/sampling/index.html`
3. `https://cran.r-project.org/web/packages/survey/index.html`
4. à savoir, l'algorithme « Mersenne-Twister » et les méthodes « Inversion » et « Rejection »

illustratif : elles jouent un rôle central pour analyser le comportement des méthodes et approfondir la compréhension des résultats théoriques.

Chaque chapitre propose de nombreux exercices, à la fois théoriques, numériques, et pratiques avec R, dont les corrigés détaillés sont accessibles sur le site github de l'ouvrage. De nombreux exemples numériques et applications à des enquêtes réelles facilitent la compréhension des résultats théoriques. Les exemples d'enquêtes, bien que tirés de situations réelles, sont volontairement très simplifiés pour illustrer les concepts sans en refléter toute la complexité méthodologique. Des études Monte-Carlo complètent l'analyse en examinant le biais et la variance des estimateurs sous différents plans de sondage, mettant en évidence les avantages, les limites et les pistes d'amélioration possibles des stratégies étudiées. Chaque chapitre se termine par une section intitulée « Synthèse et compléments », qui récapitule les principaux résultats présentés et propose des pistes de réflexion ainsi que des références bibliographiques pour approfondir et prolonger les notions exposées. Cet ouvrage ne prétend pas couvrir l'ensemble des techniques de sondage. Il se concentre sur l'erreur d'échantillonnage et n'aborde pas les erreurs de couverture ou de non-réponse, les plans à deux phases, les méthodes de variance par rééchantillonnage (jackknife, bootstrap), l'estimation sur domaines, etc.

Le chapitre 1 du tome 1 pose les concepts fondamentaux de la théorie des sondages et présente les principes d'échantillonnage et d'estimation en population finie. Ces principes sont illustrés par de nombreux exemples d'enquêtes réelles. Les notions de plans de sondage probabilistes et d'estimateurs de Horvitz-Thompson et de Hansen-Hurwitz pour un total ou une moyenne sont détaillées, ainsi que l'estimation de leur variance. Ces notions sont expliquées dans un formalisme mathématique rigoureux, complété par des exemples numériques. Les trois jeux de données utilisés dans les implémentations R de l'ouvrage sont également décrits. Ce chapitre met en lumière les liens et les différences entre la théorie des sondages et la statistique classique, dans laquelle l'inférence est habituellement basée sur l'hypothèse forte que les variables aléatoires étudiées sont indépendantes et identiquement distribuées. En particulier, la notion de stratégie d'échantillonnage, constituée d'un plan de sondage et d'un estimateur, est fondamentale et propre à la théorie des sondages.

Une fois introduits les concepts de base, ce premier tome propose dans les chapitres 2 à 6, une étude des plans de sondage les plus classiques de la théorie des sondages. Dans chacun de ces chapitres, on donne une définition des plans et/ou d'algorithmes permettant leur mise en œuvre. Les fonctions en R pour ces implémentations sont extraites de packages existants ou écrites pour l'occasion. Pour chaque plan, on étudie les estimateurs de totaux ou moyennes classiques : Horvitz-Thompson et Hansen-Hurwitz, ainsi que les formules de variance et leurs estimateurs. Les propriétés de ces différentes stratégies sont étudiées en détail avec le même formalisme mathématique et exemples numériques qu'au chapitre 1. Des simulations Monte-Carlo sur les jeux de données introduits au chapitre 1 com-

plètent l'étude des différentes stratégies en comparant en particulier les coefficients de variation empiriques (Monte-Carlo) des estimateurs.

Le chapitre 2 présente les plans de sondage à probabilités égales sans remise : simple, Bernoulli et systématique. Ces trois plans de base, qu'il faut maîtriser pour comprendre la suite, servent de briques élémentaires aux chapitres suivants. On insiste sur leurs limites, qui justifient le recours aux stratégies plus complexes présentées ensuite.

Le chapitre 3 aborde les plans à probabilités inégales (avec et sans remise). Ce chapitre présente un statut particulier en raison de la complexité des algorithmes, notamment pour les tirages sans remise. Il expose les plans les plus connus : multinomial, Poisson, systématique à probabilités inégales et Poisson conditionnel à la taille.

Le chapitre 4 détaille les plans stratifiés, incluant la construction de strates et le choix de l'allocation de la taille d'échantillon aux strates. Ces plans, à probabilités égales ou inégales, sont très utilisés en pratique. On montre en particulier l'intérêt de la stratification, en termes de variance, pour l'estimation des totaux ou des moyennes des variables dispersées, et on discute le cas de l'estimation d'une proportion (variable dichotomique et donc de faible dispersion).

