وزارة التعليم العالي والبحث العلمي

Université Badji Mokhtar
Annaba

Badji Mokhtar University -
Annaba

جامعة باجي مختار

عنابة

Faculté des Sciences
Département de Mathématiques

THESE

Présenté en vue de l'obtention du diplôme de
Magister en Mathématiques

Option : Mathématiques financières

Enquête par sondage : Avantages et inconvénients

Par :
MANSOURI RAMDHANE

Sous la direction de :
Pr.Seddik Ameur Nacira

Devant le jury

PRESIDENT :	LASKRI. Y	Pr	U.B.M- ANNABA
RAPPORTEUR :	SEDDIK AMEUR.N	Pr	U.B.M- ANNABA
EXAMINATEUR :	DJELLAB. N	M.C	U.B.M- ANNABA
EXAMINATEUR :	CHADLI. A	M.C	U.B.M- ANNABA
INVITE :	ZEGHDOUDI.H	M.C	U.B.M- ANNABA

Année : 2011

Table des matières

1 Historique des sondages **3**

1.1 Introduction 3
1.2 Les prérequis sociologiques 4
1.3 Les conditions d'émergence au plan statistique : 9
1.4 Les prérequis techniques et technologiques 10
1.4.1 -Les ordinateurs : 10
1.4.2 -Le téléphone : 11
1.5 Conclusion : 11

2 Préliminaires **12**

2.1 Définitions : 12
2.2 Principes généraux des sondages : 13
2.3 La loi normale ou loi de gauss-Laplace : 14
2.4 Théorème de la limite centrale : 14
2.4.1 Définition : 14
2.4.2 Théorème de la limite centrale : 14
2.5 Plan général d'un sondage : 15
2.6 L'estimation : 16
2.7 Qualité d'un estimateur : 16
2.8 Méthodes de construction des estimateurs : 17
2.8.1 La méthode du maximum de vraisemblance : 17
2.8.2 La méthode des moments : 18
2.9 Estimation ponctuelle : 19
2.9.1 Estimation d'une moyenne : 19
2.9.2 Estimation d'une variance : 20
2.9.3 Estimation d'une proportion : 22
2.10 Estimation par intervalle de confiance : 22
2.10.1 Estimation d'une proportion : 23

2.10.2 Estimation d'une moyenne : 24
2.10.3 Estimation d'une variance : 25

3 Elaboration d'un questionnaire 27
3.1 Introduction : 27
3.2 L'utilisation de sondages : 28
3.3 Les étapes de la mise en œuvre d'une enquête par sondage : . 29
3.4 Elaboration d'un questionnaire : 31
3.5 Coût de l'enquête : 32
3.6 Plan de sondage : 33
3.6.1 sélection d'un échantillon : 33
3.6.2 Construction des échantillons : 33
3.7 Différents types d'erreurs rencontrées dans les enquêtes par sondage : 34
3.7.1 Erreur d'échantillonnage : 34
3.7.2 Erreur d'observation : 35
3.7.3 Défaut de couverture et non-réponse : 35
3.8 Traitement des non-réponses : 37
3.8.1 Le problème de non réponse : 37
3.8.2 Le traitement de la non-réponse : 38
3.9 Avantages et inconvénients d'une enquête par sondage : . . . 39
3.9.1 Avantages des enquêtes par sondage : 39
3.9.2 Inconvénients et limites des enquêtes par sondage : . . 40
3.10 Réalisation d'une enquête par sondage : 40
3.10.1 La méthode de tirage : 40
3.10.2 l'expression de l'estimateur : 41
3.11 Conclusion : 41

4 Méthodes des sondages : 42

I Les Sondages probabilistes : 43
4.1 Les Sondages probabilistes simples : 44
4.1.1 Le sondage aléatoire simple : 44
4.1.2 Le sondage systématique : 47
4.2 Les sondages probabilistes complexes : 48
4.2.1 Le sondage stratifié : 48
4.2.2 Le sondage par grappes : 55

4.2.3 Sondage par « Redressement » : 60
4.2.4 Le sondage à plusieurs degrés : 62
4.2.5 Le sondage à plusieurs phases : 64

II Les sondages non probabilistes : 65
4.3 Introduction : 66
4.4 Avantages et inconvénients des méthodes empiriques : 66
4.5 Différentes méthodes de sondages non probabilistes : 67
4.5.1 Sondage de commodité (à l'aveuglette) : 67
4.5.2 Sondage au jugé : 67
4.5.3 Sondage volontaire : 67
4.5.4 Sondage par la méthode de quotas : 67
4.5.5 Méthode des itinéraires : 68
4.5.6 Méthode des unités types : 69
4.6 Conclusion : 69

Résumé

Ce mémoire porte essentiellement sur l'étude des différentes méthodes d'enquête par sondage.

Dans le but d'optimiser l'estimation des paramètres inconnus de la population, l'échantillon doit être choisi de manière adéquate : représentativité et précision.

Les conditions de l'échantillonnage diffèrent d'une situation à une autre ; c'est pour cela que nous avons essayé de décrire dans ce travail certaines méthodes de sondage, ainsi que leurs avantages et leurs inconvénients.

L'utilisation de nouvelles technologies ont contribué à minimiser et à mieux maitriser les divers types d'erreurs pouvant affecter la qualité des résultats d'enquêtes.

De nombreuses avancées ont eu lieu également dans l'élaboration des questionnaires, les techniques d'estimation ou le calcul de la précision des résultats d'enquêtes.

Abstract

This thesis focuses on the study of different methods of survey.

In order to optimize the estimation of unknown parameters of the population, the sample must be chosen appropriately : representativeness and accuracy.

The conditions of sampling differ from one situation to another that is why we tried to describe some work in these survey methods and their advantages and disadvantages.

The use of new technologies helped to minimize and better control the types of errors that can affect the quality of survey results.

Many advances have taken place also in developing surveys, techniques for estimating or calculating the accuracy of survey results.

Remerciements

Au terme de ce travail le fruit de ma formation, le plus grand merci tout d'abord revient à ***«ALLAH »*** *le tout puissant pour son aide,pour le courage, la patience et la confiance en soi qu'il m'a donné pour surmonter toutes les difficultés durant mes études ainsi que l'endurance pour terminer ce mémoire.*

Je remercie infiniment mes parents , ma famille,et toutes mes connaissances à Alger.

Je tiens à adresser mes remerciements les plus chaleureux et ma profonde gratitude à ma chére enseignante et mon encadreur Madame ***Seddik Ameur Nacira*** *,Professeur à l'université d'Annaba, pour m'avoir proposé le sujet de ce mémoire,vraiment les mots ne suffiront jamais pour la remercier , elle toujours m'a consideré comme son fils,c'est grâce à son soutien moral, ses orientations, ses conseils et ses encouragements que j'ai pu mener à bien ce travail.*

Mes remerciements vont également à ***Madame Laskri.Y,*** *Professeur à l'université d'Annaba, pour avoir bien voulu me faire le grand honneur d'accepter de présider le jury.*

De même je remercie vivement Madame ***Djellab.N,****M.C à l'université d'Annaba et Madame* ***Chadli.A,****M.C à l'université d'Annaba,ainsi que Monsieur* ***Zeghdoudi.H****,M.C à l'université d'Annaba, pour l'honneur qu'ils m'ont fait de bien vouloir accepter de faire partie du jury.*

Je remercie beaucoup tous les gens qui m'ont appris le sens de la vie et les clefs du succés ;ainsi que l'art de communication illimitée.

Je ne saurai oublier toutes les personnes ,les amis et les collègues que je ne pourrai citer et qui de près ou de loin m'ont ont apporté leur aide ,leur soutien et leurs encouragements et qui ont cru à mes compétences et mes ambitions .Aussi je remercie tous ceux que j'estime et qui m'estiment.

Introduction

La théorie des sondages est une discipline des statistiques qui constitue un intérêt certain dans tous les domaines de la vie courante, surtout avec les nouvelles technologies et les données qui deviennent de plus en plus volumineuses. Les sondages connaissent un réel élan et permettent de mieux décrire et comprendre les questions d'actualité.

En tant qu'outil de recherche et de connaissance, le sondage étant une méthode d'enquête sur échantillon, son apparition est liée d'une part à l'intérêt, au besoin de faire des enquêtes auprès de la population (et non pas des seules élites) et d'autre part au développement des moyens nécessaires à la réalisation des enquêtes ; ces moyens sont à la fois théoriques et statistiques, techniques et technologiques. Ils relèvent également de la gestion des ressources humaines et matérielles.

L'enquête par sondage permet aux staticiens d'estimer les caractéristiques d'une population en observant directement une partie de celle-ci.

Dans ce travail nous faisons une étude détaillée sur la théorie des enquêtes par sondage :

Tout d'abord on introduit Dans le $1^{èr}$ chapitre un historique sur le sondage qui précise les conditions d'apparitions des sondages dans le monde.

Ensuite, dans le $2^{ème}$ chapitre nous exposons les concepts de base sur les enquêtes par sondage puis nous abordons des rappels sur la statistique mathématique ;ainsi nous traitons la théorie de l'estimation, la qualité d'un estimateur, les méthodes de construction des estimateurs et nous présentons l'estimation ponctuelle et par intervalle de confiance avec de plus amples détails.

Dans le $3^{ème}$ chapitre, on introduit une initiation aux différentes techniques des enquêtes par sondage et leurs limites, ainsi que leurs utilisations dans tous les domaines liés à la vie courante.

Dans le $4^{ème}$ chapitre nous abordons les différentes méthodes des sondages ; leurs avantages et leurs inconvénients.

Ce chapitre est subdivisé en deux parties :

-La premiére partie concernant les sondages probabilistes : simples et complexes.

-Quant à la deuxiéme partie,elle est consacrée aux diverses méthodes non probabilistes (empiriques) d'enquête par sondage et les conditions d'application de ces méthodes sur le terrain.

Enfin nous terminons par une conclusion générale pour donner une idée globale sur un domaine très vaste et intéressant dans les diverses études statistiques : la théorie des enquêtes par sondage.

Chapitre 1

Historique des sondages

1.1 Introduction

Il existe plusieurs histoires des sondages, selon les disciplines et les besoins sociaux qui ont contribué à leur développement. Il y a une histoire statistique, et ceci dans deux domaines, celui de l'échantillonnage – développement des probabilités, théorie de l'échantillonnage, élaboration de la pratique liée à l'échantillonnage, estimation des biais et des marges d'erreurs associés aux divers types d'échantillons - et celui de l'analyse des données – corrélations, régressions, analyses de classification, factorielle, tables de survie et tests statistiques associés.

Il y a une histoire socio-politique, celle de l'opinion publique, de sa mesure, de son "irruption" à part entière dans la vie politique ; c'est, en parallèle l'histoire de l'avènement de la démocratie mais aussi, celle de la légitimation de la "science sociale".

Il y a également une histoire économique, celle de la mise en marché, des stunts publicitaires, de la "market research" qui vise à faire vendre des produits, dont des journaux. Il y a une histoire technique, celle du développement des technologies de l'information, Il y a également des histoires, qui se différencient selon les pays.

L'évolution des sondages, la manière dont ils sont conduits, leur plus ou moins grande légitimité sociale et politique, leur plus ou moins grand développement varie selon les sociétés et leur culture.

1.2 Les prérequis sociologiques

Plusieurs conditions d'émergence sociologiques et sociopolitiques président à l'apparition des sondages.

Le besoin de connaissance d'abord objectives puis subjectives sur la population pour des raisons d'ordre financier (impôts) puis politiques (prévenir les révoltes) est certainement la première condition. Dès 1745, l'administration de Louis XV en France dans ses directives aux intendants de province fait semer le bruit qu'il y aura une augmentation du tiers des droits d'entrée et une levée de 2 milices par village , Il demande aux intendants de recueillir et de noter les réactions de la population.

A la fin du 18 ème siècle et durant tout le 19 ème siècle, la France en particulier est secouée par de nombreuses révolutions et révoltes de la population, de la Révolution française de 1789 à la Commune de Paris (1871) en passant par les révoltes de 1830 et 1848. Il apparaît alors de plus en plus important à la classe intellectuelle comme aux administrateurs et à la police de connaître la situation de la population et ses réactions. Quételet (1825) parlera de "par le calcul, maîtriser suffisamment les faits sociaux pour éviter de coûteuses révolutions".

Même si le besoin de connaissance est d'abord le fait des élites et des administrations, l'idée d'une nécessaire connaissance de la société fait son chemin. Karl Marx propose une enquête ouvrière en 1880. Il s'agit d'un questionnaire de 100 questions imprimé à 25.000 exemplaires et expédié "à toutes les sociétés ouvrières, à tous les groupes ou cercles socialistes ou démocratiques, à tous les journaux français et à toutes les personnes qui en font la demande.

Il est intéressant de regarder ce questionnaire pour voir le chemin parcouru depuis. Les notions de neutralité et de précision des questions, les notions d'échantillon probabiliste ou même simplement "représentatif " ne sont pas encore apparues.

Si on y pense bien, une autre condition est essentielle à la dissémination des sondages comme mode d'enquête et éventuellement de connaissance. Cette connaissance doit apparaître utile.

Avant l'avènement du modèle démocratique, lorsque les élites prenaient seules l'ensemble des décisions relatives à l'organisation de la société sans qu'il soit nécessaire de consulter la population, il était moins utile de connaître l'opinion du peuple. Avec l'avènement de la démocratie et du vote universel, la situation commence déjà à changer sérieusement, Si on veut se faire élire, on doit convaincre les gens de voter pour nous.

Les idées commencent à se débattre sur la place publique. Il faut savoir ce que le peuple pense de nos positions ,ce sont d'abord les journalistes qui serviront de courroie de transmission des réactions de la population mais leur connaissance peut apparaître biaisée ou à même d'entraîner des biais étant donné leur audience. Par ailleurs, pendant longtemps, l'identification partisane était relativement stable. Les changements d'allégeance étant moins fréquents, il devenait moins impératif de mesurer régulièrement l'opinion de la population.

Comme on peut le constater, du point de vue historique, l'accent est mis sur les sondages d'opinion qui ont constitué dès le début, la partie visible de l'iceberg. Il ne faut pas oublier que la méthodologie de sondage s'est répandue à bien d'autres domaines des sciences sociales, en particulier les domaines liés à la psychologie sociale et à la psychologie du travail, à la sociologie des organisations, à l'administration et à la gestion.

Ce n'est pas un hasard si les sondages d'opinion apparaissent et se répandent d'abord aux Etats-Unis, dès les débuts du 19ème siècle. La jeune démocratie américaine, comme on l'appelait à l'époque, a déjà bien intégré la règle d'"Un homme, un vote". Bien sûr, il s'agit à l'époque uniquement d'hommes blancs mais il demeure que les usages y sont moins fixés, que la noblesse est absente, que l'immigration amène sans cesse de nouveaux citoyens intéressés à recommencer à neuf, ouverts aux nouvelles idées.

Ce n'est pas un hasard non plus si les premiers sondages d'opinion apparaissent sous la forme de votes de paille (votes simulés) initiés par des journaux, entre autres le Harrisburg Pennsylvanien et le Raleigh Star. Lors de l'élection de 1824, ils utilisent diverses méthodes pour tenter de connaître à l'avance les résultats du vote. Il y a des coupons à découper et à retourner dans les journaux ; on met des "urnes" à la sortie des bureaux de vote ; on demande aux journalistes d'interroger les passants.

Déjà à cette époque, ces méthodes sont utilisées pour promouvoir les journaux ainsi que les idées des propriétaires de journaux. Il n'est évidemment pas encore question d'échantillon.

En 1896, le Chicago Tribune organise un scrutin anticipatif auprès d'un électeur sur huit tiré au hasard. Les résultats de ce premier "sondage" s'approchent à 0,4% du vote effectif. Cet événement est d'autant plus exceptionnel que la notion même d'échantillon tiré au hasard n'est pas encore acceptée y compris dans les milieux académiques.

Ce n'est finalement qu'en 1935 que les premiers instituts de sondage seront créés aux Etats- Unis, Gallup (par George Gallup), puis Fortune Surveys (par Elmo Roper) et Crossley Poll (par Crossley). L'influence de Gallup sera telle qu'en France on utilisera le terme de gallup pour parler d'un sondage au début de leur apparition. Il est évident dès le départ que ces firmes ne peuvent vivre en faisant uniquement des sondages électoraux, Elles ont d'ailleurs débuté par des études de marketing (le pain et le beurre de la plupart des firmes de sondage).

Enfin, ces premiers sondages n'ont rien à voir avec les "règles de l'art"qui sont idéalement utilisées maintenant. Les sondeurs utilisent comme interviewers, des amis, des maîtres d'école, des voyageurs de commerce qui interrogent les gens sur leur lieu de travail, dans les commerces ou sur la rue.

En France, Jean Stoetzel, un psychologue social de formation sociologique qui enseigne à l'époque dans un lycée, tente d'introduire le sondage suite à une rencontre avec George Gallup. Il fondera l'Institut français d'opinion publique (IFOP). A la même époque, Alfred Max fonde le Geop (Groupe d'étude de l'opinion publique). En France, l'accent est donc dès le départ mis sur l'opinion publique. Stoetzel a 25 ans à l'époque et Max 28 ans. Ils sont fortement à contre courant dans la culture française de l'époque.

Il est très intéressant de lire sur cette question les textes de Loïc Blondiaux. Il montre que, contrairement à ce qui se passe aux Etats-Unis à la même époque, les premiers sondeurs français rencontrent une forte résistance des milieux académiques et intellectuels en général. L'idée que l'opinion du simple citoyen puisse avoir une importance dans le débat public n'est pas acceptée d'emblée et menace le pouvoir des médias dont le rôle dans l'"opinion publique"est très important.

La résistance existe également du côté de la population elle-même qui ne se prête pas à la question aussi aisément qu'aux Etats-Unis. C'est l'avènement du Front populaire en 1936 qui donnera l'impulsion aux grandes enquêtes sociales, impulsion aussitôt freinée par la Deuxième guerre mondiale. Ce n'est qu'après la guerre que l'IFOP reprend le flambeau et se donne pour mandat de recueillir les attitudes, les opinions et les besoins de la population. Le fait que Jean Stoetzel est un psychologue social explique sans doute qu'il met plus l'accent sur les attitudes que les autres sondeurs de l'époque.

