**REPUBLIQUE DE COTE D’IVOIRE**

**Union-Discipline-Travail**

**THEORIES DES SONDAGES**

***Janvier 2016***

\*\*\*\*

***NOTES DE COURS***

\*\*\*\*

**Ministère de l’Enseignement Supérieur et de la Recherche Scientifique**

****

**ECOLE NATIONALE SUPERIEURE DE STATISTIQUE ET D’ECONOMIE APPLIQUEE D’ABIDJAN**

**Ministère D’ETAT, MINISTERE DU PLAN ET DU DEVELOPPEMENT**

[Chapitre 1 : Généralités 5](#_Toc442467799)

[1.1- Erreurs d’observations et erreur de sondage 6](#_Toc442467800)

[1.2 Notions statistiques et vocabulaire de bases 6](#_Toc442467801)

[1.2.1 Individu 6](#_Toc442467802)

[1.2.2 Population 6](#_Toc442467803)

[1.2.3 Echantillon 7](#_Toc442467804)

[1.2.4 Base de sondage 7](#_Toc442467805)

[1.2.5 Variable d’intérêt 7](#_Toc442467806)

[1.2.6 Estimateur et estimation 8](#_Toc442467807)

[1.2.7 Le biais 8](#_Toc442467808)

[1.2.8 Le plan de sondage 8](#_Toc442467809)

[Chapitre 2 : Les méthodes empiriques d’échantillonnage 9](#_Toc442467810)

[2.1- Méthode des quotas 9](#_Toc442467811)

[2.2- La méthode des unités-types 13](#_Toc442467812)

[2.3- La méthode des itinéraires 13](#_Toc442467813)

[2.4- La méthode "boule de neige" 13](#_Toc442467814)

[2.5- Les enquêtes par téléphone 14](#_Toc442467815)

[2.6 Les enquêtes par lettre ou email 14](#_Toc442467816)

[NB : La méthode du volontariat 14](#_Toc442467817)

[Chapitre 3 : Les fondements des méthodes probabilistes de sondage 15](#_Toc442467818)

[3.1 Définition 15](#_Toc442467819)

[3.2 L’inégalité de Tchebytchev 15](#_Toc442467820)

[3.3 La loi des grands nombres 16](#_Toc442467821)

[3.4 Théorème central limite 17](#_Toc442467822)

[3.5 Intervalle de confiance 17](#_Toc442467823)

[3.6 Les grandes étapes d’une enquête par sondage 17](#_Toc442467824)

[Chapitre 4 : Sondage aléatoire simple 19](#_Toc442467825)

[4.1 Définition et principe du sondage aléatoire simple 19](#_Toc442467826)

[4.2 : Le tirage systématique de l’échantillon 19](#_Toc442467827)

[4.3 : Caractéristiques de la population et notations 20](#_Toc442467828)

[4.4 Expression des estimateurs dans le cas d’un SAS 21](#_Toc442467829)

[4.4.1 : Estimateur de la moyenne et du total 21](#_Toc442467830)

[4.4.2 : Variance des estimateurs de la moyenne et du total 21](#_Toc442467831)

[4.4.3 : Estimateur de et 23](#_Toc442467832)

[4.4.4 : Estimation des intervalles de confiance 24](#_Toc442467833)

[4.4.5 : Estimation d’une proportion 24](#_Toc442467834)

[4.5 : Détermination de la taille de l’échantillon 25](#_Toc442467835)

[4.5.1 : Détermination de la taille de l’échantillon par la méthode budgétaire 25](#_Toc442467836)

[4.5.2 : Détermination de la taille de l’échantillon par la méthode non budgétaire 25](#_Toc442467837)

[Chapitre 5 : Sondage stratifié 27](#_Toc442467838)

[5.1 Principe et justification du sondage stratifié 27](#_Toc442467839)

[5.2 Description de la population et notations 27](#_Toc442467840)

[5.3 Estimation 28](#_Toc442467841)

[5.4 Calcul de précision 28](#_Toc442467842)

[5.5 Taille de l’échantillon par strate 29](#_Toc442467843)

[5.5.1 : Allocation proportionnelle 29](#_Toc442467844)

[5.5.2 : Répartition de Neyman 29](#_Toc442467845)

[5.5.3 : optimalité des coûts 30](#_Toc442467846)

[Chapitre 6 : Sondage à probabilités inégales 31](#_Toc442467847)

[6.1 Principe et justification du sondage à probabilités inégales 31](#_Toc442467848)

[6.2 Probabilités d’inclusion 31](#_Toc442467849)

[6.3 Calcul des estimateurs dans le cas d’un tirage sans remise 32](#_Toc442467850)

[6.4 Méthodes de tirage de l’échantillon 33](#_Toc442467851)

[6.4.1. Tirage systématique à partir des totaux cumulés 33](#_Toc442467852)

[6.4.2. Tirage systématique à partir des probabilités cumulées 33](#_Toc442467853)

[Chapitre 7 : Le sondage à plusieurs degrés 34](#_Toc442467854)

[7.1. Principe et justification du sondage à plusieurs degrés 34](#_Toc442467855)

[7.2. Description de la population et notations 35](#_Toc442467856)

[7.3. Calcul et propriétés des probabilités d’inclusion des unités secondaires 36](#_Toc442467857)

[7.4. Calcul des estimateurs 36](#_Toc442467858)

[7.4.1. Cas du tirage des UP avec des probabilités égales 36](#_Toc442467859)

[7.4.2. Cas du tirage des UP à probabilités inégales 36](#_Toc442467860)

[7.4.3. Cas particulier du tirage du sondage auto-pondéré 37](#_Toc442467861)

[7.5. L’effet de grappe 38](#_Toc442467862)

[7.6. Détermination du nombre d’unités primaires 39](#_Toc442467863)

[Chapitre 8 : les redressements et corrections des non-réponses 40](#_Toc442467864)

[8.1.- Principe et justification des redressements 40](#_Toc442467865)

[8.2.- La post stratification simple 40](#_Toc442467866)

[8.3.- La post stratification sur critère multiple : la méthode du raking ratio 41](#_Toc442467867)

[8.4. Estimation par le quotient(ou par le ratio) 43](#_Toc442467868)

[8.5. Estimation par la régression 44](#_Toc442467869)

[8.6. Le traitement des non réponses 44](#_Toc442467870)

[8.6.1.- Les non réponses totales 45](#_Toc442467871)

[8.6.2.- Les non réponses partielles 45](#_Toc442467872)

[8.7.- Correction des non réponses 45](#_Toc442467873)

[8.7.1.- Correction des non réponses totales 45](#_Toc442467874)

[8.7.2.- Correction des non réponses partielles 47](#_Toc442467875)

[8.7.2.1- La méthode économétrique 47](#_Toc442467876)

[8.7.2.2- La méthode déductive 47](#_Toc442467877)

[8.7.2.3- La méthode du cold deck 47](#_Toc442467878)

[8.7.2.4- La méthode du hot deck 48](#_Toc442467879)

[8.7.2.5- La méthode d’imputation par prédicteur 48](#_Toc442467880)

# Chapitre 1 : Généralités

Une enquête par sondage est une technique d’observation qui permet de produire de l’information statistique sur une population donnée à partir de l’observation d’une partie seulement de cette population. Elle se justifie particulièrement dans l’étude de populations très nombreuses ou très dispersées pour lesquelles elle peut remplacer utilement les recensements, souvent très lourds et très onéreux et s’impose dès lors que l’observation conduit à la destruction de l’unité statistique.