Les chapitres 2 à 4 traitent de l'échantillonnage direct dans la population cible, qui suppose l'accès à une base de sondage pour cette population. En pratique, cet accès est souvent impossible ou trop coûteux : on recourt alors à des plans indirects, dont le plus simple est le plan à plusieurs degrés. Ce plan hiérarchique peut utiliser, à chaque niveau, l'un des plans de sondage présentés aux chapitres 2 à 4. Dans cet ouvrage, on se concentre sur les plans à deux degrés (niveaux), le chapitre 5 traitant du cas particulier d'un recensement au deuxième degré (plans en grappes) et le chapitre 6 traitant du cas général.

Dans le tome 1, les différentes stratégies d'échantillonnage reposent sur les estimateurs classiques de Horvitz-Thompson ou de Hansen-Hurwitz. Lorsque de l'information auxiliaire est disponible, elle est uniquement intégrée au niveau du plan de sondage (plans proportionnels à la taille ou stratifiés). Le tome 2 prolonge cette approche en introduisant des méthodes d'estimation plus élaborées et basées sur l'information auxiliaire (estimateurs de type ratio, post-stratifié et par calage). Alors que dans le tome 1 l'inférence statistique est uniquement sous le plan de sondage, dans le tome 2, on prend aussi en compte une inférence basée sur un modèle de super-population. L'inférence assistée par un modèle permet de construire de nouveaux estimateurs et de les étendre à des méthodes de régression de type ridge, sur composantes principales ou non paramétriques. Le tome 2 aborde également d'autres problématiques avancées, telles que l'échantillonnage équilibré et l'échantillonnage spatial, ainsi que l'estimation de paramètres non linéaires, en présentant les résultats de théorie asymptotique nécessaires à leur étude.

NOTATIONS

$U = \{1, \ldots, k, \ldots, N\}$	Population de taille N
$U_d \subset U$	Domaine de taille N_d, $d = 1, \ldots, D$
$U_h \subset U$	Strate de taille N_h, $h = 1, \ldots, H$
$U_i \subset U$	Unité primaire (ou grappe) de taille N_i, $i = 1, \ldots, N_I$
$s \subseteq U$	Echantillon (partie de U)
$\mathcal{S}$	Ensemble des parties de U
n_s	Taille de l'échantillon s (notée n si le plan est de taille fixe)
$f = n/N$	Taux de sondage
$p(\cdot)$, $p(s)$	Plan de sondage, probabilité de sélection d'un échantillon s
I_k	Indicatrice d'appartenance de l'unité $k \in U$ à un échantillon
$\pi_k,\ k \in U$	Probabilité d'inclusion d'ordre un
$\pi_{kl},\ k, l \in U$	Probabilité d'inclusion d'ordre deux
$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$	Covariance des indicatrices I_k et I_l
s_o	Echantillon ordonné, m-uple des individus tirés après m tirages avec remise
R_k	Nombre de fois que l'individu $k \in U$ a été sélectionné après m tirages avec remise
$p_k,\ k \in U$	Probabilité de sélection, tirage avec remise
y, y_k	Variable d'intérêt et sa valeur pour le k-ème individu, $k \in U$
x, x_k	Variable auxiliaire et sa valeur pour le k-ème individu, $k \in U$
t_{yU}, t_{xU}	Totaux des variables y et x sur U
$\overline{y}_U$, $\overline{x}_U$	Moyennes des variables y et x sur U
$\overline{y}_s$, $\overline{x}_s$	Moyennes des variables y et x sur s
S^2_{yU}, S^2_{ys}	Variance corrigée de y dans la population et dans l'échantillon
$\hat{t}_{y\mathrm{HT}}$, $\hat{\overline{y}}_{\mathrm{HT}}$	Estimateurs de Horvitz-Thompson de t_{yU}, $\overline{y}_U$
$\hat{t}_{y\mathrm{HH}}$, $\hat{\overline{y}}_{\mathrm{HH}}$	Estimateurs de Hansen-Hurwitz de t_{yU}, $\overline{y}_U$
$\mathbb{E}(\hat{t})$	Espérance d'un estimateur $\hat{t}$
$\mathrm{V}(\hat{t})$	Variance d'un estimateur $\hat{t}$
$\mathrm{AV}(\hat{t})$	Variance asymptotique d'un estimateur $\hat{t}$
$\mathrm{Cov}(\hat{t}_1, \hat{t}_2)$	Covariance entre deux estimateurs $\hat{t}_1$ et $\hat{t}_2$
$u \tilde{\sim} \mathrm{Unif}(0,1)$	Réalisation d'une variable aléatoire $\mathcal{U}$ distribuée selon la loi uniforme continue sur $]0;1[$
$z_{1-\alpha/2}$	Quantile d'ordre $(1-\alpha/2)$ d'une loi normale centrée et réduite