Par ailleurs, en France comme aux Etats-Unis, l'entreprise est au début très artisanal. Les interviewers sont des bénévoles. Ils ne reçoivent pas de formation et il n'est pas clair qu'ils reçoivent des instructions sur les personnes à interviewer.

Quant à la situation canadienne et québécoise, elle est au départ fortement liée à la situation américaine. Les premières firmes de sondage qui apparaissent sont, en 1932, Réalités Canadiennes puis en 1941, la filiale de Gallup. Elles semblent concentrer leurs efforts sur les études de marché. On retrace également la présence à Montréal dès 1935 de firmes d'études de marché qui sondent l'ensemble du Canada. (Ref : Robinson, The Measure of democracy).

Au Québec, le premier sondage d'opinion dans le cadre d'une élection est réalisé en 1959 pour le Parti Libéral par le groupe de recherche sociale avec la collaboration de Maurice Pinard. Le sondage prédit la victoire de l'Union Nationale et ce sont les Libéraux (Jean Lesage) qui l'emportent. En 1962, le même groupe prédira correctement la victoire de Libéraux.

L'histoire des sondages au Québec est émaillée d'"événements"qui expliquent en partie pourquoi les sondages politiques semblent y prendre une place plus grande qu'ailleurs au Canada. En 1966, La Société de Mathématique appliquée réalise un sondage qui prédit la victoire des Libéraux ; l'Union Nationale l'emporte avec moins de votes mais plus de sièges.

En 1976, la firme IQOP prédit la victoire du Parti Libéral alors que CROP (dirigé par Claude Corbeil) prédit correctement la première victoire du Parti Québécois. L'arbitrage est proposé mais non accepté. Les résultats trancheront en faveur de CROP.

Une situation similaire (Goldfarb contre CROP) se produit l'année suivante. Suite à ces événements, un comité conjoint SCSP-ACSALF) sera formé pour élaborer des normes sur la conduite des sondages et la publication des informations méthodologiques (voir CODEX pour le rapport de ce comité), suite aux recommandations de ce comité, les médias utiliseront des normes plus strictes dans les compte-rendus des sondages.

Entre temps, le Centre de sondage de l'Université de Montréal a été créé en 1969 suite à une contribution de la Fondation Doner. Le Centre réalisait entre autres la partie "Québec"des sondages réalisés par l'Université York à Toronto. Son initiateur fut Jacques Dofny (du département de sociologie) et ses directeurs Jacques Brazeau, Serge Carlos et finalement Victor Tremblay.

Le Centre ne faisant pas ses frais, il a été fermé en 1986. Entretemps, il aura énormément contribué à la formation d'une génération de sondeurs. Il aura également contribué à initier et à maintenir des normes méthodologiques strictes.

En 1980, la firme IQOP prédira la victoire du OUI (avec 52%) au référendum et ce deux jours avant le scrutin. Le NON l'emportera par 60%. La firme IQOP refuse de donner des informations sur sa méthodologie et de donner accès à ses données. Cela contribuera fort probablement à la mort d'IQOP dont la crédibilité avait été fortement entachée.

Lors du référendum de 1995, la moyenne de l'ensemble des sondages donnait 50,5% au NON, c'est-à-dire exactement le résultat qu'il obtiendra. Toutefois, ces résultats sont obtenus en accordant 75% des répondants discrets au NON. En 1998, la très grande majorité des sondages publiés pendant la campagne accorde une nette majorité au Parti Québécois alors que le Parti Libéral obtiendra finalement plus de votes que le Parti Québécois.

Les recherches menées suite à cette élection (Durand, Blais et Vachon avec la collaboration de CROP et CREATEC) amènent à penser qu'il faut encore raffiner la méthode, entre autres pour ce qui est de l'échantillonnage.

Les sondages politiques sont très présents dans les médias au Québec, surtout en période électorale lors d'élections générales et de référendums. L'importance des enjeux référendaires et la répartition presque égale des pour et des contre a contribué à leur donner une très grande importance.

Il ne faut toutefois pas oublier que la très grande majorité des sondages sont faits dans le cadre d'études de marché et que, même pour les sondages en sciences sociales, les sondages électoraux constituent une infime partie de l'ensemble des sondages réalisés.

En résumé, l'avènement de la démocratie, c'est-à-dire l'acceptation du fait que tous les votes s'équivalent et par conséquent toutes les opinions est à la base même de l'apparition des sondages. La fragilisation des identifications partisanes contribue fortement à l'importance que prennent les sondages dans des sociétés données à des époques données.

Cette fragilisation fait que les élus ne prennent plus leur siège pour acquis et tenteront de se mettre au diapason de leurs électeurs. Le fait que l'identification partisane ait été plus fragile aux Etats-Unis étant donnée la forte immigration et le mode de représentation politique a sans doute contribué à en faire le "berceau" des sondages tels qu'on les connaît.

Si ces conditions expliquent l'avènement et la dissémination des sondages d'un point de vue socio-politique, elles ne sont toutefois pas suffisantes, des conditions théoriques et techniques étant également essentielles au développement et à la fiabilité de l'"outil".

1.3 Les conditions d'émergence au plan statistique :

D'un point de vue théorique, le nerf de la guerre réside dans l'élaboration de la théorie des probabilités et du théorème central limite. Le développement de la théorie des probabilités débute à la fin du 18 ème siècle (Pascal, Fermat, Bernouilli). Les premières inférences statistiques sont toutefois le fait de l'Ecole anglaise (Graunt et Petty) qui utilise le terme d'arithmétique politique.

Les premiers essais se font sur l'estimation du nombre d'habitants à Londres, Petty utilisera plusieurs informations qu'il combinera : le nombre de foyers et le nombre moyen de personnes par foyer, le nombre de morts dans un an et le taux de mortalité, le nombre de naissance et le taux de natalité.

Fin 18ème, les techniques d'extrapolation se répandent, visant à remplacer "les coûteux et monstrueux dénombrements". En 1783, Laplace élabore le "théorème central limite"qui constitue la base même de l'inférence statistique :

"Plus la taille des échantillons est grande, plus la distribution d'échantillonnage des moyennes s'approche d'une distribution normale, et ceci quelle que soit la forme de la distribution".

Il s'agit d'une percée extrêmement importante. Elle permet de calculer l'erreur à laquelle on peut s'attendre lorsque l'on tire un échantillon d'une certaine taille plutôt que de prendre l'ensemble de la population. Par la suite, Quételet (1796-1874) sera le premier à concevoir que la statistique peut être fondée sur la probabilité.

La méthodologie de sondage étant basée sur le fait que l'on interroge un échantillon d'une population, les bases théoriques nécessaires à l'utilisation de l'échantillon sont donc toutes connues dès la fin du 18 ème siècle. Toutefois, pour des raisons sans doute liées à l'application de la théorie, le développement s'arrête au 19 ^{ème} siècle.

L'échantillonnage recule au profit de l'exhaustif mais paradoxalement, on a tendance à accepter de généraliser à partir d'échantillons clairement non représentatifs, c'est-à-dire constitués de personnes ayant des caractéristiques spécifiques.

Il s'agit d'une certaine manière d'une période de transition où les avancées théoriques ne se sont pas encore traduites au plan des conséquences pratiques,

c'est aussi une époque où on recueille énormément de données que l'on

analyse. C'est ainsi que l'on commence à découvrir des régularités : taux de mortalité et variation de ce taux en fonction de l'âge, taux de criminalité, on en conclura que ce sont là des questions collectives plutôt que relevant des individus et que donc, la société peut et doit intervenir (Crépel, 1999).

L'acceptation du recours à l'échantillon plutôt qu'à l'interrogation de toute la population est relativement récente. Lorsque le statisticien norvégien Kiaer proposera en 1895 au congrès de l'Association internationale de Statistique, que le recours à l'échantillon peut être utilisé pour recueillir des statistiques relativement fiables sur une population, il provoquera un débat houleux et sa proposition sera rejetée d'emblée. On alla jusqu'à déclarer que jamais une telle chose ne se produirait (Droesbeke et coll. 1987).

Pourtant, à peine sept ans plus tard, en 1903, au congrès de la même association, le principe de l'échantillonnage est accepté. Il reste alors à déterminer comment on doit s'y prendre pour faire un échantillon. Neyman travaillera plus précisément à cette question. Au 20ème siècle, les analyses et les tests statistiques se développent.

Au plan des analyses, on verra l'apparition des corrélations et régressions, de l'analyse de classification, de l'analyse de variance et plus tard de l'analyse factorielle. Au plan des tests, les test F de Fischer, t de Student, et le test de Pearson apparaîtront.

1.4 Les prérequis techniques et technologiques

1.4.1 -Les ordinateurs :

Plusieurs avancées techniques ont permis de rendre les sondages utiles.

En ce sens qu'elles permettront de traiter plus d'informations de façon plus rapide et moins chère et donc utile pour les décideurs. L'apparition de l'informatique est évidemment de toute première importance puisqu'elle permettra de traiter l'information et d'appliquer les tests statistiques.

Toutefois, au début, l'accès aux ordinateurs est réservé à peu de personnes.Les ordinateurs sont très chers et peu accessibles ; les programmes sont difficiles à apprendre. Les premiers utilisateurs des ordinateurs y auront souvent accès seulement de nuit et de week- end et un simple tableau croisé pourra coûter jusqu'à 80$ au début des années 70 au Québec.

L'apparition de la micro-informatique dans les années 80 permettra de démocratiser l'accès au traitement des informations. Ce développement s'est accéléré par la suite pour englober non plus seulement le traitement de l'information mais également la collecte de l'information. La plupart des sondeurs utilisent maintenant une collecte de données assistée par ordinateur.

Ces développements permettent de rendre l'accès à l'information produite à la fois de plus en plus rapide et de moins en moins chère, les sondages apparaîtront donc de plus en plus utiles et seront de plus en plus répandus.

1.4.2 -Le téléphone :

La dissémination du téléphone comme mode de communication est aussi essentielle aux sondages tels que nous les connaissons, du moins en Amérique du Nord puisqu'elle permet de faire fi des barrières géographiques à la collecte des informations. Ce n'est que lorsque la presque totalité des ménages ont eu accès au téléphone et s'est appropriée ce mode de communication que l'on a pu constituer des échantillons représentant vraiment l'ensemble de la population.

Le développement de nouveaux modes de communication - téléphones portables associés aux individus plutôt qu'aux ménages, Internet - pose des défis majeurs à l'industrie du sondage en ce qui a trait à l'échantillonnage.

1.5 Conclusion :

La pénétration du téléphone dans les ménages et le coût des communications, l'importance des origines "marketing" plutôt que "socio-politique" du développement de l'industrie, la présence et le rôle joué par les firmes académiques, les préoccupations des pionniers et leur formation, tous ces éléments ont joué un rôle dans le développement des sondages dans les divers pays.

Chapitre 2

Préliminaires

Dans ce chapitre, nous donnons les définitions et les préliminaires nécessaires qui nous permettent de mener notre étude.

2.1 Définitions :

- *L' enquête :* est une activité organisée et méthodique de collecte de données sur des caractéristiques d'intérêt d'une partie ou de la totalité des unités d'une population à l'aide de concepts, de méthodes et de procédures biens définis.

- *Le recensement :* tous les individus de la population sont mesurés.

- *L'échantillonnage :* c'est l'ensemble des opérateurs amenant à définir la population visée par l'enquête, puis à choisir l'échantillon.

- *Variable statistique :* c'est ce que peut être mesuré sur les mêmes individus statistiques, éléments d'une population.

- *Variable de contrôle :* ce sont les caractères retenus pour assurer la conformité de l'échantillon à l'ensemble de la population.

- *Un élément :* est un objet sur lequel on va mesurer les caractéristiques d'intérêt.

- *Les unités d'échantillonnage :* sont des entités disjointes dont l'union est égale à la population.

- *La population :* c'est l'ensemble des individus (unités statistiques) que l'on voudrait étudier et auxquels s'appliquent les résultats d'une enquête.

- *La population cible :* est une population sur laquelle l'information est requise.

- *La base de sondage :* donne les moyens d'identifier les unités d'échantillonnage de la population cible c'est-à-dire que la base de sondage est le répertoire ou fichier dans lequel sont consignées les unités de sondage de l'ensemble à étudier. C'est à partir de laquelle on va sélectionner les unités d'échantillonnage qui vont être enquêtées.

- *Un échantillon :* est un ensemble d'unités d'échantillonnage sélectionné à partir de la base de sondage.

2.2 Principes généraux des sondages :

Pour réaliser un sondage juste et adapté à la population visée, il faut choisir la technique d'échantillonnage adéquate et garder certains principes :

a) La planification : le premier aspect de la théorie de sondage est la planification de collecte des données, on distingue essentiellement deux types des méthodes de tirage soit les méthodes probabilistes ou les méthodes empiriques.

b) L'estimation : le deuxième aspect de la théorie des sondages est l'estimation de ce qu'on appelle les fonctions d'intérêt(moyennes, totaux , autres). La qualité d'un estimateur dépend presque toujours du problème à traiter. Introduire, dans les estimateurs, des données issues d'un recensement est une des spécifités de la théorie des sondages qui contribue grandement à sa complexité.

c)Traitement des erreurs de mesure : fait également partie des techniques d'estimation, il est d'ailleurs utile de décrire de manière explicite le modèle pour que ses hypothèses sous-jacents soient clairement enoncées.La modélisation de l'erreur de mesure introduit alors une deuxième source d'aléas dont il faut tenir compte à l'étape de l'estimation.

d) Les specifités de la théorie des sondages : L'étude des populations finies génère un certain nombre des difficultés très particulières .Dans le modèle général examiné, les unités sont identifiables sans ambigüité au moyen d'une étiquette. Cette propriété permet de repérer si une unité a été sélectionnée plusieurs fois dans l'échantillon, car l'étiquette est supposée connue et unique. L'identifiable des unités pose de nombreux problèmes théoriques très spécifiques comme la réduction par exhaustivité ou la méthode de maximum de vraisemblance, ne débouchent pas sur le choix de la procédure d'estimation.

2.3 La loi normale ou loi de gauss-Laplace :

Une variable aléatoire continue "Y" est distribuée selon une loi normale si sa densité de probabilité $f(y)$ est :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-m)^2/2\sigma^2}$$

$$Où : \begin{cases} m = E(y) \\ \sigma = écart - type \end{cases}$$

2.4 Théorème de la limite centrale :

2.4.1 Définition :

En probabilité, le théorème de la limite centrale est un ensemble de résultats sur la convergence faible d'une suite des variables aléatoires.Intuitivement, d'après ces résultats, toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire.

Le résultat le plus connu et le plus important est simplement appelé « théorème de la limite centrale » qui concerne une somme de variables aléatoires dont le nombre tend vers l'infini.

2.4.2 Théorème de la limite centrale :

Soit $Y_1, Y_2, ..$ un ensemble de variables aléatoires définies sur le même espace de probabilité, suivant la loi D et indépendantes. Supposons que l'espérance μ et l'écart- type σ de D existent et sont finis ($\sigma \neq 0$).

Considérons la somme $S_n = Y_1 + ... + Y_n$.Alors l'espérance de S_n est $n\mu$ et son écart type vaut $\frac{\sigma}{\sqrt{n}}$, de plus, pour parler de manière informelle, la loi de S_n tend vers la loi normale $N(n\mu, \sigma^2 n)$ quand n tend vers l'infini.

Afin de clarifier cette idée de convergence, nous allons poser

$$Z_n = \frac{S_n - n\mu}{\sigma/\sqrt{n}}$$

Alors la loi de Z_n converge vers la loi normale centrée réduite $N(0,1)$ lorsque n tend vers l'infini (il s'agit de la convergence en loi).Cela signifie

que si ϕ est la fonction de répartition de $N(0,1)$, alors pour tout réel z :

$$\lim_{n\to\infty} P(Z_n \leq z) = \phi(z)$$

De façon équivalente :

$$\lim_{n\to\infty} P(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} \leq z) = \phi(z)$$

Où

$$\overline{Y}_n = \frac{S_n}{n} = \frac{Y_1 + Y_2 + ... + Y_n}{n}$$

2.5 Plan général d'un sondage :

a) Le paramètre :

La fonction des Y_i qui nous intéresse est un paramètre, c'est-à-dire une grandeur fixée mais inconnue .Notons la :

$$\theta = f\,(Y_1, Y_2, ..., Y_N)$$

où N est la taille connue de la population. Pour le sondeur, elle représente la "vraie valeur" qu'il faut estimer. Dans le cas d'un total, par exemple :

$$f\,(Y_1, Y_2, ..., Y_N) = \sum_{i=1}^{N} Y_i$$

Dans le cas d'une moyenne :

$$f\,(Y_1, Y_2, ..., Y_N) = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Une proportion est une moyenne particulière.

b) l'échantillon :

L'information est collectée sur un échantillon de taille n tiré par une méthode appropriée. L'échantillon traditionnellement noté s ("sample"), doit être décrit par la totalité des identifiants des individus qu'il contient.

c) Les mesures des erreurs d'échantillonnage :

Un point important est la nature de l'aléa dans notre problème. L'aléa se situe exclusivement au niveau des identifiants des individus de l'échantillon, notre étude porte essentiellement sur les "dégâts" occasionnés par l'aléa.

2.6 L'estimation :

La théorie de l'estimation est l'étude des liens qui existent entre les caractéristiques connues des échantillons et celles qui correspondent de la population.

La notion de précision est très importante.Dans ce cas on détermine un intervalle dans lequel se trouve le paramètre à estimer avec une probabilité fixée.

En statistique mathématique, on cherche à estimer la valeur inconnue d'une statistique d'une population. On utilise une statistique d'un échantillon prélevé au hasard ou par d'autres méthodes aléatoires.

On peut donner une estimation : ponctuelle et par intervalle de confiance.

La qualité de l'estimation dépend de :

a) De la méthode de prélèvement de l'échantillon : quand l'échantillon est prélevé au hasard ou par une méthode aléatoire, l'estimation sera aléatoire, c'est-à-dire que les différentes valeurs possibles seront assorties de probabilités.

b) De l'information statistique : disponible dans la population d'où est extrait l'échantillon.