Dans les pays développés où les techniques de sondages sont très couramment utilisés, avant chaque élection importante, les instituts de sondage présentent les prévisions des résultats du vote. De même, avant la publication des résultats définitifs, des estimations très précises des résultats sont effectués par des instituts de sondage. Les sondages « sortie des urnes » sont si proches de la réalité que les média les donne comme résultats quasi définitifs du vote.

Très souvent, les résultats du scrutin sont très proches des estimations. Cependant, dans des cas rares, les estimations s’en écartent et même significativement parfois. Ces cas « d’échecs » des sondages sont aujourd’hui exceptionnels. Dans le passé, ils se produisaient plus souvent puisque les précautions devant entourer la technique n’étaient pas toujours prise. Par exemple, en 1936, lors des élections présidentielles américaines, le candidat républicain était donné gagnant par une large majorité alors que finalement il obtint moins de 40% des suffrages. Le sondage était basé sur une liste tirée de l’annuaire téléphonique et parmi les lecteurs du quotidien « Literacy Digest » lu par l’intelligentsia américaine et donc très favorable aux républicains. De telles erreurs sont connues et évitées de nos jours. De nombreux pièges de ce genre doivent en effet être évités pour que le sondage donne des résultats fiables. En particulier, les statisticiens doivent répondre à une série de questions avant sa mise en œuvre :

* Quelle est la meilleure méthode de sondage à utiliser ?
* Combien d’individus faut-il enquêter pour avoir des résultats généralisables à l’ensemble de la population ? (taille de la population)
* Comment constituer l’échantillon (l’échantillonnage) ?
* Comment agréger les données pour avoir la meilleure information (calcul de l’estimateur) ?
* Quelle est la variabilité de cet estimateur, autrement dit quelle est la précision des estimations faites et donc leur crédibilité ? (variance, marge d’erreur) ?

En réalité, la technique des sondages est utilisée partout et beaucoup plus souvent qu’on ne le croit. En effet, les sondages sont utilisés dans des domaines aussi divers que :

* La médecine (essais cliniques) ;
* La biologie (prévalence d’une IST) ;
* La recherche de gisements pétroliers ou miniers ;
* L’agriculture (évaluation du volume de la production agricole nationale, estimation des superficies cultivées, etc…) ;
* La vérification d’une comptabilité ;
* Les contrôles fiscaux ;
* Les contrôles de la qualité des produits sur les chaînes d’usines ;
* Le contrôle de la qualité d’un recensement (enquête de couverture) ;
* Le contrôle anti dopage en sport ;
* Les préparations culinaires (la femme qui remue et goute sa sauce) ;
* Etc.

## Erreurs d’observations et erreur de sondage

Les données collectées lors des opérations de collecte sont toujours entachées d’erreur quel que soit le type d’opération retenue (recensements ou enquêtes). Il y en a deux types : l’erreur d’observation et l’erreur d’échantillonnage. Un recensement induit un seul type d’erreur : l’erreur d’observation commise lors de la collecte des données. Quant à l’enquête par sondage, l’erreur d’observation s’associe à celle due à l’échantillonnage.

Erreur totale

Erreur d’échantillonnage

Erreur d’observation

Erreur d’observation

**Recensement**

**Enquête par sondage**

## 1.2 Notions statistiques et vocabulaire de bases

### 1.2.1 Individu

Un individu est une unité statistique. Il peut s’agir de l’unité de base sur laquelle porte l’observation statistique. C’est alors l’unité d’observation. Il peut s’agir aussi de l’unité d’échantillonnage désignée lors de la constitution de l’échantillon pouvant fournir l’information recherchée sur l’unité d’observation (ex : unité d’échantillonnage=mère et unité d’observation= enfants de moins de un an). Les individus peuvent être de différentes natures.

### 1.2.2 Population

Encore appelée univers statistique, la population est l’ensemble sur lequel porte l’étude statistique. C’est l’ensemble des unités statistiques ou individus qui peut être un ensemble d’hommes et de femme mais aussi d’arbres, de parcelles, etc.

Quatre facteurs déterminent une population donnée :

* Sa nature (êtres humains, parcelles, entreprises, arbres, etc.)
* Ses caractéristiques (sexe, âge, surface, nombre d’employés, diamètre, etc.)
* Sa localisation (quelle ville, quelle localité, quelle usine, etc.)
* La date à laquelle la population est considérée.

Définir une population est une tâche difficile car il faut éviter les erreurs d’observations liées à la sous couverture ou au contraire à la sur couverture de la population.

Lorsqu’on donne la liste exhaustive d’une population, on la définit en extension. Aucune erreur n’est alors possible mais avoir une telle liste est parfois impossible. On peut définir également une population en compréhension. On donne alors une caractéristique commune aux unités statistiques. Cette technique est plus simple mais beaucoup plus périlleuse. Considérons par exemple la population des villas de Cocody. Comment distinguer une villa d’une simple maison ? Où commence et où fini Cocody ? Ainsi la définition précise d’une population est souvent difficile mais absolument nécessaire pour effectuer une bonne enquête.

### 1.2.3 Echantillon

Un échantillon d’un univers est un sous ensemble d’unités statistiques provenant de cet univers qui permet d’en étudier certaines caractéristiques.

La théorie des sondages permet de passer des résultats obtenus sur un échantillon à des estimations sur tout l’univers. Ce passage, appelé extrapolation ou inférence, obéit à des règles précises que détaille la théorie des sondages.

### 1.2.4 Base de sondage

La base de sondage est la liste complète des individus de l’univers statistique étudié et duquel l’échantillon doit être prélevé.

Elle est nécessaire pour mettre en œuvre un échantillonnage probabiliste pour lequel chaque individu a une probabilité connue, non nulle et fixée à l’avance, de faire partie de l’échantillon.

La base de sondage doit présenter certaines propriétés :

* Elle doit donner la liste de toutes les unités d’échantillonnage faisant partie du champ de l’enquête ;
* Elle doit comporter un identifiant permettant de repérer précisément et sans ambiguïté chaque unité statistique de l’univers ;
* Elle doit être à jour, c'est-à-dire sans omission.

### 1.2.5 Variable d’intérêt

La variable d’intérêt Y est une variable qui formalise l’information objet du sondage, elle est définie sur chaque unité statistique de l’univers.

Dans un échantillon, la variable d’intérêt collectée sur chaque individu permet de partir des observations yi faites dans l’échantillon pour estimer la valeur de cette variable sur l’ensemble de l’univers.

En général, l’on cherche à déterminer une fonction des Yα. Cette fonction, définie sur la population mère constitue un paramètre. C’est une valeur inconnue, mais bien déterminée dans la population. C’est la vraie valeur que l’on cherche à estimer. Ce paramètre peut être une moyenne, une variance, un total, etc.

### 1.2.6 Estimateur et estimation

Un estimateur d’une grandeur θ est un outil ou formule mathématique qui permet d’avoir une approximation c'est-à-dire une estimation de cette grandeur θ à partir de l’échantillon.

On notera l’estimateur de cette grandeur θ.

La valeur calculée à partir de en utilisant les données de l’échantillon fournit une estimation de θ.

### 1.2.7 Le biais

L’estimateur est une variable aléatoire ayant une loi de probabilité donnée. On peut donc calculer son espérance mathématique c'est-à-dire la valeur qu’elle prend « en moyenne ». L’écart s’appelle le biais de l’estimateur. C’est une première mesure de l’erreur que l’on commet en recourant au sondage plutôt qu’à un recensement c'est-à-dire l’erreur d’échantillonnage.