Table des matières

Chapitre 1

Concepts clés de la théorie des sondages

Les enquêtes par sondage permettent de tirer des conclusions sur une population d'individus à partir de l'observation d'une partie de celle-ci (un échantillon). Elles jouent un rôle essentiel dans de nombreux domaines (sciences sociales, économie, écologie, santé publique ou marketing) lorsque l'on cherche à décrire des phénomènes à partir de données recueillies auprès d'un échantillon. Les méthodes de sondages considérées dans cet ouvrage concernent la mise en œuvre et l'analyse d'enquêtes basées sur un échantillon d'individus tirés *aléatoirement* dans une population finie. C'est le cas en particulier des enquêtes de la statistique officielle réalisées par les offices nationaux comme l'Insee (Institut national de la statistique et des études économiques) en France, l'OFS (Office fédéral de la statistique) en Suisse, ou Statistique Canada au Canada, ou plus généralement par les grands instituts de statistique publique comme l'Ined (Institut national des études démographiques), l'Inserm (Institut national de la santé et de la recherche médicale) ou Santé publique en France. Ces enquêtes couvrent une grande variété de thèmes, tels que l'emploi (pour évaluer le taux de chômage par exemple), les entreprises (pour connaître leur chiffre d'affaires moyen), la santé (pour mesurer la prévalence des pathologies), l'éducation (pour connaître les effectifs d'étudiants). Elles s'étendent également à l'occupation du territoire (pour analyser l'usage des sols), l'inventaire forestier (pour évaluer le volume total de bois), l'élevage (pour quantifier les effectifs des cheptels) ou l'environnement (à travers des études sur l'exposition à la pollution ou l'étude de la biodiversité).

Néanmoins, l'utilisation des enquêtes par sondage dépasse largement la sphère de la statistique publique. Elles sont aussi beaucoup utilisées par des organismes privés, par exemple pour mesurer des indices médiatiques d'audience TV ou radio (Médiamétrie en France), ou pour mesurer le trafic postal des objets tels que les lettres ou les colis (La Poste en France). Les enquêtes par sondage sont aussi massivement utilisées pour connaître l'opinion publique sur diverses questions comme

les intentions de vote lors des élections présidentielles.
Au-delà de ces exemples traditionnels, ces dernières années ont vu l'émergence de nouvelles applications des sondages, en particulier dans le contexte des données massives (*big data*). Le recours aux appareils connectés, tels que les capteurs automatiques ou les smartphones, a engendré la disponibilité de bases de données très volumineuses. Les techniques de sondage permettent d'obtenir des estimations de qualité pour des quantités d'intérêt à des coûts raisonnables, tout en évitant de stocker l'intégralité du jeu de données. Ainsi, il est possible de sélectionner un sous-ensemble de capteurs intelligents afin d'analyser les schémas de consommation d'énergie pour un réseau électrique. Un deuxième exemple consiste à sélectionner aléatoirement un sous-ensemble de publications sur des réseaux sociaux pour analyser le sentiment du public ou identifier des sujets tendance. Un troisième exemple concerne l'analyse de données massives de transactions de ventes en ligne. Au lieu de traiter l'ensemble des transactions, on peut prendre un échantillon aléatoire de transactions pour analyser le comportement d'achat des clients ou identifier les produits populaires.
La réalisation d'une enquête par sondage est un processus complexe qui implique plusieurs étapes cruciales, succinctement résumées comme suit :

1. Définir précisément les objectifs de l'enquête : quelles informations souhaite-t-on obtenir ? Sur quelles unités ? Pour quelle période ? Les réponses à ces questions permettent de définir la population concernée par l'enquête ainsi que les quantités à estimer.

2. Procéder au tirage de l'échantillon d'unités (individus), qu'il soit probabiliste ou non probabiliste.

3. Concevoir le questionnaire et le formulaire de réponse, collecter les données auprès des unités échantillonnées, effectuer la saisie, le codage et la vérification des données collectées.