Définition :

Soient $y_1, ...y_i, ...y_n$ les n valeurs prises par la v.a.Y de loi $L(y, \theta)$ dans un échantillon de taille n prélevé dans la population mère.

On appelle estimateur de θ, et l'on note T_n, la fonction qui aux valeurs y_i de l'échantillon fait correspondre la valeur du paramètre θ. On note la valeur numérique de cette estimation par :

$$\widehat{\theta} = T_n(y_1, ...y_n)$$

Par définition, T_n est une fonction des réalisations d'une v.a.

2.7 Qualité d'un estimateur :

Définitions :

Le biais : Un estimateur T_n d'un paramètre θ est dit sans biais, si son espérance mathématique est la vraie valeur du paramètre :

$$E(T_n) = \theta \text{ ou } E(\widehat{\theta}_n) = \theta$$

On appelle biais d'un estimateur T_n de θ, l'écart : $biais(T_n) = E(T_n) - \theta$
Un estimateur T_n de θ est asymptotiquement sans biais

$$\Longleftrightarrow \lim_{n\to\infty} E(T_n) = \theta.$$

La convergence : Un estimateur T_n d'un paramètre θ est dit convergent s'il tend en probabilité vers θ.

$$T_n \to \theta \iff \lim_{n\to\infty} P(|T_n - \theta| > \varepsilon) = 0$$

Théorème : Un estimateur T_n d'un paramètre θ dont l'espérance mathématique tend vers θ et la variance tend vers 0 lorsque $n \to \infty$, est un estimateur convergent.
$E(T_n) = 0$ et $V(T_n) = 0$ pour $n \to \infty$ alors $T_n \to \theta$.
On en déduit l'intégralité de Bienaymé-Tchebychev :

$$P(|T_n - \theta| > \varepsilon) \leq \frac{V(T_n)}{\varepsilon^2}$$

Où :

$$P(|T_n - \theta| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

Définition d'un estimateur efficace :
La variance d'un estimateur sans biais représente sa précision. Pour tous les estimateurs (ayant même moyenne), il est possible de trouver celui dont la précision sera la meilleure, i.e. dont la variance sera la plus faible. On parle alors d'estimateurs à variance minimum.

Lorsqu'on compare deux estimateurs, on dira également que T_n est plus efficace que T_n^* si : $V(T_n) \leq V(T_n^*)$

Remarque :
Si deux estimateurs sans biais et convergents du paramètre θ, alors le plus précis ,donc le meilleur est celui qui possède la variance la plus faible.

2.8 Méthodes de construction des estimateurs :

2.8.1 La méthode du maximum de vraisemblance :

La fonction de vraisemblance : On peut aussi s'en servir pour construire un estimateur. Soit Y une variable aléatoire de densité de probabilité $f(y, \theta)$

connue analytiquement mais dont l'un des paramètres θ est inconnu (numériquement). Le problème consiste donc à construire une expression analytique en fonction des réalisations de cette variable dans un échantillon de taille n, permettant de trouver la valeur numérique la plus vraisemblable pour le paramètre θ.

Si $\{y_1, ..., y_n\}$ sont des réalisations indépendantes de la variable aléatoire Y,L'approche retenue consiste à chercher la valeur de θ qui rend le plus probable les réalisations que l'on vient d'obtenir. La probabilité d'apparition a priori de l'échantillon en question peut alors être caractérisée par le produit des probabilités d'apparition de chacune des réalisations (puisque celles-ci sont supposées indépendantes deux à deux).

La fonction de vraisemblance est notée $L(Y_1, Y_2, ..., Y_n; \theta)$ tel que :

$$L(Y_1, Y_2, ..., Y_n; \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

Définition : On appelle estimateur du maximum de vraisemblance d'un paramètre inconnu, la valeur qui rend maximum la fonction de vraisemblance L de θ sur la réalisation d'un échantillon.

$$\widehat{\theta} = \text{e.m.v} = \max_{\theta} L(Y_1, Y_2, ..., Y_n; \theta)$$

La recherche d'un estimateur e.m.v se fait directement par la recherche de maximum de L ou bien si L est deux fois dérivable par rapport à θ,donc l'e.m.v est solution du système suivant :

$$\begin{cases} \frac{\partial L}{\partial \theta} = 0 \\ \frac{\partial^2 L}{\partial^2 \theta} < 0 \end{cases}$$

Remarques :

-Cependant, on préfère remplacer ce problème par un problème équivalent par la log-vraisemblance puisque la fonction log est strictement croissante.

-Si $\widehat{\theta}_n$ e.m.v du paramètre θ alors $g(\widehat{\theta}_n)$ est un e.m.v pour toute fonction g.

2.8.2 La méthode des moments :

Dans le cas où le paramètre à estimer est $\theta = E_0(Y)$ la moyenne théorique, l'estimateur naturel est la moyenne empirique c'est-à-dire la moyenne de l'échantillon "Y_n" de même que pour σ^2.

Généralement si l'un des moments d'ordre $k, k \in N^*$,existe

$$E_\theta\left(Y^k\right) = m_k\left(y\right); \left(m_1 = E\left(Y\right)\right)$$

Nous allons chercher un estimateur par résolution de l'équation obtenue en égalant le moment théorique et le moment empirique :

$$m_k\left(y\right) = m_{k_n}\left(y\right); m_{k_n} = \frac{1}{n}\sum_{i=1}^{n} y_i^k$$

Et si cette solution est unique, elle sera appelée estimateur par la méthode des moments.Cette méthode intuitive ne justifie pas les propriétés de convergence des moments empiriques vers les moments théoriques.

2.9 Estimation ponctuelle :

Définition :
C'est la valeur observée dans l'échantillon, ou une valeur unique calculée à partir de la valeur observée dans l'échantillon, dans le voisinage de laquelle la valeur inconnue de la statistique qu'on cherche à estimer a une très grande probabilité de se trouver.

2.9.1 Estimation d'une moyenne :

Soit donc une variable aléatoire Y suivant une loi normale de moyenne μ inconnue et d'écart-type σ. On dispose d'un échantillon de n réalisations de cette variable aléatoire ,la moyenne empirique $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ est un estimateur de μ.

Qualité de l'estimateur :

$$E\left(\overline{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\right) = \frac{1}{n}.n.\mu = \mu$$

$E\left(\overline{Y}\right) = \mu$, alors $\overline{Y}$ est un estimateur sans biais de μ.

$$V\left(\overline{Y}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} Var \sum_{i=1}^{n}\left(Y_i\right) = \frac{1}{n^2}.n\sigma^2 = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} V\left(\overline{Y}\right) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

$\lim_{n\to\infty} V(\overline{Y}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$.Donc $\overline{Y}$ est un estimateur convergent et sans biais de μ.

2.9.2 Estimation d'une variance :

Nous n'aborderons que le cas de l'estimation de la variance σ^2 d'une v.a.Y normale de moyenne μ à partir d'un échantillon de n valeurs.

• Si μ est connue, la variance de la v.a.Y est :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu)^2$$

• Donc :

$$E\left(S_n^2\right) = \frac{1}{n} E\left[\sum_{i=1}^{n} (Y_i - \mu)^2\right] = \frac{1}{n}.\frac{1}{n^2}\left[\sum_{i=1}^{n} E\left((n-1)\, Y_i - \sum_{k=1,k\neq i}^{n} Y_k\right)^2\right]$$

En posant :

$$X_i = (n-1)\, Y_i - \sum_{k=1,k\neq i}^{n} Y_k$$

$$E(S_n^2) = \frac{1}{n^3}\sum_{i=1}^{n} \left(X_i^2\right)$$

• Calculons l'espérance de X_i :

$$E\left(X_i\right) = E\left[(n-1)\, Y_i - \sum_{k=1,k\neq i}^{n} Y_k\right] = (n-1)\, E\left(Y_i\right) - \sum_{k=1,k\neq i}^{n} E\left(Y_k\right)$$

$$E\left(X_i\right) = (n-1)\,\mu - (n-1)\,\mu = 0, E\left(X_i\right) = 0$$

• Calculons la variance de X_i :

$$V\left(X_i\right) = E\left(X_i^2\right) - \left[E\left(X_i\right)\right]^2 = E\left(X_i^2\right)$$

• Donc on obtient :Pour $i = \overline{1,n}$ et $\mu_i = E(Y_i)$

$$\begin{aligned} E\left(S_n^2\right) &= \frac{1}{n^3}\sum_{i=1}^{n} V(X_i) \\ V(X_i) &= Var\left[(n-1)Y_i - \sum_{k=1,k\neq i}^{n} Y_k\right] \\ &= (n-1)^2 V(Y_i) + (n-1)V(Y_k) \\ &= (n-1)^2\sigma^2 + (n-1)\sigma^2 = n(n-1)\sigma^2 \end{aligned}$$

• Alors :

$$E\left(S_n^2\right) = \frac{1}{n^3}\sum_{i=1}^{n} n(n-1)\sigma^2$$

$$E\left(S_n^2\right) = \frac{1}{n^3} n.n(n-1)\sigma^2 = \frac{(n-1)}{n}\sigma^2$$

$\Longrightarrow S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)^2$ n'est pas un estimateur sans biais, d'autre part :

$$\lim_{n\to\infty} V\left(S_n^2\right) = \lim_{n\to\infty}\frac{1}{n^2}V\left(\sum_{i=1}^{n}(Y_i-\mu)^2\right) = \lim_{n\to\infty}\frac{1}{n^2}nV\left[\left(Y_i-\overline{Y}\right)^2\right] = \lim_{n\to\infty}\frac{1}{n}V\left[(Y_i-\mu)^2\right] = 0$$

Donc S_n^2 est un estimateur convergent de σ^2

• Si μ est inconnue, la variance de la v.a.Y est :

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 = \frac{n}{n-1}S_n^2$$

D'après les calculs de l'espérance et de la variance de S_{n-1}^2,on obtient :

$$E\left(S_{n-1}^2\right) = \sigma^2$$

Donc S_{n-1}^2 est un estimateur sans biais, et :$V\left(S_{n-1}^2\right) = 0$, donc S_{n-1}^2 est un estimateur convergent.

2.9.3 Estimation d'une proportion :

On veut estimer la proportion p d'individus ayant ou non un caractère donné **A** dans une population de taille n. le nombre d'individus ayant un caractère donné **A** est np, on tire un $n-$*échantillon* $Y_1, Y_2, \ldots, Y_n$ de cette population. On considère :

$$F_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

F_n est un estimateur naturel de p (Fréquence de l'échantillon).

$$F_n \longmapsto N\left(p, \tfrac{pq}{n}\right), \text{ où } q = 1-p$$

Qualité de l'estimateur :

$$E\left(F_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right)\right) = \frac{1}{n}.n.p = p$$

(Y_i est une Bernoulli) donc F_n est un estimateur sans biais.

$$V\left(F_n\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(Y_i\right) = \frac{1}{n^2}n.V\left(Y_i\right) = \frac{1}{n^2}.npq = \frac{pq}{n}.$$

(Y_i est une Bernoulli), $\lim\limits_{n\to\infty} V\left(F_n\right) = \lim\limits_{n\to\infty} \frac{pq}{n} = 0$. D'où F_n est un estimateur convergent.

2.10 Estimation par intervalle de confiance :

Cette nouvelle approche est souvent préférée dans la pratique car elle introduit la notion d'incertitude. On cherche à déterminer l'intervalle $[a, b]$ centré sur la valeur numérique estimée du paramètre inconnu θ contenant la valeur vraie avec une probabilité α fixée a priori.

Cette probabilité permet de s'adapter aux exigences de l'application :

$$P\left[a < \theta < b\right] = 1 - \alpha$$

L'intervalle $[a, b]$est appelé intervalle de confiance et $1-\alpha$ est le coefficient de confiance($\alpha \in]0, 1[$).Une estimation par intervalle de confiance sera

d'autant meilleure que l'intervalle sera petit pour un coefficient de confiance grand.

La donnée de départ, outre l'échantillon, sera la connaissance de la loi de probabilité du paramètre à estimer. Comme il n'existe pas de résolution générale de ce problème, nous allons aborder successivement les cas les plus fréquents (estimation d'une proportion, d'une moyenne, d'une variance de la loi normale).

2.10.1 Estimation d'une proportion :

Soit une population dont les individus possèdent un caractère **A** avec une probabilité p.On cherche à déterminer cette probabilité inconnue en prélevant un échantillon de taille n dans cette population. On constate que y parmi les n individus possède le caractère **A**.

La proportion $f_n = \frac{y}{n}$ Approxime la valeur vraie p, mais avec quelle confiance ?

Soit , $F_n = \frac{y}{n}, F_n$ est une v.a. construite par la somme de n variables aléatoires et de même paramètre p. C'est donc, d'après le théorème central limite, une variable aléatoire dont la loi de probabilité tend vers une loi normale de moyenne p et d'écart-type$\sqrt{\frac{p(1-p)}{n}}$ Cette approximation est valable uniquement si la taille de l'échantillon est suffisamment grande (i.e. $n > 30$ en pratique).

Construisons l'intervalle de confiance autour de p :

$$P\left(|f_n - p| < t\right) = 1 - \alpha$$

Où α est le risque (a priori, on construit un intervalle symétrique). f_n est une réalisation d'une v.a :$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. Donc on peut par normalisation en centrage obtenir une nouvelle v.a.**U** :

$$u_\alpha = \frac{f_n - p}{\sqrt{\frac{p(1-p)}{n}}} \longmapsto N(0,1)$$

On en déduit donc l'intervalle de confiance sous la forme :

$$P\left[a < \theta < b\right] = P\left[f_n - u_\alpha\sqrt{\frac{p\left(1-p\right)}{n}} < p < f_n + u_\alpha\sqrt{\frac{p\left(1-p\right)}{n}}\right] = 1 - \alpha$$

La valeur $t = u_\alpha \sqrt{\frac{p(1-p)}{n}}$ est donc un résultat de calcul. La valeur de u_α sera lue sur une table de loi normale $N(0,1)$.

Il existe par ailleurs différentes manières pour approximer la valeur de p :

• Soit par la proportion f_n.

• Soit par majoration : en effet, quelque soit la valeur de p, le produit $p(1-p)$ est majoré par $\frac{1}{4}$.

2.10.2 Estimation d'une moyenne :

• **Cas des grands échantillons** $(n > 30)$

Théorème :

Quelle que soit la loi de $Y, U = \frac{\overline{Y}-\mu}{\frac{s}{\sqrt{n-1}}}$ suit sensiblement la loi normale centrée réduite.

Utilisation:

La valeur de u_α sera lue sur une table de loi normale $N(0,1)$ tel que :

$$P\left[-u_\alpha < U < u_\alpha\right] = 1 - \alpha$$

On en déduit comme intervalle de confiance de la moyenne μ de la population, au risque α :

$$\left]\overline{y} - u_\alpha \frac{s}{\sqrt{n-1}}; \overline{y} + u_\alpha \frac{s}{\sqrt{n-1}}\right[$$

• **Cas des petits échantillons :**

Théorème :

Si Y suit une loi normale,$\overline{y}$ suit une loi normale $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Si σ est connu, l'intervalle de confiance de μ se construit alors comme dans le cas des grands échantillons. Mais en général σ est inconnu et estimé par s. dans le cas des petits échantillons, en remplaçant σ par s on modifie la nature de la loi suivie par $\overline{y}$.

Théorème :

La variable aléatoire $T = \frac{Y-\mu}{\frac{s}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté.

Utilisation :

Une loi de Student est une loi de probabilité continue dont la densité est une fonction paire et qui dépend d'un paramètre appelé nombre de degrés de liberté **(d.d.l.)**. Le cœfficient de risque α étant choisi et le nombre de degrés de liberté étant connu, la table de Student permet de lire t_α tel que :

$$P\left[-t_\alpha < T < t_\alpha\right] = 1 - \alpha$$

.On en déduit comme intervalle de confiance de la moyenne μ de la population :

$$\left]\overline{y} - t_\alpha \frac{s}{\sqrt{n-1}}; \overline{y} + t_\alpha \frac{s}{\sqrt{n-1}}\right[$$

2.10.3 Estimation d'une variance :

Théorème :

Si Y suit une loi normale, la variable aléatoire $X^2 = \frac{n-1}{\sigma^2}s^2$ suit la loi du Khi-deux à $n-1 = \nu$ degrés de liberté.

Utilisation :

• **Si n $\leq$ 31 :**

Le coefficient de risque α étant choisi et le nombre de degrés de liberté étant connu, la table de Khi-deux permet de déterminer a et b tels que :

$$P\left[a < X^2 < b\right] = 1 - \alpha$$

• Avec :

$$P\left[X^2 \geq b\right] = \frac{\alpha}{2} \quad et \quad P\left[X^2 \geq a\right] = 1 - \frac{\alpha}{2}$$

La seule valeur connue de S^2 étant s^2, on obtient comme intervalle de confiance de la σ^2 au risque α :

$$\left]\frac{n-1}{b}s^2; \frac{n-1}{a}s^2\right[$$

• **Si n > 31 :**

Le théorème cité est vrai quelque soit n. mais les tables de Khi-deux s'arrêtent habituellement au degré de liberté $\nu = 30$. On ne peut pas donc les utiliser si $n > 31$.

Théorème :

Si X^2 est une variable aléatoire qui suit une loi de Khi-deux à ν degrés de liberté et si $\nu > 30$, alors la variable aléatoire $U = \sqrt{2X^2} - \sqrt{2\nu - 1}$, suit sensiblement la loi réduite de Gauss.

Utilisation :

Ici on a :$U = \sqrt{\frac{2(n-1)}{\sigma^2}}S^2 - \sqrt{2n-3}$. Après avoir choisi α, on détermine u_α tel que $P\left[-u_\alpha < U < u_\alpha\right] = 1 - \alpha$, et on en déduit l'intervalle de confiance de la σ^2 :

$$\left]\frac{2(n-1)s^2}{(\sqrt{2n-3}+u_\alpha)^2}; \frac{2(n-1)s^2}{(\sqrt{2n-3}-u_\alpha)^2}\right[$$

Chapitre 3

Elaboration d'un questionnaire

3.1 Introduction :

Le sondage s'oppose au recensement qui est l'étude exhaustive de toute la population. Durant ces vingt dernières années, la pratique des sondages a été largement popularisée grâce à la médiatisation des sondages préélectoraux. Dans tous les pays démocratiques, l'utilisation intensive des techniques de sondage a incontestablement modifié l'art de gouverner. Ces sondages préélectoraux sont pour la plupart réalisés par des sociétés privées à des fins publicitaires. Cependant, ce type d'application est relativement marginal et peut donner une idée erronée de l'intérêt de la théorie des sondages essentiellement développée par les instituts de statistique officielle.