Un estimateur sans biais est un estimateur dont les résultats des estimations sont, en moyenne, égaux à la valeur que l’on cherche à estimer c'est-à-dire telle que le biais est nul. Pour un estimateur sans biais, on a donc

### 1.2.8 Le plan de sondage

Le plan de sondage est un document qui définit les modalités du tirage probabiliste de l’échantillon. Il précise :

* Le type de sondage probabiliste retenu et sa justification
* La taille de l’échantillon et les modalités de son calcul
* Les contraintes de la base de sondage
* Le taux de sondage
* La procédure adoptée pour le tirage de l’échantillon
* La formulation de l’estimateur.

C’est un document fondamental qui doit être rédigé pour une mise en œuvre correcte du sondage.

# Chapitre 2 : Les méthodes empiriques d’échantillonnage

Les méthodes empiriques d'échantillonnage, encore appelées méthodes de sondages à choix raisonné, sont l'ensemble des méthodes pour lesquelles les unités d’échantillonnage d'un univers sont choisies de manière à rendre l'échantillon représentatif.

Ces méthodes sont souvent utilisées lorsque l'on ne dispose pas de toute l'information nécessaire pour effectuer un autre type de sondage, quand l'investigateur ne se préoccupe pas particulièrement des erreurs d'échantillonnage ou bien lorsque le chercheur dispose de fonds particulièrement limités.

On choisit l’échantillon de telle sorte que les unités statistiques présentent globalement les mêmes caractères (du moins ceux que l’on connaît) que l’ensemble statistique dont elles sont issues.

Par même caractère, il faut entendre que l’échantillon présentera sur certaines variables connues statistiquement les mêmes proportions que l’ensemble des unités statistiques.

Sous réserve de faire preuve d'une grande objectivité et de faire preuve d'imagination, le sondage à choix raisonné donne des résultats de précision analogue au sondage basée sur des méthodes plus scientifiques.

Le principal avantage de ces méthodes est le moindre coût de revient de l'opération en comparaison des méthodes aléatoires de sondage. Ces méthodes ont cependant un inconvénient majeur : elles peuvent être critiquées par les personnes que les estimations faites n'arrangent pas car il est impossible de déterminer les marges d'erreur des estimations proposées.

### 2.1- Méthode des quotas

L’échantillon est une représentation à l’échelle réduite de l’univers statistique à étudier, du moins pour certains caractères (mêmes proportions des sous-populations dans l’échantillon que dans l’univers à étudier). L’échantillon doit être une image aussi fidèle que possible de la population mère, la population de référence du point de vue des caractères de contrôles qui ont servi à sa constitution.

La méthode suppose la connaissance de la structure de l'univers étudié en particulier la répartition des individus qui la composent selon des critères corrélés avec la variable d'intérêt, celle que l'on cherche à étudier. Ainsi, la méthode des quotas s’appuie sur la connaissance qu’on a de l’univers à étudier à partir d’une opération antérieure, un recensement par exemple.

Pour mettre cette méthode en œuvre, le chercheur calcule les différentes proportions des sous-populations de l'univers qui sont aussi celles de l'échantillon. A chaque enquêteur on alors donne le nombre d'individus à interroger ainsi que leur répartition par rapport aux critères. L'enquêteur a la latitude de choisir qui il veut du moment où il respecte les différentes répartitions. Il va probablement choisir la façon la moins pénible pour lui au risque de provoquer des distorsions graves au résultat. Ainsi, une trop grande liberté est laissée à l’individu chargé de prendre les quotas. Il faut donc lui imposer des contraintes supplémentaires (lieux d'enquêtes, date et heures, etc.)

La méthode demande une très bonne connaissance de certaines variables. De plus les caractères connus sur lesquels on s’appuie ne sont pas toujours en corrélation avec les caractères étudiés.

**Exemple** : On veut étudier le salaire moyen des travailleurs d'une très grande entreprise. L'on sait que le salaire est lié au sexe du travailleur, à son ancienneté et à sa catégorie socioprofessionnelle.

Un recensement antérieur d'une dizaine d'années plus tôt a donné les résultats suivants :

**Tableau 1**: Répartition des femmes selon l'ancienneté et la catégorie socioprofessionnelle

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| 0-5 ans | 1500 | 5000 | 20000 | 10000 |
| 6-10 ans | 3000 | 15000 | 15000 | 15000 |
| 11 ans et plus | 5000 | 7500 | 3500 | 5500 |

Tableau 2 : Répartition des hommes selon l'ancienneté et la catégorie socioprofessionnelle

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| 0-5 ans | 4500 | 2500 | 3000 | 5000 |
| 6-10 ans | 7000 | 5000 | 5000 | 6000 |
| 11 ans et plus | 8000 | 17500 | 6500 | 16500 |

On souhaiterait constituer un échantillon de taille 1000 selon la méthode des quotas.

L'entreprise vous fait appel pour l'aider à constituer l'échantillon. Que proposez-vous?

**Réponse**

On détermine tout d’abord les tableaux complets avec les totaux

**Tableau 1** : Répartition des femmes selon l'ancienneté et la catégorie socioprofessionnelle

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 1500 | 5000 | 20000 | 10000 | 36500 |
| 6-10 ans | 3000 | 15000 | 15000 | 15000 | 48000 |
| 11 ans et plus | 5000 | 7500 | 3500 | 5500 | 21500 |
| Total | 9500 | 27500 | 38500 | 30500 | 106000 |

**Tableau 2** : Répartition des hommes selon l'ancienneté et la catégorie socioprofessionnelle

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 4500 | 2500 | 3000 | 5000 | 15000 |
| 6-10 ans | 7000 | 5000 | 5000 | 6000 | 23000 |
| 11 ans et plus | 8000 | 17500 | 6500 | 16500 | 48500 |
| Total | 19500 | 25000 | 14500 | 27500 | 86500 |

Le pourcentage des femmes dans la population totale est :

Les femmes seront donc 55,1% dans l’échantillon soit un total de

Le pourcentage des hommes dans la population totale est :

Les femmes seront donc 44,9 % dans l’échantillon soit un total de

On détermine les différentes proportions des catégories parmi les femmes

**Tableau 3** : Répartition des femmes selon l'ancienneté et la catégorie socioprofessionnelle (en pourcentage)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 1,4 | 4,7 | 18,9 | 9,4 | 34,4 |
| 6-10 ans | 2,8 | 14,2 | 14,2 | 14,2 | 45,3 |
| 11 ans et plus | 4,7 | 7,1 | 3,3 | 5,2 | 20,3 |
| Total | 9,0 | 25,9 | 36,3 | 28,8 | 100,0 |

En appliquant ces différents pourcentages aux 551 femmes on obtient le tableau suivant

Tableau 4 : Répartition des femmes dans l’échantillon

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 8 | 26 | 104 | 52 | 190 |
| 6-10 ans | 16 | 78 | 78 | 78 | 249 |
| 11 ans et plus | 26 | 39 | 18 | 29 | 112 |
| Total | 49 | 143 | 200 | 158 | 551 |

On détermine les différentes proportions des catégories parmi les hommes

**Tableau 5**: Répartition des femmes selon l'ancienneté et la catégorie socioprofessionnelle (en pourcentage)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 5,2 | 2,9 | 3,5 | 5,8 | 17,3 |
| 6-10 ans | 8,1 | 5,8 | 5,8 | 6,9 | 26,6 |
| 11 ans et plus | 9,2 | 20,2 | 7,5 | 19,1 | 56,1 |
| Total | 22,5 | 28,9 | 16,8 | 31,8 | 100,0 |

En appliquant ces différents pourcentages aux 449 hommes on obtient le tableau suivant

**Tableau 6** : Répartition des hommes dans l’échantillon

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | Total |
| 0-5 ans | 23 | 13 | 16 | 26 | 78 |
| 6-10 ans | 36 | 26 | 26 | 31 | 119 |
| 11 ans et plus | 42 | 91 | 34 | 86 | 252 |
| Total | 101 | 130 | 75 | 143 | 449 |

### 2.2- La méthode des unités-types

Elle consiste à subdiviser la population à étudier en un certain nombre de sous-populations relativement homogènes. L'échantillon sera formé d'unités-types provenant de chacune de ces sous-populations.