4. Une fois les données récupérées et nettoyées, les étapes finales consistent à traiter les valeurs manquantes et à estimer les quantités d'intérêt, idéalement en fournissant la précision des estimateurs via des intervalles de confiance associés.

À chacune de ces étapes, la capacité à atteindre les objectifs désirés est étroitement liée au budget disponible. Comme mentionné au point 2 ci-dessus, les individus peuvent être sélectionnés de façon probabiliste ou non. La théorie développée dans ce livre concerne les enquêtes par sondage utilisant des méthodes probabilistes pour le tirage aléatoire d'un échantillon d'individus. Nous nous concentrons principalement sur le cadre aléatoire défini par le *plan de sondage*, lequel détermine la probabilité d'inclusion de chaque individu dans un échantillon et permet de calculer les poids de sondage associés. Ces poids sont ensuite utilisés pour estimer les paramètres d'intérêt, de manière ponctuelle et par intervalles de confiance. La théorie des sondages, étayée par la théorie des probabilités, va permettre d'inférer, de manière rigoureuse, les résultats obtenus sur un échantillon tiré aléatoirement à l'ensemble de la population. À l'opposé, les enquêtes par sondage non probabiliste, telles que la méthode de la boule de neige (« snowball sampling »), consistent à

sélectionner les individus en fonction de critères plutôt subjectifs.
Pour plus de détails sur les différentes étapes d'une enquête par sondage, le lecteur pourra se référer au livre Valliant *et al.* (2013), ainsi qu'aux ouvrages Ardilly (2006) pour les enquêtes menées par l'Insee, Lohr (2010, 2021) pour les enquêtes américaines, Benedetti *et al.* (2015) pour les enquêtes agricoles.
Dans ce chapitre, nous introduisons les principes fondamentaux de la théorie des sondages. Nous commençons par définir plusieurs notions de base telles que population cible, base de sondage, variables et paramètres d'intérêt, ainsi que la notion d'estimateur de ces paramètres. Nous définissons aussi des concepts plus avancés tels que stratégie d'échantillonnage, information auxiliaire et domaine. Ces notions sont illustrées par plusieurs exemples concrets, issus notamment de la statistique publique et d'enquêtes européennes. Nous abordons ensuite, dans un cadre général, la formalisation du plan de sondage probabiliste ainsi que les fondements de l'estimation. Une attention particulière est portée à l'estimateur de Horvitz–Thompson d'un total ou d'une moyenne, avec l'étude de son biais, de sa variance et de l'estimation de cette dernière. L'estimateur de Hansen-Hurwitz est également présenté dans le cadre d'un tirage avec remise. Enfin, ce chapitre introduit trois jeux de données qui servent, dans la suite de l'ouvrage, à illustrer les fondements théoriques par des applications concrètes avec R.

1.1 Quelques principes de base en population finie

Dans une enquête, l'objectif est de mesurer des informations sur une population finie d'intérêt ou *population cible*, notée U, qui doit être définie précisément. Un élément de la population cible s'appelle *individu* ou unité statistique. La population considérée peut être composée d'humains, d'animaux, de végétaux, d'entreprises ou d'objets, selon le contexte de l'enquête (voir les exemples ci-après). La terminologie *champ de l'enquête* désigne aussi parfois l'ensemble des individus de la population cible. Lorsque l'ensemble des individus de la population cible est accessible, l'enquête prend la forme d'un *recensement*. Cependant, comme il est généralement difficile, voire impossible ou trop coûteux, de réaliser un recensement, on opte plutôt pour le tirage aléatoire d'un échantillon d'individus, c'est-à-dire d'une partie de la population. Les informations ou *estimations* sur des quantités inconnues de la population sont alors déduites de cet échantillon. Le caractère aléatoire du tirage des individus permet d'obtenir des propriétés mathématiques des méthodes d'estimation et d'*inférer* rigoureusement les résultats obtenus sur l'échantillon à l'ensemble de la population.
Pour tirer aléatoirement un échantillon d'individus, il est nécessaire de disposer d'une liste exhaustive et sans double compte des individus qui forment la population cible et que l'on souhaite échantillonner. On appelle *base de sondage* la liste des individus constituant la population cible et, si possible, complétée par des informations connues concernant ces individus. Il n'est pas toujours simple de définir une population cible et il est souvent impossible de disposer d'une base de sondage de qualité parfaite. On parle de *défaut de couverture* lorsque la base de sondage et