Ceux-ci ont pour mission de produire des données impartiales relatives à la vie économique et sociale d'un pays. Comme il est souvent impossible d'effectuer un recensement sur tous les sujets, le recours aux sondages constitue une option intéressante.

Les résultats ainsi obtenus sont, bien sur, entachés d'une erreur d'échantillonnage due à une sélection aléatoire d'une partie de la population. Al'issue de l'enquête, ces erreurs de mesure peuvent néanmoins être réduites en réalisant de manière plus fiable les interviews, la transcription, l'encodage et le traitement des données. Un sondage exécuté avec rigueur fournit généralement des estimateurs plus précis qu'un recensement. C'est pourquoi l'utilisation de données de sondages a connu un développement très important depuis la Seconde guerre mondiale.

La théorie des sondages a joué un rôle prépondérant dans le développement de la statistique ; cependant l'utilisation de techniques d'échantillonnage n'a été acceptée que très récemment. Parmi les polémiques qui ont animé cette théorie, on retrouve certains débats classiques de la statistique mathématique, comme la place de la modélisation et la discussion sur des techniques d'estimation.

La théorie des sondages a été tiraillée entre les grands courants de la statistique et a donné naissance à de multiples approches : basé sur un modèle prédictif, bayesienne...etc.

3.2 L'utilisation de sondages :

Les sondages font partie de disciplines qui, tout en étant très mal connues dans leurs fondements par le grand public, n'en demeure pas moins abondamment mises en œuvre dans les aspects les plus divers de la réalité quotidienne. Ce sont les sondages d'opinion et les sondages sur les modes de vie qui, bien adaptés à la médiatisation, constituent la forme extérieure la plus "envahissante" des sondages.

Cette diffusion, parfois excessive a pour conséquence directe d'avoir familiarisé une très grande majorité des citoyens avec le concept de sondage.

Cependant, il serait absurde de restreindre l'emploi du sondage au cas des enquêtes sur les opinions et les modes de vie. On peut donner des exemples moins connus d'utilisation des techniques de sondage :

-La recherche de gisements pétroliers, la recherche minière donnent lieu à un sondage au sens étymologique du terme.
-Recherche des conditions opérationnelles optimales.
-Enquêtes sociologiques et démographiques.
-Enquêtes statistiques gouvernementales(salaires et chômage, maladies..).
-La détermination du volume de certaines productions agricoles.
-La vérification d'une comptabilité d'entreprise importante.
-Les contrôles fiscaux.
-Etudes commerciales et industrielles (production, distribution, inventaire, gestion de stock).

-Etudes de trafic (rail, routes, air).

-Les calculs de grands indices "médiatiques", comme l'indice des prix à la consommation, l'indice du coût à la construction.

-Le contrôle de la qualité de fabrication des automobiles sur une chaîne d'usine.

-L'estimation du pourcentage d'individus qui n'ont pas été recensés lors d'une opération de recensement de population.

-Recherche médicale (traitements, études de corrélation entre maladies et causes possibles).

Même en sport, les techniques de sondage sont utilisés :les contrôles antidopage dans certaines compétitions, au moins en partie, réalisés sur des athlètes sélectionnés par sondage .

3.3 Les étapes de la mise en œuvre d'une enquête par sondage :

La mise en œuvre d'une enquête par sondage peut être présentée comme une succession d'étapes :

a) Définir les objectifs de l'enquête et ses contraintes : on y précise ce que l'on veut etudier, quel est le champ de l'enquête.Les contraintes sont essentiellement les contraintes de coût, les contraintes de collecte et les contraintes de disponibilité d'informations auxiliaires ;il peut aussi s'agir de contraintes de nature organisationnelle, ou tout simplement de contraintes dûes à la limitation des techniques aujourd'hui maitrisées par les spécialistes des sondages.

b) Recherche de toute l'information disponible sur le thème de l'étude :

En effet, toute information relative à la population étudiée, aux thèmes abordés, ou issue d'enquêtes antérieures peut être mise à profit pour construire le plan de sondage et les estimateurs.

c) Définition des thèmes abordés : Cette étape débouche sur la définition des variables à étudier et de la population cible.

d) Définition de la population et de la base de sondage : Cette phase, souvent conjointe à la précédente, est moins évidente qu'il n'y parait a priori. Comment, par exemple, définir la population d'un pays donné ? Faut-il y inclure les étrangers domiciliés dans ce pays, les nationaux expatriés ? Il convient dès lors de procéder en deux temps : d'une part, définir avec précision la population cible sur un plan conceptuel. D'autre part, construire

une procédure permettant d'avoir accès à toutes les unités de la population ; identifiées appelée la base de sondage en tenant compte des aspects de qualité et de coût.

e)Bases de sondage :

Pour pouvoir réaliser un tirage probabiliste proprement, c'est-à-dire où chaque individu a une probabilité connue et fixée à l'avance de faire partie de l'échantillon, il est nécessaire de disposer d'une liste de toutes les unités d'échantillonnage faisant partie du champ de l'enquête. Cette liste doit avoir trois qualités principales :

- Elle doit permettre de repérer l'unité sans ambiguïté. (Exemple : nom, prénom, adresse, date de naissance).
- Elle doit être exhaustive
- Elle doit être sans double compte.

f)Construction du plan de sondage :

Il s'agit de la détermination de la manière dont les unités vont être sélectionnées. Si les unités sont sélectionnées selon une procédure aléatoire, le plan est dit probabiliste, sinon il est dit empirique. Souvent, le plan de sondage est construit en fonction d'informations connues pour l'ensemble de la population.

g) Représentativité des échantillons :

L'objectif d'un sondage est de fournir un certain nombre d'informations sur une population en n'examinant qu'une partie de celle-ci, appelée l'échantillon. On dit souvent qu'un échantillon est représentatif d'une population s'il en constitue le modèle réduit. La "représentativité" est ainsi invoquée en tant qu'argument de validité :

Un bon échantillon devrait "ressembler" autant que possible à la population à étudier.

h) Concevoir le questionnaire : on s'attache à la formulation des questions, à leur ordre, à leur pertinence, à la durée moyenne de remplissage, etc.

*i) La collecte des données :*est le processus qui permet d'obtenir l'information nécessaire. Les enquêtes peuvent être réalisées par des enquêteurs, par courrier, par téléphone ou par voie électronique. La collecte des données est évidement une étape cruciale qui doit être suivie rigoureusement.

j) Codage et création d'un fichier de données : en enchainant une codification automatique et une phase de reprise manuelle des échecs de codification automatique. Saisir au préalable les données au cas où le questionnaire serait sous forme papier.

k) Estimation des paramètres.

l) Analyser les données : il s'agit d'aller plus loin que l'étape précédente d'estimation des simples paramètres en analysant, par des méthodes appropriées (tests, analyse factorielle ou ajustement de courbes à des nuages de points..).quel que soit le traitement statistique qu'on appliquera aux données d'enquête, il est préférable de tenir compte de la façon dont ces données ont été recueillies, c'est-à-dire de tenir compte du plan de sondage.

m) Publier et procéder à une évaluation finale globale : à ce stade, on s'intéresse aussi à la confidentialité des données.

3.4 Elaboration d'un questionnaire :

Un questionnaire est un groupe ou une séquence de questions conçues afin d'obtenir d'un répondant de l'information sur un sujet. Les questionnaires ont un rôle central dans les enquêtes car ils ont des répercussions importantes sur la qualité des données. Ils ont aussi des répercussions sur l'image de marque que l'organisme statistique projette dans le public.

Les questions posées doivent être conformes à l'énoncé des objectifs de l'enquête et permettre la collecte d'information utile pour l'analyse des données. Elles doivent répondre à tous les besoins d'information, mais chaque question devrait avoir une justification explicite pour être inscrite dans le questionnaire. Il faut savoir pourquoi chaque question est posée et à quoi servira l'information.

La formulation de la question doit être claire. Les questions doivent être réparties en séquences logiques pour le répondant. Le processus de conception du questionnaire commence par la formulation des objectifs de l'enquête et des besoins d'information selon les étapes suivantes :

a) Consultation avec les utilisateurs des données et les répondants.

b) Examen des questionnaires précédents.

c) Élaboration d'une version provisoire du questionnaire.

d) Examen et révision du questionnaire.

e) Mise à l'essai et révision du questionnaire.

f) Touche finale apportée au questionnaire.

Un questionnaire bien conçu devrait :

- Permettre la collecte des données avec efficience et le résultat devrait comprendre un nombre minimal d'erreurs et de données incohérentes.

- être convivial pour l'intervieweur et le répondant.

- Diminuer dans l'ensemble le coût et le temps de la collecte des données.

On distingue essentiellement les questions ouvertes et les questions fermées.

- *Une question ouverte :* est une question à laquelle l'interviewé répond comme il le désire ; ce qu'il dit est en général intégralement enregistré par l'enquêteur.

- *Une question fermée :* est une question à laquelle les réponses possibles sont déterminées à l'avance par le rédacteur du questionnaire, proposées à l'enquêté souvent sur un carton-réponse en lui précisant le nombre de réponses qu'il peut choisir.

Les questions ouvertes permettent de recueillir une information parfois plus nuancée que les questions fermées. Elles donnent à l'interviewé le sentiment que son avis compte et permettent d'aérer un questionnaire complexe.

On essaie cependant d'éviter les questions ouvertes dans les études quantitatives. On y obtient souvent des réponses superficielles ou stéréotypées, l'interviewé citant les premières raisons qui lui viennent à l'esprit . D'autre part, l'analyse des questions ouvertes s'avère lourde et coûteuse malgré l'existence de logiciels « ad hoc » : il faut en effet regrouper les réponses ayant une signification semblable en un nombre limité de thèmes, ce qui implique de surcroît une certaine part de subjectivité. Mieux vaut donc « fermer » les questions en utilisant les résultats d'une étude qualitative préalable.

3.5 Coût de l'enquête :

La plupart des budgets supportent, malheureusement, assez mal les recensements. Il est alors nécessaire de limiter ses ambitions et de collecter l'information Y sur une partie de la population, en constituant un échantillon d'individus, réalisant ainsi par définition une enquête par sondage.

Trois conditions sont nécessaires pour qu'une enquête par sondage donne une description satisfaisante de la population étudiée selon le budget de l'enquête :

- Un choix judicieux du mode de désignation de l'échantillon.

- Une observation correcte et exacte des données à collecter auprès de chaque individu de l'échantillon.

- L'extrapolation des résultats de l'échantillon à l'ensemble de la population à partir d'estimateurs appropriés.

3.6 Plan de sondage :

Du point de vue du sondeur, le premier temps fort du processus d'enquête par sondage réside dans la sélection de l'échantillon d'individus. Il existe de nombreuses méthodes de tirage concurrentes pour produire un échantillon.

Sa seconde préoccupation est l'agrégation des réponses recueillies auprès des individus :

- Etape des estimateurs, c'est-à-dire des expressions mathématiques qui permettent de proposer une valeur pour la fonction des Y_i.

- Calcul de la précision de cet estimateur, qui conforte le sondeur dans son approche, car il revendique la possibilité de faire presque aussi bien que le recensement avec un coût beaucoup plus faible.

La méthode de tirage adoptée est la partie déterminante du coût global résultant du processus d'enquête. Déterminer la méthode de sélection de l'échantillon et la formulation de l'estimateur c'est déterminer par définition *le plan de sondage.*

3.6.1 sélection d'un échantillon :

L'échantillonnage permet aux statisticiens d'estimer les caractéristiques d'une population en observant directement un échantillon ; on distingue essentiellement deux types d'échantillonnage :

a) L'échantillonnage non probabiliste :

Dans ce type d'échantillonnage on considère que la distribution de caractéristiques à l'intérieur de la population est égale, ce que fait que le chercheur croit que n'importe quel échantillon est représentatif et par conséquent que les résultats sont exacts.De plus il n'y a aucun moyen de mesurer la précision de l'échantillonnage.

b) L'échantillonnage probabiliste :

Dans l'échantillonnage probabiliste chaque unité a une chance d'être sélectionnée ; c'est-à-dire que les individus sont sélectionnés selon une procédure aléatoire.

3.6.2 Construction des échantillons :

Toutes les considérations développées jusqu'à présent sur les méthodes statistiques reposent sur le tirage "au hasard" des individus composants un échantillon.

Il reste le problème d'effectuer un prélèvement qui soit rigoureusement aléatoire. Pour ce tirage on se sert généralement de résultats pré calculés fournis par :

- *Les tables de nombres au hasard :* ces tables proviennent de l'analyse, le plus souvent au moyen d'un ordinateur, de phénomènes obéissant à un hasard aussi parfait que possible, on a :

a) Table de M. G. Kendall et Barrington Smith : fournit 25.000 Nombres de 4 chiffres présentés sous forme de 100 tableaux de 250 nombres de 4 chiffres

b) Table du rand corporation : fournit ainsi 500 nombres au hasard, compris entre 0 et 99999.

Construction des échantillons : tables de nombres au hasard

Remarque :

Les tables de nombres au hasard ne servent pas uniquement à prélever des échantillons aléatoires sur une population inconnue, mais permettent également dconstruire des échantillons artificiels, dont les individus obéissent à telle loi de base qu'on choisit, en vue de reconstituer un processus aléatoire qu'on veut étudier.

Simulation : construction des échantillons artificiels

3.7 Différents types d'erreurs rencontrées dans les enquêtes par sondage :

3.7.1 Erreur d'échantillonnage :

Le sondage est caractérisé par une erreur d'échantillonnage mesurable, et qui se traduit par l'existence d'une loi de l'estimateur au travers d'une densité de probabilité. On peut assimiler à des erreurs d'échantillonnage certaines erreurs résultant des imperfections de la base de sondage que sont la présence des doubles comptes et le vieillissement de l'information auxiliaire. Il s'en suit une perte de précision en cas de redressement, c'est-à-dire que la variance augmente mais cela ne génère pas de biais.

L'erreur d'échantillonnage existe dans toutes les enquêtes par sondage, mais elle est nulle dans les enquêtes par recensement.

3.7.2 Erreur d'observation :

Dans la réalité, il existe une seconde famille d'erreurs appelées erreurs d'observation ou erreurs de mesure, qui tient au fait que la valeur que l'on recueille lors de l'enquête peut être une valeur différente de la vraie valeur qui est attachée à l'individu i .Ce type d'erreur survient surtout dans les questions sensibles. Elle ne peut être évitée systématiquement sur des sujets délicats comme le revenu, la sexualité, la consommation de drogue, etc...

Cette erreur est plus souvent volontairement introduite par l'enquêté, il existe parallèlement à cela une longue liste des causes ; on peut citer pêle-mêle, et sans être exhaustif :

- Erreurs de bonne foi de l'enquêté, c'est le cas si on fait appel à sa mémoire et que celle-ci est défaillante.

- Erreurs introduites par l'enquêteur qui interprète les questions et ″souffle″ éventuellement les réponses grâce au défaut de formation de l'enquêteur, qui n'a pas à influencer la réponse de l'enquêté.

- *Erreurs d'écriture sur un questionnaire :* on se trompe de case en cochant, ou erreur de saisie des réponses ″ on frappe sur la mauvaise touche du clavier″.

- *Erreurs de codification :* c'est le cas systématiquement délicat des ″questions ouvertes″ où, au lieu de demander à l'enquêté de choisir entre plusieurs possibilités, celui-ci formule sa réponse ″en clair″ comme il l'entend .Alors le logiciel de codification (par exemple : SICORE à l'INSEE) hésite ou ne sait pas conclure.

- Mauvaise compréhension de la question par l'enquêté ″personne âgée, un étranger, personne handicapée″, ou simplement question ambigüe, mal formulée.

- L'individu échantillonné a été remplacé, volontairement ou par erreur, par un autre individu ; exemple : l'enquêteur ne trouve pas $Mr.X$ à son domicile, et va sonner chez son voisin, $Mr.Y$ en pensant qu'il fera bien l'affaire...

3.7.3 Défaut de couverture et non-réponse :

On peut distinguer une troisième famille d'erreurs, liée à l'existence d'une base de sondage incomplète″ c'est-à-dire non exhaustive″ et à la non-réponse de certains individus aux questions posées. En effet, une base de sondage incomplète est une situation qui donne lieu à un défaut de couverture de la population.

-*Erreurs de couverture* : Ce type d'erreur survient lorsque la base de sondage et la population cible ne coïncident pas parfaitement. Il y a sous-couverture lorsque certaines unités de la population cible ne sont pas sur la base de sondage et ne peuvent donc pas être sélectionnées. Il y a sur-couverture lorsque certaines unités sur la base de sondage ne sont pas dans la population cible.

-*Erreurs de non-réponse :* Malgré tous les efforts mis de l'avant pour maximiser la quantité d'information recueillie, il y aura inévitablement de la non-réponse. On distingue deux types de non-réponse :

- *La non-réponse totale :* elle survient lorsque l'unité échantillonnée ne répond à aucune des questions posées ; elle peut être causée par un refus de subir l'interview ou parce que l'organisme en charge de l'enquête a été incapable de contacter l'unité échantillonnée. La non-réponse totale est généralement traitée par l'utilisation de méthodes de re-pondération qui consistent à hausser le poids de sondage des répondants (qui est le nombre d'individus que le répondant dans l'échantillon représente dans la population) pour compenser l'absence des non-répondants.

- *La non-réponse partielle :* elle survient lorsque l'unité échantillonnée a répondu à certaines questions, mais pas toutes. Elle peut être causée parce que la question est perçue comme indiscrète. La non-réponse partielle est généralement traitée par l'imputation qui consiste à déterminer une valeur de remplacement pour la valeur manquante.

- *Bilan :* en résumé, pour une enquête par sondage.