Plus la population est homogène et plus la méthode permet d’avoir des résultats précis. Cette méthode présente aussi l'avantage d'être très rapide. Il faut avoir l’effectif des différents groupes pour pouvoir faire une bonne extrapolation sinon en avoir une bonne approximation.

### 2.3- La méthode des itinéraires

Elle consiste à indiquer à l'enquêteur, les itinéraires à suivre ainsi que les lieux où il devra s'arrêter pour faire une interview.

Cette méthode permet de prendre en compte les caractéristiques géographiques de l'univers (types de quartiers par exemple). Son coût est peu élevé et le problème des non-réponses peut être résolu facilement.

### 2.4- La méthode "boule de neige"

Elle est surtout appliquée pour constituer un échantillon basé sur une caractéristique relativement rare dans l'univers et pour laquelle on dispose de peu d'information quant à leur localisation.

L'idée est alors de partir d'un individu identifié et de retrouver certains autres possédant cette caractéristique à partir des indications que ce dernier donne à la demande de l'enquêteur.

On identifie au départ un individu appartenant à cette population. On l’enquête. On lui demande ensuite de donner le contact d’une autre personne du groupe que la personne enquêtée connaît. On va enquêter aussi cette personne et ainsi de suite. L’échantillon grandit au fur et à mesure que l’enquête progresse. L’enquête s’arrête lorsqu’on n’arrive plus à trouver quelqu’un d’autre à enquêter.

**Exemple** : Enquête sur les migrants de retour en Afrique de l'Ouest.

Il s'agissait de mener une investigation sur les migrants ivoiriens et ghanéens ayant choisi (ou ayant été contraints) de revenir au pays après avoir résidé en Europe ou en Amérique du Nord.

On retrouve un premier migrant de retour, on l'interroge. A la demande de l'enquêteur, il donne le contact d'une connaissance dans la même situation. L'enquête se poursuit donc au fur et à mesure des identifications et la taille de l'échantillon augmente comme une boule de neige qui roule.

Cette méthode peut amener des biais lors de l'extrapolation des résultats du fait que les personnes qui se connaissent ont plus certainement des profils semblables. Aussi faut-il partir de plusieurs individus ayant les profils les plus différents possibles. Il faut aussi veiller à ne pas avoir des doubles comptes.

### 2.5- Les enquêtes par téléphone

Ce type d'enquête préserve l'anonymat des répondants tout particulièrement le gène à répondre à certaines questions. Leurs coûts sont assez faibles généralement. Mais de sérieux problèmes de représentativité sont leurs principaux inconvénients surtout dans les pays en développement où une faible proportion de la population dispose d'un téléphone.

Cependant, cette dernière contrainte est en train d’être levée car le téléphone portable se généralise rapidement en Afrique. Bientôt, le problème de couverture pourrait être levé. Le sondage par téléphone sera alors une méthode très prometteuse car les bases de sondage existent chez les opérateurs ainsi que certaines caractéristiques des abonnés pouvant aider à un ciblage spécifique.

### 2.6 Les enquêtes par lettre ou email

Les enquêtes par lettre ou email préservent aussi l’anonymat des personnes enquêtées. Le principal problème peut être la représentativité de l’échantillon. En effet, le contexte peut être tel que de nombreux individus ne peuvent être joints par lettre. C’est le cas dans nombre de pays africains. De même, tout le monde ne possède pas encore une adresse émail et d’autres n’ont même pas accès à l’ordinateur ou à l’internet. Certains enfin sont analphabètes.

Dans le cas où l’enquête porte sur un groupe spécifique où ces contraintes n’existent pas (des entreprises, des ONG, etc.), les taux de non réponse sont souvent trop importants.

### NB : La méthode du volontariat

Lors d’une enquête par sondage ou un recensement, l’enquêteur a le libre choix de participer ou non à l’opération. En principe, en dehors des enquêtes de type administratives ou financière, notamment les enquêtes fiscales, etc., l’enquêté peut refuser de se prêter à l’enquête.

En particulier, l’éthique exige le volontariat pour certaines enquêtes sensibles telles que les enquêtes sur le VIH/SIDA. L’enquêteur n'interroge que les personnes qui veulent bien participer à l'enquête. On peut alors avoir des taux de refus exceptionnellement élevés. Dans ce cas, on peut avoir un biais de sélection de l'échantillon dont il convient de tenir compte lors de l’extrapolation des résultats.

# Chapitre 3 : Les fondements des méthodes probabilistes de sondage

## 3.1 Définition

Les méthodes de sondage aléatoire sont caractérisées par le fait que l’échantillon est constitué de manière à ce que chaque unité d’échantillonnage ait une probabilité connue mais non nulle d’appartenir à l’échantillon.

Elle se fonde essentiellement sur les deux théorèmes importants que sont l’inégalité de Tchebychev et la loi des grands nombres.

## 3.2 L’inégalité de Tchebytchev

Soit une variable aléatoire X quelconque d’espérance mathématique m et de variance σ². Quelle est la probabilité P pour que la variable aléatoire X appartienne à l’intervalle I = [m-tσ, m+tσ] avec t > 0 ?

On suppose que X est discrète (sans perte de généralité)

Soit E l’ensemble des indices des points extérieurs à I et F l’ensemble des indice des points intérieurs à I.

On a

C’est l’inégalité de Tchébicheff

## 3.3 La loi des grands nombres

Soit une urne contenant N boules dont NB boules blanches avec p=.= proportion de boules blanches dans l’urne.

On tire au hasard n boules. Soit X la variable aléatoire égale au nombre de boules blanches tirées.

Soit Fn= égale à la proportion de boules blanches tirées.

On suppose que les tirages sont effectués avec remise. Alors X suit une loi binomiale de paramètres n et p.

On a : E(X)=np et V(X)=npq

E(Fn)=(=

σ²=V()=

NB : Dans le cas d’un tirage avec remise, X suit une loi hypergéométrique

et

d’où

En appliquant l’inégalité de Tchebicheff, on a :

Ainsi, en choisissant n suffisamment grand, on peut faire en sorte que Fn soit aussi proche que l’on veut de p. C’est la loi des grands nombres.

De la même manière, on a :

Si X1 , X2 , X3 …, Xn sont n variables indépendantes suivant une loi de probabilité quelconque d’espérance mathématique m et d’écart type σ.

La moyenne est une variable aléatoire d’espérance mathématique m et d’écart type on a donc, en appliquant l’inégalité de Bienaymé-Tchebicheff

## 3.4 Théorème central limite

Si (Xn ) est une suite de variables aléatoires indépendantes et de même loi admettant des moments d’ordre un et deux notés m=E(Xn ) et σ²=V(X), alors,

(Converge en loi vers la loi normale centrée réduite)

C’est en partie en raison de ce théorème fondamental que la loi normale est si importante en statistique.

## 3.5 Intervalle de confiance

Soit un estimateur sans biais de θ, d’écart type qui suit une loi normale. On a :

Ainsi, pour α ϵ] 0,1 [, le fractile d’ordre 1- de la loi normale est telle que la probabilité que U soit comprise dans l’intervalle . Comme est un estimateur sans biais de θ, Ainsi,

Ainsi, pour α=5%, la probabilité que la vraie valeur du paramètre soit comprise dans un intervalle I tel que I=[ , ] est 0,95.

C’est pour cela qu’il est important, après avoir déterminé l’estimateur, de calculer son écart type afin de calculer l’intervalle de confiance.