la population cible diffèrent. Nous avons volontairement simplifié la présentation de la notion de base de sondage et le lecteur pourra se référer par exemple à Ardilly (2006) ou Lohr (2010) pour plus de détails. Le répertoire *SIRENE*[1] (le Système national d'Identification et du Répertoire des ENtreprises et de leurs Etablissements), par exemple, est utilisé pour créer la base de sondage pour les enquêtes auprès des entreprises françaises qui sont indexées selon leur numéro siret/siren. Ce répertoire, mis à jour quotidiennement, contient de nombreuses informations concernant les entreprises, telles que leur localisation, activité, tranche d'effectif, date de création, etc. Un autre exemple de base de sondage est le répertoire des 25 millions d'adresses françaises, appelée *Base Adresse Nationale*[2] et utilisée notamment dans les enquêtes par sondage menées par La Poste pour mesurer le trafic postal.

Dans cet ouvrage, nous considérons que la population cible U peut être énumérée de la façon suivante :

$$U = \{1, \ldots, k, \ldots, N\},$$

où chaque individu est identifié de façon unique par son identifiant k. On écrit « $k = 1, \ldots, N$ » ou le plus souvent « $k \in U$ » pour faire varier l'indice k dans la population. La taille N finie de la population peut être connue ou inconnue. L'approche des sondages est parfois appelée approche *en population finie* car le caractère fini de la population est ce qui la différencie de la statistique classique.

Si une base de sondage existe et qu'elle couvre correctement la population cible, alors il est possible d'effectuer un échantillonnage *direct* des individus dans la base de sondage (chapitres 2 à 4). Le répertoire SIRENE constitue un exemple de base de sondage permettant un échantillonnage direct des entreprises (selon un tirage stratifié tel que présenté au chapitre 4). Sinon, on peut envisager de passer par une population intermédiaire qui, idéalement, regroupe les individus de la population cible et pour laquelle une base de sondage existe (par exemple dans le cadre d'un tirage en grappes ou à deux degrés, tels que présentés dans les chapitres 5 et 6). On effectue dans ce cas ce qu'on appelle un échantillonnage *indirect* des individus.

Pour une variable d'intérêt y, on note y_k la valeur de la variable pour l'individu $k \in U$,

$$t_{yU} = \sum_{k \in U} y_k,$$

le total de la variable y pour l'ensemble des individus de la population, et

$$\overline{y}_U = \frac{t_{yU}}{N}$$

sa moyenne. Le total et la moyenne pour une variable y sur une population U sont en général inconnus et sont considérés comme des *paramètres d'intérêt*. La variable y peut être quantitative ou qualitative, comme d'habitude en statistique. Si y est

1. Le répertoire SIRENE est mis à disposition gratuitement par l'Insee (« open data ») : https://www.insee.fr/fr/information/6675111.
2. La Base Adresse Nationale est également mise à disposition gratuitement (« open data ») : https://adresse.data.gouv.fr/donnees-nationales.

qualitative dichotomique, avec $y_k = 1$ si l'individu k a une certaine caractéristique A et 0 sinon, alors la moyenne de y dans la population U est la proportion P des individus ayant la caractéristique A dans la population U. Il existe d'autres paramètres d'intérêt que le total ou la moyenne. On peut, par exemple, s'intéresser à un ratio R de deux totaux t_{yU} et t_{xU} :

$$R = \frac{t_{yU}}{t_{xU}},$$

aux quantiles d'une variable quantitative y, aux coefficients d'une régression linéaire ou logistique, ou encore à l'indice de Gini. Dans le cadre du présent tome, nous considérons uniquement des paramètres de type totaux ou moyennes alors que le tome suivant aborde l'estimation de paramètres plus complexes tels que les ratios ou les paramètres d'une régression.
Examinons quelques exemples pour illustrer les notions précédentes.

Exemple 1.1
Considérons une course de vélo pour laquelle on s'intéresse à la proportion de cyclistes prenant un certain produit dopant. Dans cet exemple, la population cible est l'ensemble des cyclistes inscrits à la course et la base de sondage est la liste de ces cyclistes dans laquelle l'échantillon de cyclistes est sélectionné. La variable d'intérêt y prend la valeur $y_k = 1$ si le cycliste k est détecté positif au produit dopant et 0 sinon. Le paramètre d'intérêt est la proportion P de cyclistes ayant pris le produit dopant et correspond à la moyenne $\overline{y}_U$ de la variable y [3]. Pour estimer cette proportion, les organisateurs sélectionnent de manière aléatoire un échantillon de cyclistes avant le départ de la course et les cyclistes sélectionnés passent un contrôle anti-dopage. Les cyclistes peuvent également être choisis de manière aléatoire à l'arrivée de la course.