ERREUR TOTALE = ERREUR D'ECHANTILLONNAGE + ERREUR D'OBSERVATION + ERREURS DUES AU DEFAUT DE COUVERTURE ET A LA NON-REPONSE

On notera qu'en général les erreurs sont inferieures dans les enquêtes par sondage à ce qu'elles sont dans les recensements, cela tient à la formation de l'enquêteur.

En général, l'enquêteur par sondage reçoit une formation plus approfondie que les agents recenseurs.

3.8 Traitement des non-réponses :

3.8.1 Le problème de non réponse :

La non-réponse entraine également une perte de précision puisque les résultats seront calculés sur un nombre de répondants plus faible que la taille d'échantillon initialement prévue.

Cette perte de précision possible peut être anticipée en estimant *a priori* un taux de non-réponse probable et en augmentant en conséquence le nombre d'individus sélectionnés.

Le taux de non-réponse ne mesure donc pas à lui tout seul l'erreur de non-réponse : un taux de non-réponse faible peut entrainer un biais important si répondants et non-répondants ont des comportements très différents en ce qui concerne les thèmes de l'enquête. Inversement, un taux de non-réponse élevé n'est pas trop grave si ces comportements sont très voisins, notons que ce biais ne diminue pas avec l'augmentation de la taille d'échantillon.

Le problème des non-réponses totales se pose différemment selon le type d'enquêtes et de méthodes de sondage utilisées :

- Dans les méthodes non aléatoires (méthode des quotas par exemple lorsque l'enquête est réalisée en face à face), l'utilisateur de résultats n'a généralement pas connaissance du problème des non-réponses, des refus que l'enquêteur a subis, etc... Il peut éventuellement en avoir une appréciation sommaire, en examinant les consignes données aux enquêteurs et le déroulement du terrain.

- Dans les méthodes aléatoires, par contre, les individus composant l'échantillon sont nommément désignés. On connaîtra donc après enquête le nombre de non- réponses, éventuellement classées par catégorie .

- Dans les enquêtes en ligne, avec questionnaire proposé aux visiteurs de sites, l'utilisateur n'a généralement aucune connaissance des taux de réponse.

- Enfin, dans les panels, les non-répondants à une vague de panel sont bien identifiés ; on possède généralement beaucoup d'informations les concernant (via le questionnaire de recrutement et leurs réponses à d'autres vagues du panel). On peut alors utiliser les informations pour corriger l'effet des non-réponses ou pour estimer les réponses manquantes. Rappelons que les panels sont des échantillons permanents ou quasi- permanents d'individus ou de ménages qui sont interrogés ou observés régulièrement sur leurs comportements.

Deux types de stratégie peuvent être utilisés pour réduire l'erreur de non-réponse :

• La première stratégie consiste à réduire au maximum le taux de non-réponse en soignant très consciencieusement les techniques de collecte (questionnaires, enquêteurs, rappels...).

• la deuxième stratégie consiste à réduire a posteriori l'erreur de non-réponse en utilisant des techniques de repondération (pour la non-réponse totale) ou d'imputation (pour la non- réponse partielle).

3.8.2 Le traitement de la non-réponse :

A posteriori, il est possible de réduire l'effet de la non-réponse par l'utilisation de méthodes statistiques adaptées.

Pour *la non-réponse totale*, une stratégie envisageable est celle de la repondération de l'échantillon des répondants : on réajuste sur des distributions connues. Formellement, les calculs sont les mêmes que ceux de la post-stratification ou du redressement sur critères multiples, mais ils n'ont pas le même statut théorique.

Dans ces dernières méthodes, on considère que les écarts entre les distributions observées sur l'échantillon et sur la population résultent des seuls aléas d'échantillonnage et les calculs relèvent de la seule théorie mathématique de l'estimation.

Dans la repondération pour non-réponse, on postule, de plus, l'hypothèse plus ou moins vérifiable que, dans chaque catégorie définie par les variables de redressement, répondants et non-répondants sont en moyenne semblables.Les méthodes de repondération permettent de réduire le biais dû à la non- réponse mais pas nécessairement la variance.

Pour *la non-réponse partielle*, on utilise des méthodes par *imputation :*

On remplace chaque donnée manquante par une donnée « prédite » en fonction des informations obtenues pour le même individu et pour des individus proches. On peut distinguer deux grandes classes de méthodes d'imputation :

• *l'imputation déterministe* est telle que la valeur imputée ne change pas si l'on relance la procédure d'imputation. Parmi les méthodes, citons l'imputation utilisant un modèle explicite de type linéaire : les coefficients du modèle sont estimés sur l'échantillon des répondants et permettent d'estimer les valeurs imputées pour les non-répondants.

• *l'imputation aléatoire :*est telle que la valeur imputée change si l'on relance la procédure d'imputation. Parmi les méthodes possibles, citons les méthodes de hot-deck dans lesquelles la valeur inconnue de la variable d'intérêt pour un non-répondant est remplacée par la valeur observée d'un répondant tiré au hasard, soit parmi l'ensemble des répondants (hot-deck d'ensemble), soit dans une classe à laquelle appartient le receveur (hot-deck par classe), etc...

3.9 Avantages et inconvénients d'une enquête par sondage :

3.9.1 Avantages des enquêtes par sondage :

- *Cout et rapidité :* une enquête sans difficulté particulière, effectuée sur un petit échantillon est moins coûteuse et peut être lancée rapidement et fournir des bons résultats dans de courts délais.

- *Souplesse dans le choix des concepts* : cet avantage est surtout sensible par rapport aux informations obtenues comme sous-produit d'une activité administrative ou de gestion.

- *La méthode de sondage* :a un champ d'application plus vaste que les recensements, les thèmes relatifs à la consommation et les budgets des ménages, la lecture de presse et à l'opinion politique ne sont réalisés que par voie d'enquêtes par sondage.

- *Les enquêtes par sondage sont moins coûteuses* et plus rapide qu'un recensement. Bien que les coûts unitaires soient plus élevés, comme le nombre d'unité à observer est plus faible ,le coût global est relativement modéré.

- *Les enquêtes par sondage* offrent plus de possibilités pour la réduction de l'erreur d'observation. Le nombre d'unité échantillon étant peu nombreux, il est possible de confier la réalisation de l'enquête à un personnel qualifié et très bien formé, ce n'est pas le cas dans un recensement.

Par ailleurs, les techniques de mesure objective et directe des phénomènes à étudier sont faciles à utiliser dans les enquêtes par sondage .Tous ces efforts de mesure sont dictés par le souci de réduire les erreurs d'observation.

3.9.2 Inconvénients et limites des enquêtes par sondage :

- Les enquêtes par sondage sont fondées sur la loi des grands nombres, les extrapolations d'un échantillon ne peuvent être réalisées avec une précision acceptable qu'à partir d'échantillons de taille suffisamment importante, c'est pour cela que la méthode des sondages n' est pas fiable et applicable pour l'étude des populations de faibles tailles, et sans intérêt dans le cas d'une population réduite et délicate. Il est très hasardeux de mener, par exemple, une enquête par sondage pour estimer le nombre de personnes souffrants d'une infirmité physique déterminée et leur répartition selon diverses caractéristiques sociodémographiques.

- Les enquêtes par sondage ne permettent pas de donner des résultats très détaillés surtout sur le plan géographique.

- Toutes les populations suffisamment bien définies et bien localisées peuvent être étudiées par sondage, pour cela, les difficultés seront encore plus grandes. Lorsqu'on procède par sondage probabiliste, on doit disposer d'une base de sondage qui donne la liste de tous les individus composant la population avec leurs adresses.

Malgré ses limites, la méthode des sondages offre de grands avantages, elle permet d'étudier des domaines hors d'atteinte des recensements et statistiques administratives ; fidélité des consommateurs à une marque, à un parti politique, de ce fait, le chercheur est beaucoup plus libre dans la définition des concepts précisant l'objectif de sa recherche.

Les sondages obéissent à des règles très précises, mais la méthode est assez souple pour s'adapter à des objectifs différents. Enfin c'est la seule technique permettant de recueillir des informations quantifiées sur des collectivités importantes.

3.10 Réalisation d'une enquête par sondage :

3.10.1 La méthode de tirage :

est le processus que l'on a choisi pour tirer l'échantillon. Parmi les données à considérer à l'issue du tirage, on s'attache fondamentalement aux valeurs de probabilité que chaque individu tiré avait d'être sélectionné dans l'échantillon avant que le tirage ne se fasse, dans un sondage dit « probabiliste »,ces

probabilités sont d'ailleurs calculables pour chaque individu de la population et pas seulement pour les individus tirés.

Elles ne sont pas caractéristiques de la méthode de tirage ; car plusieurs méthodes de tirage différentes peuvent conduire au même jeu de probabilités, elles sont fondamentales car systématiquement présentes dans le calcul des estimateurs (sur ce point, seulement celles qui concernent les individus de l'échantillon).

3.10.2 l'expression de l'estimateur :

Par contre,c'est la « formule » choisie pour estimer le paramètre inconnu θ qui nous intéresse, pour une méthode de tirage donnée. Il existe de nombreux estimateurs concurrents. Réciproquement, un estimateur donné peut être appliqué à des échantillons tirés selon des méthodes différentes.

Le choix de la méthode de tirage et de l'estimateur dépend du biais, de la variance, du coût, ainsi que de la disponibilité d'information adéquate.

3.11 Conclusion :

Les discussions historiques autour des méthodes de sondage ont suscité la naissance de deux courants de pensée voire même deux schémas de sondage. La distinction entre les deux tient à la technique de tirage de l'échantillon et oppose d'une part les organismes statistiques et les instituts d'opinion.

Les premiers estiment que la seule méthode valable, scientifique, rigoureuse, etc. est celle qui permet l'utilisation du calcul des probabilités, le prélèvement de l'échantillon s'opérant selon les lois du hasard. Cette technique est qualifiée de sondage probabiliste ou scientifique.

Les seconds pratiquent des sondages *non probabilistes ou Empiriques*. L'examen de ces techniques montrera la signification et les conséquences de l'un et de l'autre choix.

Chapitre 4

Méthodes des sondages :

Première partie

Les Sondages probabilistes :

Par définition, les sondages probabilistes sont ceux pour lesquels chaque individu de la population a une probabilité donnée connue d'avance d'appartenir à l'échantillon dite *probabilité d'inclusion*, ou *probabilité de sélection.*

Le procédé le plus élémentaire consiste à tirer avec probabilités égales les n individus-échantillon parmi les N individus composant la population.

Les sondages probabilistes ont l'avantage de permettre des études de précision des estimateurs utilisant essentiellement la théorie et le calcul des probabilités. Ils ont un caractère ″scientifiquement rigoureux″ et une théorie mathématique qui les justifie, permettant de faire des choix fondés sur des hypothèses nécessairement formalisables.

4.1 Les Sondages probabilistes simples :

Nous commencerons par présenter les aspects principaux du sondage aléatoire simple, qui a pour atout principal de ne pas nécessiter l'utilisation de l'information auxiliaire.

4.1.1 Le sondage aléatoire simple :

Le sondage aléatoire simple consiste à tirer dans une population de taille N un échantillon de taille n sans remise de façon à ce que chaque individu ait la même probabilité d'appartenir à l'échantillon.

Le tirage se fait soit :

- Avec remise : un individu peut être choisi plusieurs fois, c.-à-d. si l'élément extrait de la population pour effectuer l'échantillonnage, est remis dans cette population après . L'échantillonnage est dit *non exhaustif.*

- sans remise : un individu déjà choisi ne peut l'être de nouveau .L'échantillonnage est dit échantillonnage exhaustif.

Un plan est dit simple si tous les échantillons de même taille ont la même probabilité d'être sélectionnée.

On considère une population bien déterminée et une variable formalisant l'information qui nous intéresse, appelée *variable d'intérêt* et que l'on notera Y, définie sur chaque individu de cette population. L'individu est l'unité de base à laquelle on s'intéresse et la population est l'ensemble des individus. Par la suite, on supposera toujours que la population est de taille finie et connue. Cette taille sera notée N.

En général, on ne cherche pas à connaître la valeur Y_i prise par chacun des individus i, l'intérêt porte plutôt sur une fonction de ces Y_i qui constitue l'information que l'on cherche :

Variables quantitatives : total, moyenne, dispersion, quantiles.

Variables qualitatives : pourcentage.

Le sondage aléatoire simple est la méthode d'échantillonnage la plus facile à appliquer et la plus couramment utilisée, ceci revient au fait qu'elle n'exige pas de données supplémentaires autre que la liste de sondage.Il existe pour cette méthode les formules types pour déterminer les estimations, les tailles,etc,....

Expressions des estimateurs du total et de la moyenne et d'un pourcentage :

a) estimateur de moyenne : On l'estime par

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

b) estimateur du total : on l'estime par

$$T = \sum_{i=1}^{N} Y_i = N\overline{Y}$$

c) estimateur d'un pourcentage :

L'estimation d'une proportion (pourcentage) est un cas particulier de moyenne. En effet, la population des individus dont on cherche la proportion constitue ce qu'on appelle un domaine. Pour chaque individu i de la population, fabriquons une variable Y_i qui vaut 1 si i est dans le domaine d'intérêt noté D (un étudiant, une femme de plus de 40 ans) et 0 sinon (Y est la variable indicatrice du domaine D). La moyenne vaut :

$$\overline{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} = \frac{N_D}{N} = p$$

Où N_D est l'effectif vrai (inconnu) d'individu appartenant au domaine D et P est le pourcentage recherché. Par conséquent, P apparaît comme une moyenne et on l'estime sans biais par :

$$\widehat{p} = \overline{y}$$

Expressions de la variance des estimateurs :

On note $f = \frac{n}{N}$,le taux de sondage. On considère l'expression :

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$$

Il s'agit de la variance " corrigée" ou "dispersion " de la variable Y.

a) Cas de l'estimateur de la moyenne :

On a alors la variance de $\overline{y}$:

$$V\left(\overline{y}\right) = (1-f) \times \frac{S^2}{n}$$

Cette formule prouve que, pour réaliser un sondage aléatoire simple qui fournisse des résultats précis il faut :

- Une taille d'échantillon n grand.
- Un taux de sondage f grand (proche de 1).
- Une dispersion S^2 faible.

On peut montrer que la moyenne simple $\overline{y}$ estime sans biais $\overline{Y}$ si le tirage est simple avec remise et que sa précision vaut :

$$V\left(\overline{y}\right) = \frac{1}{n} \times \frac{N-1}{N} \times S^2$$

b) Cas de l'estimateur d'une proportion :

On a :

$$S^2 = \frac{N-1}{N} \times P(1-P)$$

On suppose N grand, soit $S^2 \approx P\left(1-P\right)$, où P est toujours la vraie proportion.

D'où :

$$V\left(P\right) = (1-f) \times \frac{P(1-P)}{n}$$

c) Cas de l'estimateur d'un total :

La variance de l'estimateur du total vaut :

$$\begin{aligned} V\left(\widehat{T}\right) &= N^2 \times V\left(\overline{y}\right); \\ V\left(\widehat{T}\right) &= N^2 \times (1-f) \times \frac{S^2}{n} \end{aligned}$$

Comparaison des plans avec remise et sans remise :

Le plan de sondage simple sans remise est toujours plus intéressant que le plan avec remise. En effet, le rapport des variances des estimateurs de moyenne pour les plans sans remise et avec remise pour une même taille d'échantillon est toujours strictement inferieur à1 ; $\forall\ n \geq 2$ On a :

$$\frac{V\left[\overline{Y}_{SR}\right]}{V\left[\overline{Y}_{AR}\right]} = \frac{\frac{(N-n)}{nN}S^2}{\frac{\delta^2}{n}} = \frac{N-n}{N-1} < 1$$

Le tableau reprend les différents estimateurs. Il est important de noter que la variance calculée dans l'échantillon n'estime pas la même fonction d'intérêt selon qu'elle est calculée dans un plan simple avec ou sans remise.

Dans le cas d'un échantillon aléatoire simple avec remise, il est toujours plus intéressant de ne conserver que les unités distinctes.

Plan de sondage simple	**Sans remise**	**Avec remise**
Taille de l'échantillon	n	m
Estimateur de la moyenne	$\overline{Y}_{SR} = \frac{1}{n}\sum_{k\in S} y_k$	$\overline{Y}_{AR} = \frac{1}{m}\sum_{k\in S} y_k$
Estimateur de la variance de la moyenne	$V\left[\overline{Y}_{SR}\right] = \frac{(N-n)}{Nn}$	$V\left[\overline{Y}_{AR}\right] = \frac{\delta^2}{m}$
Esperance de la variance de l'échantillon	$E\left(s^2\right) = S^2$	$E\left(\widetilde{s}^2\right) = \delta^2$
Estimation de la variance de l'estimateur	$\widehat{V}\left[\overline{Y}_{SR}\right] = \frac{N-n}{Nn}$	$\widehat{V}\left[\overline{Y}_{AR}\right] = \frac{\widetilde{s}^2}{m}$

4.1.2 Le sondage systématique :

Les unités de la population mère ayant été numérotées de 1 à N, un sondage systématique consistera à prélever n unités numérotées : $h, h+k,$ $\ldots, h+(n-1)k$.

h : s'appelle l'origine, étant pris au hasard tel que $1 \leq h \leq k$

k :pas de tirage, tel que $k = \frac{N}{n}$.

h et k étant des nombres entiers satisfaisant à la relation :

$$h + (n-1)k \leq N \leq h + nk.$$

On dispose à l'origine d'un fichier pas nécessairement informatique comprenant des individus dans un ordre fixé. Le principe consiste à calculer d'abord un nombre appelé″ **PAS** de tirage″,à tirer un individu au hasard en début de fichier et partant de celui-ci ,à″ descendre″ ensuite le long du fichier en retenant un individu tous les **PAS** individus ,aux arrondis près.

On utilise souvent cette méthode dans l'industrie où l'on sélectionne une unité pour des essais à partir d'une chaîne de production, pour vérifier que la machinerie et l'équipement sont d'une qualité uniforme.

Sondage systématique : individus régulièrement espacés

Avantages et inconvénients de sondage systématique :

- L'avantage de sondage systématique vient du fait que la sélection d'un échantillon peut être plus facile.

- Cette méthode est extrêmement simple à programmer.