## 3.6 Les grandes étapes d’une enquête par sondage

Pour réaliser une enquête par sondage, on procède par différentes étapes.

**1ère étape : conception du sondage**

A cette étape, on écrit le protocole de l’enquête. C’est un document qui pose la problématique de l’enquête, en donne les objectifs, définit le champ de l’enquête c’est-à-dire la population d’étude, la variable d’intérêt, etc. Il précise aussi la méthodologie et la taille de l’échantillon pour atteindre la précision souhaitée, le calendrier et le budget de l’opération. Le plan de sondage, la base de sondage et ses imperfections éventuelles et le mode de collecte de l’échantillon sont les principaux points de cette méthodologie.

**2ème étape : conception du questionnaire**

Elle se fait sur la base des objectifs fixés par l’enquête. Un module est en général dédié à chaque objectif spécifique. Les différents manuels doivent être rédigés pour aider à la collecte des données. Le questionnaire doit être testé ainsi que toute la méthodologie de l’enquête lors d’une enquête pilote.

**3ème phase : préparation et mise en œuvre de l’opération de collecte**

Elle consiste à rédiger et envoyer des lettres administratives et à faire des missions d’exploration ainsi que la publicité de l’opération.

La phase de collecte consiste à recueillir les données sur le terrain en tenant compte des enseignements de l’enquête pilote.

**4ème phase : Traitement des données et publication des résultats**

Cette phase consiste à codifier, saisir et apurer les données puis à rédiger le document puis à le diffuser.

**5ème phase : Archivage des données et des autres documents**

L’archivage est en principe la dernière phase de l’opération. Elle consiste à classer et ranger convenablement les documents papiers mais surtout à une sauvegarde électronique des outils de collecte, des programmes, des rapports de terrain, des rapports d’analyse et de tout autre document en rapport avec l’enquête.

# Chapitre 4 : Sondage aléatoire simple

## 4.1 Définition et principe du sondage aléatoire simple

Le SAS est une méthode de sondage qui consiste à tirer directement dans une population de taille N et sans tenir compte d’aucune information auxiliaire, un échantillon de taille n de façon à ce que chaque individu de la population ait la même probabilité d’appartenir à l’échantillon.

Cette probabilité, appelée probabilité d’inclusion, a pour valeur

La grandeur est le taux de sondage. Le SAS est la méthode de sondage la plus simple. Il n’exige que la disponibilité de la base de sondage.

Cette méthode est recommandée lorsque les informations à disposition laissent supposer que la dispersion des individus est faible relativement à la variable d’intérêt.

## 4.2 : Le tirage systématique de l’échantillon

Toute méthode de tirage garantissant la condition du SAS c’est-à-dire donnant la même probabilité d’inclusion peut être utilisée pour la constitution de l’échantillon.

La méthode de tirage la plus utilisée est le tirage systématique. Son algorithme de tirage est le suivant.

1°) On numérote les individus de 1 à N

2°) On calcule le pas de tirage

3°) On tire de façon aléatoire un nombre compris entre 1 et p (1 et p inclus). Soit u1 ce nombre. Ce sera le numéro du premier individu tiré.

4°) On retrouve les autres numéros des individus tirés en ajoutant le pas p au dernier numéro tiré.

Ainsi u2 = u1 +p ;

u3 =u2 +p= u1 +2p ;

u4 =u3 +p= u1 +3p ;

…

Un =un-1 +p= u1 +(n-1)p ;

**Exemple**

On désire tirer 20 individus parmi 100. Le pas est p=5.

On tire un n ombre au hasard entre 1 et 5. Soit par exemple 4. C’est le premier individu tiré.

On ajoute le pas 5 à 4 pour trouver le numéro du second individu (soit 9). Les numéros des autres individus tirés se retrouvent en ajoutant à chaque fois p=5 au dernier numéro obtenu. La liste finale est : 4 ; 9 ; 14 ; 19 ; 24 ; 29 ; 34 ; 39 ; 44 ; 49 ; 54 ; 59 ; 64 ; 69 ; 74 ; 79 ; 84 ; 89 ; 94 ; 99. Ce sont les numéros des individus de l’échantillon.

## 4.3 : Caractéristiques de la population et notations

La variable d’intérêt sera notée Y.

Les lettres en majuscule désigneront les caractéristiques de la population mère et les lettres minuscules seront réservées aux caractéristiques de l’échantillon. Les lettres grecques en indice sont réservées pour les la population générale. Les lettres latines en indice sont réservées à l’échantillon.

Ainsi, Yα désigne la valeur de Y pour l’individu numéro α au sein de la population mère et yi la valeur de Y pour l’individu numéro i appartenant à l’échantillon. N est la taille de la population et n la taille de l’échantillon. L’échantillon tiré sera désigné par .

On a alors les formules suivantes :

Moyenne dans la population

Total dans la population

Variance dans la population

Variance corrigée dans la population

Ainsi,

Pour l’échantillon on a respectivement :

Moyenne dans l’échantillon

Total dans l’échantillon

Variance dans l’échantillon

Variance corrigée dans l’échantillon

et,

## 4.4 Expression des estimateurs dans le cas d’un SAS

### 4.4.1 : Estimateur de la moyenne et du total

La moyenne de Y dans l’échantillon tiré est un estimateur sans biais de . En effet, soit

, Notons que

On a :

Ainsi, l’estimateur est un estimateur sans biais du total T. En effet,

Ainsi, dans l’expression de l’estimateur du total, le poids d’un individu de l’échantillon vaut . C’est le coefficient d’extrapolation ou de pondération.

### : Variance des estimateurs de la moyenne et du total

**4.4.2.1 Calcul de**

Soit Comme on le sait, et donc

V(=

Cov(

Calculons maintenant

Or,

D’où

Comme

Ainsi, on a

Et par suite

Ainsi, est d’autant plus faible (et donc l’estimation d’autant plus précise) que :

- S² et donc V(Y) est faible

- n est important

- f proche de 1

**4.4.2.2 Calcul de**

On déduit de

Comme ,

et ne sont pas connues puisque S² n’est pas connu. Il faut donc les estimer à leur tour.

### 4.4.3 : Estimateur de et

Pour trouver il suffit de trouver un estimateur de S².

Un estimateur sans biais de S² est s².

En effet,

On a d’une part

Comme

D’autre part, on a:

D’où finalement,

Ainsi,

s² est donc un estimateur sans biais de S²

Nous pouvons donc écrire, en remplaçant S² par s² dans :

### 4.4.4 : Estimation des intervalles de confiance

L’intervalle de confiance réel à 95% pour est :

Celui de T est :

Comme S² est inconnu, on ne peut calculer ces intervalles de confiance. On les estime donc en remplaçant S² par s², son estimateur sans biais.

Les estimateurs de l’intervalle de confiance de à 95% pour et T sont respectivement :

et :

### 4.4.5 : Estimation d’une proportion

Il est très fréquent d’estimer une proportion dans une population. On utilise alors une variable indicatrice Yα qui vaut 1 si l’individu α appartient à la sous population considérée et 0 sinon. Si N’ est l’effectif de cette sous population, on a :

Ainsi, estimer P revient à estimer et donc :

On montre, dans le cas d’une proportion que et donc

Comme , en remplaçant S² par s² dans V(p), on a l’expression de l’estimateur de V(p) soit :

En général, si n est grand et négligeable, on fait l’approximation

### 4.5 : Détermination de la taille de l’échantillon

Deux méthodes existent pour la détermination de la taille de l’échantillon : la méthode budgétaire et la méthode basée sur la précision.