Exemple 1.2
Dans l'enquête française SIVIS (Système d'information et de vigilance sur la sécurité scolaire), on s'intéresse au taux moyen d'incidents graves dans les établissements scolaires. La population U est l'ensemble des établissements scolaires, la variable d'intérêt y est le nombre d'incidents graves pour 1 000 élèves au sein d'un établissement et le paramètre d'intérêt est la moyenne $\overline{y}_U$ [4]. Pour estimer ce paramètre, un échantillon d'établissements est sélectionné et les chefs d'établissements déclarent le nombre d'incidents survenus dans leur établissement (pour 1 000 élèves).

Exemple 1.3
Si on s'intéresse au nombre de chômeurs dans l'Union européenne avec une fréquence trimestrielle, la population cible U est constituée des individus en âge de

3. Pour plus de détails sur la collecte de ce type de données, voir le rapport de l'Agence mondiale anti-dopage : https://www.wada-ama.org/sites/default/files/2022-12/isti_2023_w_annex_k_final-clean.pdf.

4. Pour plus de détails sur l'enquête 2019-2020, voir https://shs.hal.science/halshs-03594464.

travailler (16 ans ou plus). Les 34 pays qui fournissent des données à Eurostat sur les forces de travail réalisent une enquête, telle que l'enquête Emploi en France mise en place par l'Insee, où la base de sondage est le fichier démographique des logements et des individus (Fidéli). La variable dichotomique d'intérêt y prend la valeur $y_k = 1$ si l'individu k est chômeur et 0 sinon. Le paramètre d'intérêt est le nombre total t_{yU} de chômeurs. Si on s'intéresse plutôt au taux de chômage, le paramètre d'intérêt est un ratio de deux totaux $R = t_{yU}/t_{xU}$ où x est la variable qui prend la valeur $x_k = 1$ si l'individu k fait partie de la population active et 0 sinon.

Exemple 1.4
L'enquête européenne LUCAS (Land Use and Cover Area frame Survey), également appelée « enquête statistique aréolaire sur l'utilisation et l'occupation des sols », se déroule tous les 3 ou 4 ans et vise à recueillir des informations sur l'utilisation (socio-économique) et la couverture (physique) des sols de l'Union européenne. Supposons que l'on souhaite estimer la surface totale des zones urbaines sur le territoire de l'Union européenne. Pour ce faire, on décrit le territoire au moyen d'une population cible composée de points régulièrement répartis sur l'ensemble du domaine européen. La variable d'intérêt associée est notée y_k, avec $y_k = 1$ si le point k est situé en zone urbaine et $y_k = 0$ sinon. Le paramètre d'intérêt, noté t_{yU}, correspond alors au nombre total de points situés en zone urbaine. Pour convertir ce total en surface, il suffit de le multiplier par l'unité de surface associée à chaque point, laquelle dépend de la densité de la grille utilisée (et donc du nombre total de points). L'échantillon est constitué de « points », qui sont en réalité des cercles d'un rayon de 1,5 mètre répartis sur le territoire européen et observés directement par les enquêteurs sur le terrain [5].

Exemple 1.5
L'Office fédéral de l'environnement (OFEV) a mis en place en 1995 un inventaire des prairies sèches en Suisse, visant à analyser la composition des différentes espèces végétales. Afin d'obtenir des informations fiables sur l'évolution des espèces, l'OFEV a lancé en 2011 un programme national et à long terme de suivi de la biodiversité des prairies suisses. Un plan de sondage complexe (à deux degrés, spatial et avec des probabilités inégales) est mis en pratique (Tillé & Ecker, 2014) pour sélectionner un échantillon des cercles de 10 m^2; la composition des espèces est collectée sur le terrain, à l'intérieur des cercles sélectionnées, tous les six ans.

Exemple 1.6
L'enquête EU-SILC (European Union Statistics on Income and Living Conditions) porte sur le revenu et les conditions de vie des ménages et des personnes de l'Union européenne. Chaque pays européen met en œuvre sa propre enquête avec des règles communes. En France, par exemple, il s'agit de l'enquête SRCV (Statistiques sur les ressources et les conditions de vie des ménages) qui est une enquête par panel où, chaque année, un nouvel échantillon de ménages répondant pour la première fois à l'enquête vient alimenter le panel et un échantillon dit « sortant » quitte

5. Pour plus de détails, voir https://ec.europa.eu/eurostat/fr/web/lucas/methodology.