- Ainsi qu'elle permet d'obtenir de forts gains en précision dès que l'on connait une variable auxiliaire bien corrélée avec la variable d'intérêt.

En sus des avantages principaux évoqués ci-dessus, le tirage systématique peut s'avérer utile dans certaines circonstances :

- Il est applicable lorsqu'on ne connait pas a priori la taille N de la population auquel cas l'échantillon final aura une taille aléatoire.

- Il fournit une méthode simple et efficace pour gérer un mécanisme de tirage d'échantillon dans le temps lorsque la base de sondage est donnée.

- Il peut être facilement mis en œuvre sur le terrain par un enquêteur qui n'est pas spécialiste de l'échantillonnage et sa bonne application peut être contrôlée.

Le plus grand inconvénient est que les échantillons risquent de ne pas être représentatifs de la population.

4.2 Les sondages probabilistes complexes :

4.2.1 Le sondage stratifié :

La population est subdivisée en diverses catégories, appelées" *Strates* ", suivant un facteur autre que la caractéristique étudiée.

Définition de Strate :

La strate est un groupement homogène d'individus tirés de la population mère et présumés reliés entre eux par un ou plusieurs caractères.

Notations :

- La population de taille N est scindée en H strates G_1, G_2,..., G_H de tailles respectives :$N_1, N_2, \ldots, N_H$ telles que :

$$N = \sum_{h=1}^{H} N_h$$

- On tire de chaque strate de taille N_h un échantillon de taille n_h telle que :

$\sum_{h=1}^{H} n_h = n$,n est la taille de l'échantillon prélevé.
On note $f_h = \frac{n_h}{N_h}$: le taux de sondage dans la strate h.

L'estimation de la moyenne de la strate h est :

$$\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i$$

La variance corrigée ou la dispersion vraie est estimée par :

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{N} (y_i - \overline{y}_h)$$

Estimation :

La stratification permet d'améliorer considérablement la précision des estimations, pour un coût ordinairement faible, d'autant qu'il est possible de déterminer une répartition optimum de l'échantillon entre les strates.
Un estimateur stratifié de la moyenne $\overline{Y}$ est donné par :

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

L'estimateur du total est évidemment :

$$\widehat{T}_{st} = N\widehat{\overline{Y}}_{st} = \sum_{h=1}^{H} N_h \times \overline{y}_h$$

L'estimation de la proportion est :

$$\widehat{P}_{st} = \sum_{h=1}^{H} \frac{N_h}{N} \times P_h$$

Calcul de précision :

Un estimateur sans biais de la variance $V\left(\widehat{\overline{Y}}_{st}\right)$ est donné par :

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times (1 - f_h) \times \frac{s_h^2}{n_h - 1}$$

Un estimateur sans biais de la variance $V\left(\widehat{P}_{st}\right)$ est :

$$\widehat{V}\left(\widehat{P}_{st}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \times \frac{P_h\,(1 - P_h)}{n_h}$$

- On se fixe le seuil α et on cherche la valeur $Z_{1-\alpha/2}$ dans la table de la loi normale centrée et réduite.

- L'intervalle de confiance de $\widehat{\overline{Y}}_{st}$ au seuil α est :

$$I.C = \left[\widehat{\overline{Y}}_{st} - Z_{1-\alpha/2}\sqrt{V\left(\widehat{\overline{Y}}_{st}\right)}\ , \widehat{\overline{Y}}_{st} + Z_{1-\alpha/2}\sqrt{V\left(\widehat{\overline{Y}}_{st}\right)}\right]$$

- L'intervalle de confiance de $\widehat{P}_{st}$ au seuil α est :

$$I.C = \left[\widehat{P}_{st} - Z_{1-\alpha/2}\sqrt{V\left(\widehat{P}_{st}\right)}\ , \widehat{P}_{st} + Z_{1-\alpha/2}\sqrt{V\left(\widehat{P}_{st}\right)}\right]$$

Un estimateur sans biais de la variance $V\left(\widehat{T}_{st}\right)$ est :

$$\widehat{V}\left(\widehat{T}_{st}\right) = \widehat{V}\left(N\widehat{\overline{Y}}_{st}\right) = N^2 \times \widehat{V}\left(\widehat{\overline{Y}}_{st}\right)$$

On obtient :

$$\widehat{V}\left(\widehat{T}_{st}\right) = \sum_{h=1}^{H} N_h^2 \times (1 - f_h) \times \frac{s_h^2}{n_h - 1}$$

Taille d'échantillon par strate :

- On suppose que les strates sont déjà constituées, comment répartir une taille d'échantillon globale entre les différentes strates ?

a)Allocation proportionnelle « Répartition Représentative » :
Cette répartition consiste à utiliser le même taux de sondage f pour toutes les strates :

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

Un sondage stratifié vérifiant ces conditions est dit à allocation proportionnelle. Ce type de sondage est très courant ; essentiel pour trois raisons :

- Il permet justement d'écrire $\widehat{\overline{Y}}_{st} = \overline{y}$,et offre un avantage de simplicité dans le calcul de l'estimateur , puisque le calcul de$\overline{y}$ est l'opération d'estimation la plus simple que l'on puisse avoir .Le sondage devient dans ce cas particulier, un sondage à probabilités égales.
- Dans tous les cas, le tirage stratifié à allocation proportionnelle est plus précis que le sondage aléatoire simple. En effet, on voit généralement que la stratification apporte un plus en s'appuyant sur un raisonnement qualitatif de décomposition de la variance. On montre que si on note la variance du sondage stratifié à allocation proportionnelle et la variance issue d'un sondage aléatoire simple de même taille, on a :

$$V_{PROP}\left(\overline{y}\right) = \frac{1-f}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N} \times S_h^2\right)$$

$$\Rightarrow V_{PROP}\left(\overline{y}\right) \overset{\frown}{=} \frac{1-f}{n}\left(S^2 - \sum_{h=1}^{H}\frac{N_h}{N}\left(\overline{Y}_h - \overline{Y}\right)^2\right)$$

$$\overset{\backsim}{=} V_{SAS}\left(\overline{y}\right) - \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left(\overline{Y}_h - \overline{Y}\right)^2$$

$$\Rightarrow GAIN \overset{\backsim}{=} \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\left(\overline{Y}_h - \overline{Y}\right)^2$$

Par conséquent, il ressort :

$$V_{PROP}\left(\widehat{\overline{Y}}_{st}\right) \leq V_{SAS}\left(\overline{y}\right)$$

- L'allocation proportionnelle ne dépend pas des Y_i : il n'y a pas besoin de connaitre quoi que ce soit a priori sur le comportement Y des individus, mais seulement les tailles N_h .Le gain de précision obtenu par rapport au sondage aléatoire simple est donc assuré quelles que soient les variables Y. Cela est assez remarquable et particulièrement bienvenu lorsqu'on traite plusieurs variables d'intérêt Y.

b) Allocation optimale « Répartition au sens de NEYMAN » :
Celle-ci consiste à respecter l'égalité suivante :

$$\frac{n_h}{N_h.S_h} = \frac{n}{\sum\limits_{h=1}^{H} N_h S_h} = cte$$

Il s'agit de trouver les n_h qui rendent minimum la variance $\widehat{V}\left(\widehat{T}_{st}\right)$ sous la contrainte :

$$\sum_{h=1}^{H} n_h = n$$

Par conséquent ,le gain est d'autant plus important, qu'il y a des différences dans les dispersions internes des strates.

Optimisation et coût d'un sondage stratifié :

Le coût de l'enquête est un critère qui peut également être pris en compte. On peut se poser la question suivante : disposant d'un budget C pour réaliser une enquête, et ayant décidé de constituer a priori H strates dans la population, on note c_h coût unitaire d'enquête en strate h ;alors quelles doivent être les tailles d'échantillon par strate ?

A priori notre seule contrainte est d'avoir :

$$C = \sum_{h=1}^{H} n_h.c_h$$

Où n_h est la taille d'échantillon cherchée dans la strate h.Il y a donc un très grand nombre de combinaisons possibles. Compte tenu de l'intérêt particulier présenté par l'allocation proportionnelle, on peut décider d'opter pour ce type d'allocation ; en déterminant la taille n globale par :

$$C = \sum_{h=1}^{H} \left(n.\frac{N_h}{N} \right) c_h$$

$$\Rightarrow n = C \Big/ \left[\sum_{h=1}^{H} \frac{N_h}{N}.c_h \right]$$

$$\Rightarrow n_h = \frac{N_h}{N}.C. \left[1 \Big/ \sum_{h=1}^{H} \frac{N_h}{N}.c_h \right]$$

Il est possible de faire mieux, en minimisant l'expression de la variance $V\left(\widehat{\overline{Y}}_{st}\right)$ sous la contrainte de budget, on montre qu'un optimum est atteint lorsque n_h est proportionnel à la valeur :

$$\frac{N_h.S_h}{\sqrt{c_h}}$$

C'est-à-dire pour le choix suivant (dit allocation de Neyman).

$$n_h = \frac{N_h.S_h}{\sqrt{c_h}}.\frac{C}{\sum\limits_{h=1}^{H} N_h.S_h.\sqrt{c_h}}$$

Sous réserve que ces n_h soient tous inferieurs ou égaux aux N_h.

Dans le cas ou $c_h = 1$ quel que soit h (coût unitaire identique dans chaque strate),on se donne la taille totale de l'échantillon notée n (c'est aussi le budget global C) ;on choisit :

$$n_h = n.\frac{N_h.S_h}{\sum_{h=1}^{H} N_h.S_h}$$

On montre qu'avec cette allocation, la précision notée $V_{OPTI}\left(\widehat{\overline{Y}}_{st}\right)$ vaut :

$$V_{OPTI}\left(\widehat{\overline{Y}}_{st}\right) = \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}.S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}.S_h^2$$

Finalement si $c_h = 1$, on a sous une contrainte de budget donnée :

$$V_{OPTI}\left(\widehat{\overline{Y}}_{st}\right) \leq V_{PROP}\left(\overline{y}\right) \leq V_{SAS}\left(\overline{y}\right)$$

Comme toujours, la dispersion S_h est inconnue. Il est alors nécessaire si on veut employer une taille voisine de la taille optimale de l'échantillon par strate, d'estimer a priori S_h . Plusieurs possibilités s'offrent : soit on utilise des estimations issues d'une enquête précédente portant sur le même sujet, soit on se fie à un expert en la matière qui communique une valeur approximative, soit on se cale sur le cas d'une proportion ,en estimant a priori la valeur P_h, soit encore on effectue le sondage en deux temps. La première enquête fournissant une estimation de la dispersion S_h .Cette dernière idée est assez peu employée en pratique, les responsables d'enquête ayant souvent l'impression de gaspiller des moyens dans le simple but de recueillir une information sur un ou plusieurs paramètres ″ici les S_h″ qui permettrait de réaliser de façon plus performante et dans un deuxième temps le tirage de l'échantillon final d'individus visités par l'enquêteur.

Même s'il existe sur de telles pratiques un blocage psychologique, assez naturel, il faut se convaincre que des études sérieuses montrent qu'il peut être bénéfique de consacrer une partie de son budget pour acquérir une connaissance nécessaire à la préparation d'un plan de sondage optimisé.

Les avantages de stratification :

- La réduction de coûts.
- Optimisation de la gestion.

Conclusion :

Un sondage stratifié est toujours plus efficace qu'un sondage aléatoire simple, ainsi que la stratification est certainement l'une des meilleures manières d'introduire une information auxiliaire dans une enquête afin d'augmenter la précision des estimateurs. De manière générale, si l'information est disponible, nous verrons qu'on a pratiquement toujours intérêt à stratifier.

4.2.2 Le sondage par grappes :

Quel que soit l'estimateur utilisé, le sondage aléatoire simple présente deux inconvénients :

- Il faut disposer d'une base de sondage complète
- Le sondage aléatoire simple peut disposer suffisamment d'échantillons pour que le coût total soit prohibitif.

Pour contourner ces deux difficultés, on peut utiliser la technique de sondage par grappes.

Le sondage en grappes est un cas particulier du sondage à plusieurs degrés. Ayant tiré un certain nombre d'unités à l'avant dernier degré de tirage, on réalise l'enquête au dernier degré de tirage auprès de tous les individus inclus dans ces unités. On dit que l'on réalise une enquête exhaustive"ou recensement"dans les unités de l'avant dernier degré de tirage, lesquelles sont appelées *grappes*.

On commence par construire des groupes d'individus qui soient disjoints et dont la réunion est la population toute entière. On tire alors (généralement par sondage aléatoire simple) un certain nombre de groupes dans la base de sondage de groupes qui à été constituée. Chaque groupe est un individu ou « unité d'échantillonnage ».

On parle de grappes. On réalise ensuite l'enquête auprès de tous les individus de la grappe *(recensement de la grappe).*

Estimation :

On note :
M : nombre total d'unités primaires constituées.
N_i : taille de l'unité primaire i.
N : taille de la population totale.

$$N = \sum_{i=1}^{M} N_i$$

$\overline{N}$: la taille moyenne des grappes

$$\overline{N} = \frac{1}{M} \sum_{i=1}^{M} N_i$$

Y_{ij} : la valeur de la variable d'intérêt pour l'individu j de l'unité primaire i .
T_i : $\sum_{j=1}^{N} Y_{ij}$: le total de la grappe i.
$\overline{Y}_i$: $\frac{1}{N_i} T_i$:la moyenne de la grappe i.
$\overline{T}$: $\frac{1}{M} \sum_{i=1}^{M} T_i$:Le total moyen.
T : $\sum_{i=1}^{M} T_i$:Le total global.
$\overline{Y}$:$\frac{T}{N}$:La moyenne générale.
m :taille de l'échantillon d'unité primaire.
L'estimateur du total est :

$$\widehat{T} = \sum_{i=1}^{M} \frac{T_i}{\frac{m}{M}} = M \times \frac{\sum_{i=1}^{M} T_i}{m}$$

Contrairement au cas du sondage aléatoire simple, il n'y a donc pas besoin de connaître la taille totale N de la population pour estimer un total T. Le sondage par grappe permet donc d'estimer N (*qui est un total particulier*). Cependant l'estimation sera de mauvaise qualité dès que les tailles N_i seront sensiblement différentes d'une unité primaire à l'autre. La connaissance de N est par contre nécessaire pour estimer la moyenne $\overline{Y}$:

$$\widehat{\overline{Y}} = \frac{1}{N} \widehat{T}$$

Dans le cas où les grappes ont des tailles différentes :

$$\widehat{\overline{Y}} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{N_i}{N}\right).\overline{Y}_i$$

Un estimateur sans biais de la moyenne est :

$$\widehat{\overline{Y}}_G = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{N_i}{N}\right).\overline{Y}_i$$

Dans le cas où les grappes sont de même taille :

$$N_i = N_0 = \overline{N} : \widehat{\overline{Y}} = \frac{1}{M}\sum_{i=1}^{M}\overline{Y}_i$$

Un estimateur sans biais de la moyenne est :

$$\widehat{\overline{Y}}_G = \frac{1}{m}\sum_{i=1}^{m}\overline{Y}_i$$

Calcul de précision :

Ces formules étant des cas particuliers des formules des tirages à plusieurs dégrés.

Soit :

$$S_1^2 = \frac{1}{M-1} \times \sum_{i=1}^{M}\left(T_i - \overline{T}\right)^2$$

La variance de l'estimateur $\widehat{T}$ est alors :

$$V\left(\widehat{T}\right) = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m}$$

De plus si on note :

$$s_1^2 = \frac{1}{m-1} \times \sum_{i=1}^{m}\left(T_i - \frac{\widehat{T}}{M}\right)^2$$

Alors s_1^2 est un estimateur sans biais de la variance. On en déduit la variance de l'estimateur sans biais de la moyenne :

$$V\left(\widehat{\overline{Y}}\right) = \left(\frac{M}{N}\right)^2 \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m} = \left(\frac{1}{\overline{\overline{N}}}\right)^2 \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m}$$

Dont un estimateur sans biais est :

$$V\left(\widehat{\overline{Y}}_G\right) = \left(\frac{M}{N}\right)^2 \left(1 - \frac{m}{M}\right) \times \frac{s_1^2}{m} = \left(\frac{1}{\overline{\overline{N}}}\right)^2 \left(1 - \frac{m}{M}\right) \times \frac{s_1^2}{m}$$

On a donc :

$$V\left(\overline{Y}_G\right) = \left(1 - \frac{m}{M}\right) \frac{1}{m-1} \sum_{i=1}^{m} \left(\frac{N_i}{N}\overline{Y}_i - \overline{Y}_G\right)$$

Cas particulier des grappes uniformes :

Dans ce cas on a : $N_i = N_0 = \overline{N}, n = m \times \overline{N}$ est une constante non aléatoire et :

$$\widehat{\overline{Y}}_G = \frac{M}{N} \sum_{i \in s} \frac{T_i}{m} = \frac{1}{m} \sum_{i=1}^{m} \frac{T_i}{N} = \frac{1}{m} \sum_{i=1}^{m} \overline{Y}_i$$

$$et \quad : \quad V\left(\widehat{\overline{Y}}_G\right) = \left(1 - \frac{m}{M}\right) \frac{1}{M-1} \frac{\sum_{i=1}^{M} \left(\overline{Y}_i - \overline{Y}_G\right)^2}{m}$$

D'où :

$$\widehat{V}\left(\widehat{\overline{Y}}_G\right) = \left(1 - \frac{m}{M}\right) \frac{1}{m-1} \frac{\sum_{i=1}^{M} \left(\overline{Y}_i - \overline{Y}_G\right)^2}{m}$$

Comparaison avec un sondage aléatoire simple :

Supposons qu'on puisse réaliser un échantillon de même taille n, sans tenir compte du groupement en grappes.