### 4.5.1 : Détermination de la taille de l’échantillon par la méthode budgétaire

Soit B le budget disponible pour l’enquête, CF l’ensemble des coûts fixes de l’opération (location des salles, location des véhicules, voyages préliminaires, préparatifs de l’enquête, etc.), Cv l’ensemble des coûts variables de l’opération, c le coût unitaire pour enquêter un individu et n la taille de l’échantillon.

On a la relation suivante :

B= CF+Cv = CF + nc d’où

On applique cette méthode lorsque la contrainte budgétaire est forte et que l’on n’est pas guidé par le souci de la précision.

### 4.5.2 : Détermination de la taille de l’échantillon par la méthode non budgétaire

On souhaite que la valeur inconnue appartienne dans 95% des cas à un intervalle du type

Avec la condition

On a :

Si on considère qu’il est négligeable. La formule de détermination de la taille n devient encore plus simple.

On

Lorsque l’erreur est donnée sous la forme ε=α, on a, en négligeant ,

avec CV=coefficient de variation

En réalité, on ne connaît ni S² ni CV. On ne peut même pas les estimer puisque l’on ne dispose pas encore d’un échantillon, étant encore au stade du calcul de la taille nécessaire de l’échantillon. On est donc obligé de faire des approximations à partir des connaissances antérieures que l’on a sur la variable d’intérêt ou sur toute autre variable corrélée à cette dernière.

Dans le cas d’une proportion, on a :

La formule simplifiée donne

En prenant P=0,5, on maximise n. On prend cette valeur maximisée de n lorsqu’aucune approximation fiable de P n’existe.

# Chapitre 5 : Sondage stratifié

## 5.1 Principe et justification du sondage stratifié

Le sondage aléatoire simple est approprié lorsque la variance de la variable d’intérêt est faible. Que faire alors lorsque l’on pense que la population à étudier est très hétérogène du point de vue de la variable d’intérêt Y ? Cette préoccupation a fait naître d’autres techniques de sondage dont le sondage stratifié.

L’idée de la stratification est relativement simple. Comme les individus sont très hétérogènes, il convient de constituer des classes relativement homogènes au sein desquelles le SAS donnerait des résultats intéressants. Stratifier une population consiste à la répartir avant le tirage de l’échantillon en un certain nombre de sous ensembles homogènes par rapport à certains caractères connus a priori et qui sont corrélés avec la variable Y étudiée. Ces sous ensembles s’appellent des strates. Le tirage dans les strates s’effectue de manière indépendante.

La stratification répond à plusieurs autres exigences

* Augmenter la précision d’ensemble
* Avoir une bonne précision au niveau de chaque strate
* Obtenir des estimations par strate en vue de comparaison des strates
* Appliquer des méthodes d’estimation différentes selon les strates (ex : population nomade et population sédentaire)

## 5.2 Description de la population et notations

La population mère est de taille N. Elle est scindée en H strates de tailles respectives avec

On considère le cas où le tirage dans chaque strate se fait selon un SAS.

On tire un échantillon de taille de taille dans la strate h et on a , taille de l’échantillon global.

= taux de sondage dans la strate h

= taux de sondage global

= moyenne dans la strate h

= moyenne globale

= variance corrigée dans la strate h

= variance corrigée dans l’ensemble de la population

On a :

Pour grand pour tout h () on a :

Comme , S² est la somme de deux moyennes pondérées

= moyenne pondérée des dispersions à l’intérieur des strates, la dispersion intra strate et

= moyenne pondérée des carrés des écarts entre la moyenne des strates et la moyenne globale. C’est la dispersion inter strate.

## 5.3 Estimation

On a :

= estimateur sans biais de puisqu’on effectue le SAS dans chaque strate

= estimateur sans biais de

En effet,

= estimateur sans biais du total global T.

= coefficient d’extrapolation dans la strate h

## 5.4 Calcul de précision

Comme ,

Ainsi,

= intervalle de confiance réel au seuil de confiance

= intervalle de confiance estimé au seuil de confiance

## 5.5 Taille de l’échantillon par strate

Il existe plusieurs manières de répartir l’échantillon global entre les strates.

### 5.5.1 : Allocation proportionnelle

On dit que le sondage stratifié est à allocation proportionnelle lorsque le taux de sondage par strate est le même que le taux de sondage global c'est-à-dire lorsque la taille de l’échantillon est telle que

Dans ce cas de figure, l’exploitation des données est plus facile puisque la moyenne de l’échantillon global devient un estimateur sans biais de la moyenne globale. En effet, on a :

Ce type de sondage est dit auto pondéré.

Rappelons que pour le sondage stratifié on a dans le cas général

Dans le cas d’un sondage à allocation proportionnelle, on a et

On montre que la variance obtenue lors d’un sondage à allocation proportionnelle est liée à la variance obtenue dans un SAS à partir d’un échantillon de même taille. Cette relation est

### 5.5.2 : Répartition de Neyman

La répartition de Neyman est une répartition correspondant à celle d’un sondage à allocation proportionnelle à la dispersion dans chaque strate. On a :

Plus une strate est hétérogène et plus on y utilise un taux de sondage important. On utilise la répartition de Neyman lorsque le phénomène étudié a une distribution très dissymétrique (Ex : exploitations agricoles de petites tailles en grand nombre et quelques grandes exploitations concentrant une partie importante de la production).

### 5.5.3 : optimalité des coûts

Le coût de l’enquête est un critère important que l’on doit prendre en compte dans l’allocation des tailles entre les strates.

Soit CV= ensemble des coûts variables de l’enquête.

Si ch est le coût unitaire par strate, on a :

Le programme à résoudre devient le suivant

La solution est

# Chapitre 6 : Sondage à probabilités inégales

## 6.1 Principe et justification du sondage à probabilités inégales

Dans le sondage aléatoire simple, les unités statistiques ont toutes les mêmes probabilités d’appartenir à l’échantillon. Cette situation peut ne pas être optimale en particulier lorsque l’on dispose d’informations auxiliaires accréditant la thèse d’une meilleure observation disponible au niveau de certaines unités.

Contrairement à ce que l’intuition peut suggérer, on peut se donner des probabilités d’inclusion différentes pour les unités d’échantillonnage. C’est d’ailleurs ce qui a été fait pour le sondage stratifié pour lequel l’on peut se donner des taux de sondage différents par strate ( = probabilité d’inclusion dans la strate h). Ce faisant, on est obligé de tenir compte de cela dans le choix de l’estimateur : un individu ayant plus de chance d’appartenir à l’échantillon devra avoir un poids moindre dans l’estimateur.

Intuitivement, plus une unité est de grande taille, plus il est intéressant de pouvoir l’inclure dans l’échantillon puisqu’il est susceptible d’apporter plus d’information. Par exemple, dans une enquête auprès des exploitants agricoles d’une région, on peut avoir intérêt à les tirer avec des probabilités inégales, proportionnelles à leurs tailles. De même lorsqu’on étudie le volume d’une production, on veillera à ce que les entreprises de grandes tailles aient plus de chance d’appartenir à l’échantillon.

## 6.2 Probabilités d’inclusion

On appelle probabilité d’inclusion de l’unité statistique α la probabilité πα que cette unité fasse partie de l’échantillon. Elle est aussi appelée probabilité de sélection d’ordre 1. C’est un nombre compris entre 0 et 1 choisi de manière à ce que l’on ait .

La probabilité notée , que les unités α et β appartiennent simultanément à l’échantillon est appelée probabilité d’inclusion double ou probabilité de sélection d’ordre 2. Elle dépend du mode de tirage.

Dans le cas d’un SAS, on a :

Dans un sondage à probabilités inégales, les sont choisies par le sondeur de manière à respecter la relation

Pour un tirage de taille fixe n.