On a :

$$n = \sum_{i \in s} N_i = mN_0$$

La moyenne $\overline{y}$ sera estimée par :

$$\overline{y} = \frac{1}{n}\sum_{i\in s} y_i$$

Dont la précision sera :

$$V(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

Où :

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(Y_{ij} - \overline{Y}\right)^2$$

Comme $N = MN_0$ et $n = mN_0$,on a :

$$V(\overline{y}) = \frac{1}{N_0}\left(1 - \frac{m}{M}\right)\frac{S^2}{m}$$

Comparons $\overline{y}$ et $\widehat{\overline{Y}}_G$, on a :

$$V\left(\widehat{\overline{Y}}_G\right) = \left(1 - \frac{m}{M}\right)\frac{1}{M}\frac{\sum_{i=1}^{M}\left(\overline{Y}_i - \overline{Y}_G\right)^2}{m} = \left(1 - \frac{m}{M}\right)\eta^2 S^2$$

Où η^2 est le rapport de corrélation inter-grappes :

$$\eta^2 = \frac{\sum_{i=1}^{M}\frac{N_0}{N}\left(\overline{Y}_i - \overline{Y}_G\right)^2}{\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(Y_{ij} - \overline{Y}_G\right)^2}$$

Par conséquent :

$$V\left(\widehat{\overline{Y}}_G\right) < V(\overline{y}) \Rightarrow \eta^2 < \frac{1}{N_0}$$

Le sondage par grappes est donc meilleur que le sondage simple si le rapport de corrélation inter-grappes est inférieur à $\frac{1}{N_0}$.

Avantages et inconvénients des sondages par grappes :

L'avantage de ce type de sondage est toujours de limiter *les coûts de déplacement.*Parfois, on y trouve des avantages d'ordre pratique, ou des gains en terme de meilleure couverture d'une population, ou des taux de réponse plus élevés, ainsi qu'on n'établit la base de sondage que dans les grappes échantillon. L'inconvénient est que, le plus souvent, il y a perte de précision des estimateurs.

"dans certains cas très particuliers correspondant à un effet de grappe ρ négatif, il peut y avoir au contraire amélioration de la qualité des estimateurs".

ρ : est le coefficient de corrélation.

En conclusion :

Pour obtenir un bon sondage en grappes, l'expression de la variance nous montre qu'il faut :

- Des grappes hétérogènes (ρ faible), c'est-à-dire regroupant des individus aussi différents que possible les uns des autres du point du vue de la variable d'intérêt Y.

- des grappes de faible taille (N_0 soit petite) et par conséquent un grand nombre des grappes dans la population.

- Des grappes de tailles voisines.

- que les moyennes de grappes $\overline{Y}_i$ soient le plus semblable possible.

- Tirer un maximum de grappes (m grand).

Les trois premiers points sont relatifs à la constitution des grappes. Le dernier est proprement lié au tirage et directement fonction du budget dont on dispose. La bonne réalisation des trois premières conditions dépend aussi du budget, car il est bien évident que la recherche d'un découpage en grappes qui ait de bonnes qualités statistiques est plus long et plus coûteux qu'un découpage réalisé n'importe comment..

4.2.3 Sondage par « Redressement » :

Principe et définition :

Soit x_i une information auxiliaire quantitative et y_i la variable d'intérêt, liée à x_i.

Une estimation par la méthode de quotient est une méthode de redressement avec une autre variable quantitative.

Estimation et calcul de précision :

L'estimation par quotient de la moyenne est notée $\overline{Y}_q$ tel que :

$$\overline{Y}_q = \overline{y} \times \frac{\overline{X}}{\overline{x}}$$

$\overline{Y}$: La moyenne vraie obtenue à partir de la variable auxiliaire Y.
$\overline{x}$: La moyenne estimée obtenue à partir de la variable auxiliaire X.
$\overline{y}$: La moyenne estimée obtenue à partir de la variable auxiliaire Y.

Le but est d'estimer une moyenne, un pourcentage, un total. On est amené à penser qu'il existe une différence de même ordre entre $\overline{x}$ et $\overline{X}$,$\overline{y}$ et $\overline{Y}$.

L'estimateur $\overline{Y}_q$ est un estimateur biaisé, mais il peut-être plus précis que $\overline{y}$, on apporte une correction $\left(\frac{\overline{X}}{\overline{x}}\right)$.

Précision :
On va estimer le rapport inconnu $R = \frac{\overline{Y}}{\overline{X}}$ par $\widehat{R} = \frac{\overline{y}}{\overline{x}}$.
Dans la pratique nous calculons les quantités : $\overline{x}$,$\overline{y}$ S_x^2, S_y^2, S_{xy}

$$S_{xy} = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

La précision est calculée par l'écart quadratique moyen.
On distingue deux cas de tirage :

- *Tirage avec remise :* on estime l'erreur par

$$EQM\left(\overline{Y}_q\right) = \frac{1}{n}\left(S_y^2 - 2RS_{xy} + R^2S_x^2\right)$$

- *Tirage sans remise :* On l'estime par

$$EQM\left(\overline{Y}_q\right) = \frac{N-n}{Nn}\left(S_y^2 - 2RS_{xy} + R^2S_x^2\right)$$

Comparaison avec un sondage aléatoire simple :

On a :

- Sondage aléatoire simple :

$$V\left(\overline{Y}\right) = \frac{N-n}{Nn} S_y^2$$

- Sondage par quotient :

$$EQM\left(\overline{Y}_q\right) = \frac{N-n}{Nn}\left(S_y^2 - 2RS_{xy} + R^2 S_x^2\right)$$

Alors les sondages par quotient sont profitables si :

$$EQM\left(\overline{Y}_q\right) < V\left(\overline{Y}\right)$$

Par conséquent, on obtient : $\frac{S_{xy}}{S_y^2} > \frac{R}{2}$

Si cette inégalité est vérifiée la précision de l'estimateur par le quotient est meilleure que l'estimation par le sondage aléatoire simple.

Conclusion :

Le sondage par ratio «ou quotient » trouve rapidement une limite en pratique lorsqu'on dispose de plusieurs variables auxiliaires concurrentes, on peut programmer un ratio pour chaque variable auxiliaire, et retenir finalement une combinaison linéaire pondérée de tous ces ratios.

4.2.4 Le sondage à plusieurs degrés :

Principe :

Le sondage à plusieurs degrés consiste à designer les unités-échantillon en cascade :

-au premier degré de tirage, on tire, dans chaque unité primaire échantillon, un échantillon d'unités secondaires.

-au troisième degré de tirage, on tire, dans chaque unité secondaire échantillon,un échantillon d'unités tertiaires, etc...

Bien entendu, chaque unité secondaire doit être rattachée à une et une seule unité primaire ; chaque unité à une et une seule unité secondaire, etc...

Lorsque l'on a été contraint de tirer un échantillon de grappes (absence de répertoire sur les unités élémentaires), il n'est pas nécessaire de soumettre

à l'enquêteur tous les éléments de chaque grappe tirée au sort. A partir de ce sondage du premier degré, il est possible de prélever (au sort) un échantillon du deuxième degré et ainsi de suite, la désagrégation s'arrêtant au stade souhaitable.

Avantages et inconvénients de sondage à plusieurs degrés :

Le sondage à plusieurs degrés permet de simplifier l'établissement de la base de sondage : il suffit, par exemple, de dresser la liste des exploitations agricoles pour les seules commune-échantillons, on évite ainsi d'avoir à l'établir pour tout le territoire.

Le tirage à plusieurs degrés assure une concentration géographique des unités à observer qui permet de réduire fortement les frais de déplacement.

En revanche, la précision des estimateurs est ordinairement moins pour un échantillon tiré à plusieurs degrés que pour un échantillon élémentaire d'effectif correspondant : les exploitations agricoles appartenant à une même commune ont le plus souvent tendance à se ressembler ; l'effet de grappe est donc habituellement défavorable.

Cependant, à coût constant, l'efficacité d'un sondage à plusieurs degrés l'emporte généralement sur celle d'un sondage à un seul degré. De fait, les principales enquêtes sont effectuées d'après un plan de sondage à plusieurs degrés.

Pour l'établissement du plan de sondage, un problème fondamental est celui de la répartition optimale de l'échantillon entre les unités primaires et les unités secondaires.

En effet, sans modifier le coût de l'enquête, il est possible, par exemple, d'augmenter le nombre d'unités primaires d'échantillon à condition de diminuer corrélativement le nombre des unités secondaires d'échantillon par unité primaire et même vraisemblablement, le nombre total des unités secondaires.

4.2.5 Le sondage à plusieurs phases :

Dans ce cas, il y a combinaison de plusieurs sondages, les fractions de sondage varient, mais les unités restent les mêmes.

En général, on se contente de deux phases. Supposons que l'on doive effectuer une enquête concernant la consommation auprès d'un échantillon de ménages. En raison de son coût assez élevé, ce genre d'enquête ne peut se faire qu'auprès d'un nombre limité de ménages. En outre, un échantillon tiré par sondage élémentaire, contiendrait surtout des ménages des classes moyennes ; les ménages dont les ressources anormalement basses, comme ceux des classes aisées, ne figureraient qu'en nombre très réduit, constituant des groupes trop peu nombreux pour faire l'objet d'analyse valides. C'est pourquoi on sera amené, à un sondage en deux phases, l'unité de sondage reste le ménage.

Dans la première phase, on tire un grand échantillon de ménages que l'on soumet à une rapide enquête, dans laquelle chaque ménage indique le montant de ses ressources. Dans la seconde phase, ce grand échantillon est stratifié selon le montant indiqué. Dans les strates extrêmes, l'échantillon intégral est soumis au questionnaire. Dans les strates intermédiaires, on applique des taux variables, déterminés de manière à interroger des ménages en nombre suffisant pour l'analyse.

Deuxième partie

Les sondages non probabilistes :

4.3 Introduction :

Lorsque les méthodes précédemment décrites ne peuvent être utilisées pour des raisons méthodologiques ou de coûts, on peut avoir recours aux méthodes dites non aléatoires ou encore méthodes empiriques.

Par opposition, les sondages empiriques sont ceux qui ne permettent pas de calculer la probabilité d'inclusion des individus.

Un sondage est non probabiliste lorsque il n'inclut pas de mécanisme de sélection aléatoire des individus de la population : la probabilité de sélection des individus de la population n'est donc pas connue, peut même être nulle pour certains, d'où l'existence de biais et l'impossibilité de calculer le marges d'erreur.

Les sondages empiriques, comme leur nom l'indique, s'inspirent de considérations moins objectives, plus discutables, et sont par conséquent moins rigoureux. Ils se justifient davantage par des raisons qualitatives et quantitatives et leur utilisation est souvent dictée par des considérations budgétaires. Pour ces raisons, on trouve de nombreux détracteurs des méthodes empiriques, qui n'ont pas bonne presse dans les grandes revues anglo-saxonnes. En France, ils sont cependant fréquemment utilisés. Dans ce type de sondage, on suppose que la distribution de caractéristiques à l'intérieur de la population est égale, ce qui fait que le chercheur croit que n'importe quel échantillon est représentatif et par conséquent que les résultats sont exacts, de plus il n'y a aucun moyen pour mesurer la précision d'échantillonnage.

4.4 Avantages et inconvénients des méthodes empiriques :

a)Avantages :

- Peuvent être utilisées quand aucune base de sondage n'est disponible.
- Souvent plus rapides et moins couteuses que les échantillons aléatoires.

b) Inconvenients :

- La validité des échantillons non aléatoires ne peut être ´évaluée que subjectivement.
- On ne peut pas calculer des variances d'estimateurs.
- Ils fournissent habituellement des échantillons biaisés.

4.5 Différentes méthodes de sondages non probabilistes :

Parmi ces méthodes, on distingue :

4.5.1 Sondage de commodité (à l'aveuglette) :

On ne sélectionne des unités d'un échantillon que si on peut y avoir facilement accès. Par exemple les reporteurs de télévision sont à la recherche d'interview des gens dans la rue pour déterminer comment la population perçoit une question. Cette méthode est utilisée pour la recherche exploratoire.

4.5.2 Sondage au jugé :

Cette méthode consiste à prélever un échantillon sur la base de jugement d'une ou plusieurs personnes qui connaissent bien la population étudiée. Cette méthode est rapide et peu coûteuse, elle est préviligiée pour des travaux expérimentaux, on fait appel à l'expertise de spécialistes.

4.5.3 Sondage volontaire :

Comme son nom l'indique, ce type d'échantillonnage intervient lorsque des gens offrent volontairement leurs services pour l'étude, *Par exemple* : tester l'efficacité de certains médicaments.

Cette méthode aboutit souvent à des résultats biaisés puisque les personnes que se portent volontaires ont souvent des idées préconçues. L'échantillon n'est pas représentatif.

4.5.4 Sondage par la méthode de quotas :

On l'appelle aussi échantillonnage à choix raisonné, c'est la méthode empirique la plus fréquemment rencontrée. Elle consiste à décomposer la population en sous groupes de sorte que la structure de l'échantillon soit exactement celle de la population toute entière selon certains critères que l'on a préalablement choisis. Elle réduit aussi les frais d'enquête et ne nécessite pas de base de sondage, mais la représentativité n'est pas assurée.

Avantages et limites de la méthode des quotas :

a)Les avantages :

- La méthode de sondage par quotas n'exige pas la constitution d'une base de sondage.

- Le coût de cette méthode est moins élevé que le coût de la méthode probabiliste.

- La méthode des quotas s'adapte mieux aux domaines où il y a des difficultés d'observation. Ces difficultés proviennent du fait qu'on prévoit beaucoup de″non réponse″ pour les individus désignés par la méthode probabiliste. Les non réponses sont généralement dues soit à l'imprécision, soit à l'incomplétude des adresses, soit à la réticence ou l'incompétence des personnes interviewées.

- La méthode des quotas s'adapte mieux aux échantillons de faible taille, contrairement à la méthode probabiliste qui exige une taille minimale pour l'échantillon, la méthode des quotas ne pose aucune contrainte.

b) Les limites :

- La méthode des quotas ne permet pas de calculer l'erreur d'échantillonnage. Par conséquent, il est impossible d'estimer la variance et le coefficient de variation.

- Le défaut fondamental de cette méthode est qu'elle repose sur une répétition de principe, à savoir que la distribution des variables de contrôle détermine la distribution des variables à étudier.

4.5.5 Méthode des itinéraires :

C'est une variante de la méthode des quotas, au lieu de laisser l'enquêteur déterminer librement les individus à interroger, on lui impose de réaliser ses interviews dans certains endroits définis par avance sur une carte. Cette contrainte prend alors la forme d'un itinéraire imposé (un ensemble de rues dans une commune *par exemple*, ou même un ensemble d'immeubles) en fonction duquel, l'enquêteur fera la gestion de ses quotas.

L'itinéraire peut être lui-même défini aléatoirement, ou bien être minutieusement choisi selon une philosophie analogue à celle des unités-types

L'avantage, par rapport à la méthode des quotas classiques est de limiter la liberté de choix de l'enquêteur. L'inconvénient est le coût de préparation de l'opération qui vient en sus.

4.5.6 Méthode des unités types :

C'est la plus empirique des méthodes. Elle consiste à choisir un individu « moyen » que l'on déclare représentatif d'un groupe d'individus possédant les mêmes caractéristiques. Sa validité repose sur le pari audacieux : que les individus semblables à l'individu moyen selon les variables qui ont servi précisément à définir cet individu seront également semblables à l'individu moyen pour toutes les variables d'intérêt de l'enquête.

Cela revient à admettre implicitement qu'il existe une relation stable reliant la variable d'intérêt aux variables servant à définir l'individu « représentatif »,et que cette relation est la même pour tous les individus d'un même groupe.

Encore plus que pour la méthode des quotas, les résultats seront sensibles au tirage au travers du choix que l'on a fait de l'individu moyen, dit ″unité type″.

Critique de la méthode des unités types :

- Cette méthode s'adapte bien aux échantillons de faibles effectifs.
- Aussi elle exige une connaissance parfaite de la population à étudier pour pouvoir choisir des unités types et l'existence d'une unité type n'est pas certaine.
- Enfin cette méthode ne permet pas d'estimer l'erreur d'échantillonnage.

4.6 Conclusion :

Il existe plusieurs méthodes empiriques, mais les résultats de ces méthodes sont très peu fiables.

CONCLUSION

Le sondage étant une méthode d'enquête sur échantillon, comme il est souvent impossible d'effectuer un recensement sur tous les sujets, le recours aux sondages constitue une option intéressante pour optimiser les estimations des paramètres inconnus d'un certain caractère de la population, calculer la précision de ceux-ci et réduire les erreurs d'échantillonnage.

Un sondage exécuté avec rigueur fournit généralement des estimateurs plus précis qu'un recensement, c'est pourquoi cette science a connu ces dernières années de profondes innovations.

Les applications de cette science sont importantes et multiples : il n'est guère un secteur d'activité qui n'ait un jour besoin d'échantillonner afin d'observer,de comparer, de déduire ou de prévoir.

Derrière cet aspect bien connu au public, se cache un savoir faire de haute technicité et une science en plein développement.

Bibliographie

[1] ANNIE FOURRIER-REGLAT, MARTHE-ALINE JUTAND, ANNE-MARIE DUSSAIX-Initiation aux méthodes des sondages, Application à la pharmaco-épidémiologie, Arme pharmacovigilance éditions, Bordeaux, (2006).

[2] ARDILLY PASCAL-Les techniques de sondage, éditions Technip, (2006).

[3] BROSSIER GILDAS, ANNE MARIE DUSSAIX -Enquêtes et sondages : méthodes, modèles, application et nouvelles approches, Dunod, Paris,(1999).

[4] DEMING.W.E -quelques méthodes de sondage : Revue de statistique appliquée, tome12,no4, p.11-55.Université de NEW-YORK.

[5] DESABIE- Théorie et pratique des sondages, Dunod, Paris, (1966).

[6] DURAND CLAIRE-Histoire des sondages,(08 janvier 2001).

[7] GOURIEROUX CHRISTIAN- Théorie des sondages, ed . Economica,(1981).

[8] GRAIS BERNARD-Méthodes statistiques, 3éme édition, Dunod, Paris, (2000).

[9] GROSBRAS, J-M, méthodes statistiques des sondages, Economica, Paris, (1987).

[10] HAZIZA DAVID-Notes de cours échantillonnage STT2000, Université de Montréal, canada, (Automne 2010).

[11] HERNIAUX GERARD-Initiation aux sondages : I.U.T carrières de l'information, Paris, Masson, (1971).

[12] JEAN-JAEQUES DROESBEKE, FICHET BERNARD, TASSI PHILIPPE-les sondages, éditeurs. Economica, (1987).