On montre que quelle que soit la méthode de tirage utilisée on a :

dépend elle du mode de tirage

## 6.3 Calcul des estimateurs dans le cas d’un tirage sans remise

On utilise l’estimateur pour estimer le total T. C’est un estimateur sans biais de T.

On estime la moyenne par

Ainsi, chaque individu de l’échantillon est pondéré par l’inverse de la probabilité d’inclusion.

est appelé estimateur de Horvitz-Thompson ou estimateurs des sommes dilatées.

NB : Aucune des probabilités d’inclusion ne doit être nulle.

Considérons la variable Y telle que

. On a :

est estimateur sans biais du total N, taille de la population mère.

NB : Si on avait , serait égale à 0 et on aurait eu un sondage parfait qui prévoit la valeur exacte du paramètre recherché. C’est le cas lorsque (Ainsi, )

C’est pourquoi on choisit des à peu près proportionnelles aux

On prend donc une variable auxiliaire X très corrélée à Y et n étant la taille de l’échantillon.

## 6.4 Méthodes de tirage de l’échantillon

### 6.4.1. Tirage systématique à partir des totaux cumulés

L’algorithme est le suivant :

1°) On numérote les unités de 1 à N

2°) On fait le cumul des tailles des unités. Soit C le cumul total

3°) On calcule le pas

4°) on génère un nombre aléatoire x entre 1 et p (1 et p compris)

5°) On situe x dans la colonne des cumuls et on retient l’unité i correspondant à x.

6°) On ajoute à chaque fois le pas pour situer et choisir les autres unités de l’échantillon.

### 6.4.2. Tirage systématique à partir des probabilités cumulées

L’algorithme de tirage est le suivant :

1°) On numérote les unités de 1 à N

2°) On calcule les probabilités d’inclusion

3°) On fait le cumul des probabilités d’inclusion.

4°) On génère un nombre aléatoire u entre 0 et 1

5°) on génère les termes ui = u +(i-1), i=1 à n

6°) On repère ces termes dans la colonne des probabilités d’inclusion cumulées pour avoir les unités correspondantes

# Chapitre 7 : Le sondage à plusieurs degrés

## 7.1. Principe et justification du sondage à plusieurs degrés

Lorsqu’on souhaite effectuer un sondage, on ne dispose pas toujours des éléments nécessaires pour faire le sondage dans de bonnes conditions. L’un des éléments essentiels et dont l’absence interdit le sondage probabiliste est la base de sondage. En Afrique, elle souvent absente ou trop vieille pour pouvoir être vraiment utile pour la constitution de l’échantillon. Le sondage à plusieurs degrés permet d’aller de l’avant lorsqu’on ne dispose pas de la base actualisée des unités qu’on souhaite échantillonner mais plutôt de la liste exhaustive d’autres unités plus grandes contenant les unités à échantillonner.

Le sondage à plusieurs degrés consiste à tirer d’abord un échantillon des grandes unités contenant les unités à échantillonner et dont on dispose de la base de sondage. Ce sont les unités primaires dans lesquelles on recense les unités qui intéressent le sondeur (unités secondaires). La base de sondage ainsi constituée (des unités secondaires) peut maintenant servir à tirer l’échantillon souhaité au départ. Un sondage peut également se faire à trois ou quatre et même plus de degrés.

Par exemple si l’on souhaite faire une enquête agricole mais qu’on ne dispose pas de base de sondage actualisée des exploitations agricoles. On peut dans un premier temps tirer des départements (Unités primaires) puis des villages dans les départements tirés (unités secondaires) et enfin des exploitations dans les villages tirés (unités tertiaires).

L’avantage du sondage à plusieurs degrés est que ce type de sondage est moins coûteux que par exemple le sondage aléatoire simple puisque, chercher à constituer directement la base de sondage de toutes les unités à échantillonner peut être financièrement dissuasif. C’est même souvent impossible puisque on n’a pas toujours le temps ni les ressources humaines d’entreprendre une telle opération. De plus, lorsque l’on tire des unités primaires, le travail de recensement des unités secondaires est plus facile puisque le nombre d’unités à dénombrer devient raisonnable et que le travail à faire reste concentré géographiquement ce qui facilite les choses.

Cette proximité géographique des unités et le fait que toutes les unités ne sont pas concernées par le tirage à partir du deuxième tirage font le sondage aléatoire simple est plus précis que le sondage à plusieurs degrés pour une même taille de l’échantillon. Cependant, pour un budget identique, on peut arriver à avoir une taille d’échantillon plus importante avec le sondage à plusieurs degrés qu’avec le sondage aléatoire simple de sorte que le sondage à plusieurs degrés peut se révéler plus intéressant du point de vue de la précision obtenue.

## 7.2. Description de la population et notations

On se place dans le cas d’un sondage à deux degrés. On adopte les notations suivantes :

N = taille de la population

M = nombre des unités primaires (UP)

m = nombre des unités primaires tirés

Nh = nombre d’unités secondaires dans l’unité primaire h, h=1 à M

n = taille de l’échantillon (nombre des unités secondaires tirés)

nh = nombre d’unités secondaires tirés dans l’unité primaire h

= taille moyenne des UP en unités secondaires

= taille de l’échantillon global

= taux de sondage au premier degré

= taux de sondage au second degré dans l’unité primaire h

= taux de sondage global

= total de Y dans l’unité primaire h

= total de Y dans la population globale

= total moyen de Y par unité primaire

moyenne de Y dans l’unité primaire h

= moyenne de Y dans la population globale

variance corrigée ou dispersion des totaux dans la population globale

=variance corrigée ou dispersion de Y dans l’échantillon tiré dans l’unité primaire h.

## 7.3. Calcul et propriétés des probabilités d’inclusion des unités secondaires

Soit la probabilité d’inclusion de l’unité primaire h et la probabilité d’inclusion d’une unité secondaire α sachant que l’unité primaire h à laquelle elle appartient est tirée.

La probabilité d’inclusion de l’unité α de l’unité primaire h est :

La probabilité d’inclusion des unités secondaires α et β des unités primaires h et l est :

## 7.4. Calcul des estimateurs

Dans la pratique, on a deux types de tirage des unités primaires : le tirage à probabilités égales et le tirage à probabilités inégales proportionnelles à leurs tailles. Les unités secondaires sont tirées avec des probabilités égales et sans remise.

### 7.4.1. Cas du tirage des UP avec des probabilités égales

On applique le tirage systématique à la sélection des UP et des US. Dans ce cas on a :

et

L’estimateur du total T est :

Et

### 7.4.2. Cas du tirage des UP à probabilités inégales

Lorsque les UP sont de tailles très inégales, il est conseillé de tirer les UP avec remise proportionnellement à leur taille. L’estimateur du total T est :

Avec

Ah = probabilité de l’unité primaire h d’être tiré à chaque tirage et

Et

### 7.4.3. Cas particulier du tirage du sondage auto-pondéré

Pour avoir un sondage auto-pondéré dans lequel on n’a plus besoin de poids de sondage, il faut, dans un sondage à deux degrés, il faut procéder par deux étapes.

On tire d’abord les UP avec des probabilités proportionnelles à leurs tailles en unités d’observation () dans un premier temps.

Dans ce cas, l’estimateur du total est égal à

Et est un estimateur sans biais de la moyenne

L’estimateur de de la variance de l’estimateur de la moyenne est donné par

Dans un second temps, on tire le même nombre d’unités secondaires dans les UP tirées (). La taille de l’échantillon est alors n=mn0

Les probabilités de sélection des unités ont alors pour expression

Par conséquent on a :

La moyenne est estimée directement par la moyenne dans l’échantillon.