[13] KAMOUN ABDESSALEM, KRIAA MOHAMED -Initiation aux techniques de sondage, Centre de publication universitaire, Tunis,(2006).

[14] PASQUIER.A -Eléments des probabilités et de théorie des sondages, Dunod, Paris, (1969).

[15] RYNKIEWICZ JOSEPH-Théorie des sondages,(7 novembre 2007).

[16] TYLLE YVES- Théorie des sondages : Echantillonnage et estimation en populations finies, Dunod, Paris, (2001).

ANNEXES

Table N°1

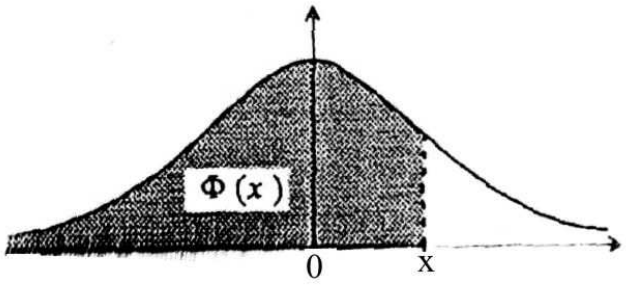
Fonction de répartition de la loi normale réduite

Si U suit la loi normale réduite, pour $x \geqslant 0$, la table donne la valeur :

$\phi(x) = P(U \leqslant x)$.

La valeur x s'obtient par addition des nombres inscrits en marge pour x<0, on a :

$\phi(x) = 1 - \phi(-x)$



x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500 0	0,504 0	0,508 0	0,512 0	0,516 0	0,519 9	0,523 9	0,527 9	0,531 9	0,535 9
0,1	0,539 8	0,543 8	0,547 8	0,551 7	0,555 7	0,559 6	0,563 6	0,567 5	0,571 4	0,575 3
0,2	0,579 3	0,583 2	0,587 1	0,591 0	0,594 8	0,598 7	0,602 6	0,606 4	0,610 3	0,614 1
0,3	0,617 9	0,621 7	0,625 5	0,629 3	0,633 1	0,636 8	0,640 6	0,644 3	0,648 0	0,651 7
0,4	0,655 4	0,659 1	0,662 8	0,666 4	0,670 0	0,673 6	0,677 2	0,680 8	0,684 4	0,687 9
0,5	0,691 5	0,695 0	0,698 5	0,701 9	0,705 4	0,708 8	0,712 3	0,715 7	0,719 0	0,722 4
0,6	0,725 7	0,729 1	0,732 4	0,735 7	0,738 9	0,742 2	0,745 4	0,748 6	0,751 7	0,754 9
0,7	0,758 0	0,761 1	0,764 2	0,767 3	0,770 4	0,773 4	0,776 4	0,779 4	0,782 3	0,785 2
0,8	0,788 1	0,791 0	0,793 9	0,796 7	0,799 5	0,802 3	0,805 1	0,807 8	0,810 6	0,813 3
0,9	0,815 9	0,818 6	0,821 2	0,823 8	0,826 4	0,828 9	0,831 5	0,834 0	0,836 5	0,838 9
1,0	0,841 3	0,843 8	0,846 1	0,848 5	0,850 8	0,853 1	0,855 4	0,857 7	0,859 9	0,862 1
1,1	0,864 3	0,866 5	0,868 6	0,870 8	0,872 9	0,874 9	0,877 0	0,879 0	0,881 0	0,883 0
1,2	0,884 9	0,886 9	0,888 8	0,890 7	0,892 5	0,894 4	0,896 2	0,898 0	0,899 7	0,901 5
1,3	0,903 2	0,904 9	0,906 6	0,908 2	0,909 9	0,911 5	0,913 1	0,914 7	0,916 2	0,917 7
1,4	0,919 2	0,920 7	0,922 2	0,923 6	0,925 1	0,926 5	0,927 9	0,929 2	0,930 6	0,931 9
1,5	0,933 2	0,934 5	0,935 7	0,937 0	0,938 2	0,939 4	0,940 6	0,941 8	0,942 9	0,944 1
1,6	0,945 2	0,946 3	0,947 4	0,948 4	0,949 5	0,950 5	0,951 5	0,952 5	0,953 5	0,954 5
1,7	0,955 4	0,956 4	0,957 3	0,958 2	0,959 1	0,959 9	0,960 8	0,961 6	0,962 5	0,963 3
1,8	0,964 1	0,964 9	0,965 6	0,966 4	0,967 1	0,967 8	0,968 6	0,969 3	0,969 9	0,970 6
1,9	0,971 3	0,971 9	0,972 6	0,973 2	0,973 8	0,974 4	0,975 0	0,975 6	0,976 1	0,976 7
2,0	0,977 2	0,977 8	0,978 3	0,978 8	0,979 3	0,979 8	0,980 3	0,980 8	0,981 2	0,981 7
2,1	0,982 1	0,982 6	0,983 0	0,983 4	0,983 8	0,984 2	0,984 6	0,985 0	0,985 4	0,985 7
2,2	0,986 1	0,986 4	0,986 8	0,987 1	0,987 5	0,987 8	0,988 1	0,988 4	0,988 7	0,989 0
2,3	0,989 3	0,989 6	0,989 8	0,990 1	0,990 4	0,990 6	0,990 9	0,991 1	0,991 3	0,991 6
2,4	0,991 8	0,992 0	0,992 2	0,992 5	0,992 7	0,992 9	0,993 1	0,993 2	0,993 4	0,993 6
2,5	0,993 8	0,994 0	0,994 1	0,994 3	0,994 5	0,994 6	0,994 8	0,994 9	0,995 1	0,995 2
2,6	0,995 3	0,995 5	0,995 6	0,995 7	0,995 9	0,996 0	0,996 1	0,996 2	0,996 3	0,996 4
2,7	0,996 5	0,996 6	0,996 7	0,996 8	0,996 9	0,997 0	0,997 1	0,997 2	0,997 3	0,997 4
2,8	0,997 4	0,997 5	0,997 6	0,997 7	0,997 7	0,997 8	0,997 9	0,997 9	0,998 0	0,998 1
2,9	0,998 1	0,998 2	0,998 2	0,998 3	0,998 4	0,998 4	0,998 5	0,998 5	0,998 6	0,998 6

Table N°2

Extrait de la table de nombres au hasard de M. G. Kendall et Babington Smith

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	14 55	61 40	03 47	70 64	42 47	23 28	93 85	76 46	96 55	51 04
2	27 34	17 89	78 39	96 49	50 22	56 25	48 06	84 42	21 48	50 82
3	13 92	63 26	15 86	71 31	80 34	80 52	19 28	00 30	88 00	07 76
4	23 86	30 95	80 76	26 02	15 77	35 69	25 81	97 51	10 70	01 55
5	12 52	17 61	34 42	69 37	93 30	35 40	85 26	18 50	61 44	31 84
6	70 00	44 52	95 70	71 71	97 01	25 73	42 70	85 14	04 38	28 77
7	01 43	12 14	84 74	07 01	86 04	36 89	74 22	42 95	28 62	34 81
8	64 96	69 50	79 22	72 03	25 36	24 69	00 13	22 84	49 82	50 89
9	34 48	73 08	62 08	17 30	99 56	00 84	09 72	67 99	04 42	27 28
10	64 68	16 86	75 63	06 40	45 22	47 92	61 85	52 35	71 63	92 73
11	39 08	84 40	14 88	58 52	04 51	46 98	33 56	14 63	85 35	03 48
12	24 84	01 27	62 21	74 23	37 11	03 66	68 00	01 14	22 07	75 19
13	33 99	80 71	61 02	75 17	40 82	43 16	39 83	45 09	12 37	78 22
14	30 02	19 47	92 98	07 77	61 62	58 26	27 33	37 82	43 81	17 56
15	07 34	12 04	54 18	32 79	09 68	40 35	35 89	25 44	31 71	97 88
16	43 03	11 35	32 45	66 61	30 80	90 53	50 74	13 88	81 31	20 83
17	15 01	18 74	93 19	94 29	84 26	14 67	74 15	38 76	26 94	44 97
18	64 09	93 89	67 41	39 72	83 15	22 65	49 63	66 86	94 99	81 30
19	39 61	04 31	29 63	25 13	99 49	11 20	68 03	19 61	12 42	43 28
20	85 37	28 81	79 66	56 75	63 25	30 59	58 85	51 22	17 41	94 33
21	49 87	98 30	28 29	94 13	46 11	20 61	03 57	29 52	68 00	98 27
22	24 45	58 33	29 99	29 55	31 12	87 14	18 03	03 32	31 39	98 81
23	52 25	19 01	74 33	75 38	54 69	69 83	07 58	83 58	02 17	59 75
24	53 57	39 44	04 74	08 43	66 88	46 50	71 55	90 72	76 28	88 17
25	79 05	38 56	47 32	08 07	04 77	27 10	34 49	57 71	79 82	90 69

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40
1	96 98	42 15	15 69	13 11	70 61	02 98	24 23	38 37	98 70	69 91
2	01 59	39 16	89 26	25 59	76 02	93 87	69 33	31 21	79 15	57 64
3	31 35	62 63	46 61	57 03	01 75	35 86	96 14	39 91	28 41	50 23
4	12 88	74 18	11 96	13 11	83 49	85 22	12 28	93 06	06 69	46 37
5	96 54	48 65	84 04	87 67	82 94	35 48	19 99	21 38	67 69	56 22
6	39 75	07 09	37 68	91 60	11 40	97 10	90 01	01 29	79 07	55 18
7	63 23	31 41	02 66	17 77	76 81	15 84	11 89	80 26	94 67	85 56
8	95 76	85 84	86 20	20 82	87 29	71 45	79 21	20 80	94 46	78 92
9	55 69	13 28	18 61	65 89	74 76	57 51	25 99	62 31	34 26	31 26
10	64 26	37 74	60 05	83 58	88 85	31 90	53 74	05 60	45 17	34 26
11	13 62	63 60	39 81	38 73	64 48	08 71	28 25	76 93	93 04	90 16
12	58 71	92 51	33 86	31 68	99 46	86 68	32 37	46 08	11 36	08 50
13	80 57	90 85	79 71	88 97	17 13	82 22	95 80	79 63	69 83	30 76
14	82 69	22 76	08 36	99 10	83 03	76 18	92 62	75 40	18 02	70 89
15	71 72	04 42	73 70	55 70	47 03	17 74	31 20	57 61	79 58	92 85
16	97 93	71 64	03 43	17 62	28 39	00 01	11 69	81 20	45 89	98 80
17	11 64	80 18	30 99	66 64	99 08	45 39	36 33	18 59	16 31	31 92
18	43 17	19 25	13 13	11 36	83 33	68 17	44 28	51 71	19 84	17 57
19	88 92	62 29	73 06	00 43	44 70	45 09	43 82	48 51	16 24	43 05
20	82 11	84 44	95 43	90 39	12 63	38 37	91 51	24 49	62 62	02 40
21	43 02	20 62	27 39	46 12	76 38	04 29	38 48	86 28	75 33	85 79
22	34 91	10 97	29 77	31 56	94 76	18 68	94 26	97 60	00 32	65 32
23	68 33	07 70	17 35	57 77	16 56	50 24	08 57	17 93	99 44	04 53
24	05 43	49 32	23 00	96 35	79 83	18 20	91 44	75 18	92 55	50 08
25	41 75	32 67	11 50	63 22	23 20	79 00	62 19	78 08	21 19	03 28

Table N°3

Table de nombres au hasard

22 719	92 549	10 907	35 994	63 461	83 659	24 494	53 825	97 047	79 069
17 618	88 357	52 487	79 816	74 600	50 436	88 823	19 806	33 960	30 928
25 267	35 973	80 231	60 039	50 253	63 457	97 444	13 799	35 853	03 149
88 594	69 428	66 934	27 705	51 262	63 941	77 660	66 418	84 755	29 197
60 482	33 679	03 078	08 047	39 891	34 068	81 957	02 985	83 113	36 981
30 753	19 458	02 849	30 366	83 892	80 912	91 335	41 703	79 401	97 251
60 551	24 788	35 764	57 453	06 341	10 178	91 896	70 619	46 440	98 356
35 612	09 972	98 891	92 625	70 509	95 484	34 858	13 499	28 966	88 287
[illegible]	18 448	45 922	55 179	18 412	31 186	91 017	37 410	76 512	79 361
73 998	97 374	66 685	06 639	34 500	17 935	79 544	15 475	74 765	11 199
14 973	68 806	49 122	16 124	61 905	22 047	17 229	46 703	39 727	16 753
[illegible]	48 382	[illegible]	[illegible]	51 686	15 537	[illegible]	[illegible]	91 783	92 825
37 868	82 916	73 732	63 230	85 306	56 988	[illegible]	98 029	42 208	00 190
61 666	48 114	95 183	02 628	05 355	97 627	74 554	91 267	31 240	34 723
56 638	70 054	19 427	24 811	37 164	71 641	50 515	88 231	99 539	75 745
43 973	07 496	17 405	08 966	65 989	68 017	56 975	94 080	93 689	98 889
05 540	72 301	36 504	00 187	90 375	22 891	22 205	27 777	84 803	39 220
95 141	07 885	94 399	41 145	50 210	92 423	13 303	09 621	94 153	18 691
75 954	68 499	42 308	38 387	52 163	64 563	02 813	45 577	93 125	25 294
97 905	05 301	98 496	20 682	[illegible]	68 537	[illegible]	78 282	02 396	10 002
[illegible]	37 782	67 537	38 813	00 377	98 873	[illegible]	[illegible]	25 457	28 716
[illegible]	[illegible]	63 187	46 191	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
32 251	06 804	85 959	20 974	[illegible]	[illegible]	[illegible]	[illegible]	24 721	04 187
62 361	59 105	39 338	59 358	69 193	15 586	57 695	80 518	59 788	04 215
54 954	90 337	99 346	60 442	90 933	58 323	83 183	90 011	44 236	90 815
70 773	03 331	84 228	01 405	61 494	72 064	24 713	39 851	01 431	60 841
68 702	08 331	08 923	83 173	67 081	87 472	47 980	08 802	95 495	78 745
39 599	33 465	96 705	41 458	34 670	55 385	25 484	71 068	15 155	85 371
54 958	34 935	16 858	16 523	54 262	63 310	50 348	53 457	39 440	80 411
98 124	08 864	36 485	78 766	52 802	56 315	43 523	06 513	50 899	86 432
43 099	88 373	80 091	35 058	35 755	47 556	98 602	71 744	70 442	92 312
88 667	44 515	80 435	17 140	32 588	98 708	93 010	98 590	23 656	85 664
87 009	95 736	76 930	71 090	27 143	95 229	24 799	02 313	17 436	20 273
70 581	40 618	16 631	54 178	44 737	02 544	81 368	08 078	46 740	52 583
03 723	25 551	03 816	97 612	99 833	06 779	47 619	12 901	60 179	23 780
19 943	30 139	07 932	29 267	01 934	19 584	13 356	35 803	90 284	97 565
71 559	30 728	83 499	65 977	37 442	72 526	53 123	99 948	59 762	19 952
25 400	16 143	79 028	81 790	57 747	87 972	54 981	10 079	17 490	15 215
59 894	59 543	13 668	27 197	51 979	38 403	23 989	38 549	92 968	53 309
29 757	26 942	08 736	15 184	73 650	51 130	59 167	89 866	06 030	88 929
87 650	08 162	90 596	70 312	84 462	07 653	80 962	96 692	07 030	62 470
84 094	70 059	86 833	23 531	31 749	23 930	04 763	89 322	67 576	38 627
92 101	17 194	06 003	99 847	12 781	38 729	88 072	92 589	61 828	36 504
26 641	99 088	65 294	37 138	75 881	12 627	19 461	69 536	64 419	82 106
04 920	91 233	46 959	14 735	15 153	28 306	76 351	28 109	86 078	45 234
25 417	97 570	91 045	09 929	75 140	43 926	90 282	99 088	93 605	03 547
98 874	96 989	84 371	87 624	74 090	71 983	62 424	62 130	44 470	74 725
02 127	82 060	84 618	58 572	56 716	79 862	59 896	50 702	31 938	18 336
[illegible]	[illegible]	[illegible]	[illegible]	34 139	27 631	[illegible]	91 504	77 617	[illegible]
76 176	[illegible]	[illegible]	84 361	63 898	97 863	[illegible]	24 931	87 983	[illegible]

Programmation :

Le logiciel SAS ***(Statistical Analysis System)*** offre, depuis peu, une procédure permettant d'effectuer un tirage systématique, sur fichier trié si on le souhaite ;la syntaxe est la suivante :

PROC SURVEYSELECT DATA=TAB M=SYS

SAMPSIZE=n; CONTROL VARTRI;

La table TAB contient les données, et n est la taille de l'échantillon, si on souhaite contrôler non plus la taille de l'échantillon mais le **PAS**, on remplace l'option SAMPSIZE par l'option RATE=r, ou **r** est l'inverse du **PAS**, c.-à-d. le taux de sondage.

L'instruction CONTROL est optionnelle et donne lieu au tri préalable de la base TAB selon la variable VARTRI.

Cela étant, lorsque la valeur du **PAS** est un entier, on peut programmer très simplement le tirage sans utiliser la procédure précédente : si on veut réaliser ce tirage avec un PAS égal par exemple à 150, et si la population est stockée dans un tableau TAB (éventuellement trié), on écrit :

```
DATA PARAM; INPUT à 1 BIDON I.; CARDS;
1
;
DATA PARAM; SET PARAM;
ALEA=1+INT (150* RANUNI (0));
CALL SYMPUT (′ALEA′,ALEA);
DATA SORTIE; SET TAB;
IF MOD (_N_-&ALEA, 150)=0;
```

Le tableau SORTIE contient l'échantillon voulu.

Si le **PAS** n'est pas entier, et si on dispose d'un fichier ou chaque enregistrement-individu i doit être sélectionné avec une probabilité n/N, on peut écrire dans un language adapté :

```
POINT=ALEA
I=1
DO WHILE I ≤ N
IF POINT≤ I*(n/N) THEN DO
SELECTION (I) = ″OUI″
POINT=POINT+1
ELSE DO
SELECTION (I) = ″NON″
END
I=I+1
END
```

ALEA est un nombre généré aléatoirement entre 0 et 1.