NB : En réalité, la taille des UP se révèle être différente de Nh au moment du dénombrement. Ainsi ; si N’h est la taille de l’UPh au moment du dénombrement, on a en réalité

Il faudra donc en tenir compte en choisissant US dans l’unité primaire h.

### 7.5. L’effet de grappe

Dans un sondage à plusieurs degrés, les réponses des personnes interrogées ont tendance à se ressembler du fait de leur proximité géographique et probablement socio-économique. L’effet de grappe traduit le fait que cette similarité réduit la quantité d’information contenue dans les réponses données par les répondants et qu’il faille en tenir compte pour améliorer la variance de l’estimateur.

Il peut être mesuré par un coefficient appelé coefficient de corrélation intra-grappe encore appelé effet de grappe. Son expression dans un plan de sondage à deux degrés est le suivant :

Le plus souvent est strictement positif, traduisant une similarité des individus d’une même UP.

On montre que, dans un plan de sondage à deux degrés sans stratification ni tirage à probabilités inégales, si toutes les UP sont de même taille et qu’on choisit un même nombre d’unités secondaires par UP tirée, on a :

La grandeur Deff= appelée effet de sondage, permet de corriger la perte de variance introduite par l’effet de grappe.

Lorsqu’on néglige le taux de sondage dans un SAS, on a :

On donne en général une estimation de l’effet de grappe à partir d’enquêtes similaires organisées antérieurement.

### 7.6. Détermination du nombre d’unités primaires

La détermination du nombre d’unités primaires tient compte des moyens dont on dispose pour l’enquête. Soit c1 le coût fixe par UP et c2 le coût variable par UP. Le coût de l’enquête est :

On minimise la variance sous la contrainte du budget. Les solutions sont :

Et

# Chapitre 8 : les redressements et corrections des non-réponses

### 8.1.- Principe et justification des redressements

Il y a presque toujours des variables quantitatives qui existent sur la population mère, disponibles pour tous les individus et qui peuvent permettre d’améliorer les résultats du sondage. Lorsque l’on dispose de telles données, il faut chercher à les utiliser pour avoir des estimateurs plus précis. On peut les utiliser aussi bien pour améliorer le sondage lors du tirage de l’échantillon (classement dans un tirage systématique, stratification, variable utilisée pour le calcul des probabilités d’inclusion dans un sondage à probabilités inégales, etc.) ou au niveau de l’estimateur seulement. Dans ce dernier cas on parle de redressement. On distingue trois grandes méthodes de redressement couramment utilisés : l’estimateur post stratifié, l’estimateur parle ratio et l’estimateur par la régression.

On tire par exemple un échantillon de taille n par le SAS. On se rend compte que l’estimateur sans biais donne une valeur manifestement fausse au regard de la valeur réelle connue d’une variable X fortement corrélée à la variable d’intérêt Y. On décide alors de tenir compte des informations à dispositions pour trouver un meilleur estimateur en se basant sur les informations qu’apporte la variable auxiliaire X. On corrige donc les poids dans l’échantillon de manière à ce que cette correction donne une valeur correcte de l’indicateur. On se dit alors que si X a été corrigée par le nouvel estimateur, Y très corrélé à X, doit l’être aussi par la méthode de correction de X.

### 8.2.- La post stratification simple

C’est la méthode de redressement la plus courante.

**8.2.1.- Estimation d’une moyenne et d’un total**

Soit H groupes constitués à partir de l’échantillon sur la base d’une information auxiliaire. Notons que les Yα ne sont pas connues sur la population mais le sont sur l’échantillon (puisque ce dernier est tiré). Les H groupes sont appelés des post-strates.

On suppose que les tailles Nh sont connues quel que soit h  et qu’au moins un individu de chaque catégorie a été tiré compte tenu de la valeur de n.

Lorsqu’on retient les mêmes notations que dans le cas de la méthode du sondage stratifié, on a :

est un estimateur sans biais de et

est un estimateur sans biais du total global T.

Ces estimateurs sont appelés estimateurs post-stratifiés respectivement de la moyenne et du total T.

Ces estimateurs, dans leur formulation ne sont pas différentes de ceux de la stratification mais il existe entre eux une différence de taille. Dans le premier cas les nh sont choisis par le sondeur et dans le second cas, ils sont une variable aléatoire et donc lui sont imposées. Ainsi, les variances de ces indicateurs ne sont pas les mêmes que dans le cas simplement stratifié.

**8.2.2.- Précision de la post stratification**

Du fait que les nh soient aléatoires dans une post stratification, la variance de la moyenne dans une post stratification n’est pas la même que dans le cas de la stratification simple.

On obtient une approximation de cette variance par la formule suivante lorsque les nh sont assez grands :

Et donc on a :

NB : Cette précision est meilleure que dans le cas d’un SAS mais est légèrement moins bonne que dans le cas d’une stratification normale

### 8.3.- La post stratification sur critère multiple : la méthode du raking ratio

Lorsque l’on dispose de deux critères qualitatifs ou plus, deux cas sont possibles. Si l’on connait la structure de la population mère selon les différents critères, on applique la méthode de la post stratification simple sur chacun des croisements des variables de post stratification. Si on ne connaît que les totaux marginaux, on adopte la méthode du raking ratio encore connue sous le nom de calage sur marges.

Le principe est le suivant. On réalise un sondage aléatoire simple. Soient deux variables qualitatives X1 et X2 dont on connait les pourcentages sur la population mère. Il se trouve que dans l’échantillon, on ne retrouve pas les structures selon X1 et X2, deux variables importantes. On peut alors penser que la structure selon Y (corrélé à X1 et X2) est également mauvaise dans l’échantillon. On décide alors de faire coïncider les structures selon X1 et X2 dans l’échantillon à ce qu’on a dans la population générale.

Soit les tableaux suivants

Tableau1 : croisement de X1 et X2 dans l’échantillon

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Colonnes j | | | Total |
| Lignes i |  |  |  |  |
|  |  |  |  |
|  | nij |  | ni. |
|  |  |  |  |
| Total |  | n.j |  | n |

Les nij sont tous connus.

Tableau 2 : croisement de X1 et X2 dans la population mère

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Colonnes j | | | Total |
| Lignes i |  |  |  |  |
|  |  |  |  |
|  | Nij |  | Ni. |
|  |  |  |  |
| Total |  | N.j |  | n |

Les Nij ne sont pas connus mais les Ni. et les N.j le sont.

On constate que

On modifie les nij pour qu’on ait les deux égalités. Ceci revient à modifier les poids des individus de l’échantillon de manière à faire coïncider sa structure à celle de la population mère selon les critères retenus.

L’algorithme est le suivant

Etape 1 : calage sur les lignes

On détermine les n’i. tels qu’on ait

Donc On modifie en conséquence le tableau 1 de l’échantillon en remplaçant les nij par

Ainsi,

Etape 2 : calage sur les colonnes

On détermine les n’.j tels qu’on ait

Donc On modifie en conséquence le nouveau tableau des de l’échantillon en les remplaçant par

Ainsi,

Après cette opération, on dérange les totaux lignes qui ne sont plus égaux aux .

Etape 3 : calage à nouveau sur les lignes

On recommence le processus jusqu’à ce qu’on obtienne une convergence.

On aboutit à des valeurs n’ij tels que

et

On a

### 8.4. Estimation par le quotient(ou par le ratio)

Elle repose sur l’existence d’une relation entre Y et une variable quantitative X connue.

On a une relation du type où r est une constante et ui un résidu très petit.

Si l’on connait les moyennes ou les totaux de X à la fois sur l’échantillon ( et sur la population totale (, l’estimateur par le quotient est :

En effet, on a :

Donc

Et

Ainsi, dans l’estimation par le quotient, le poids dans un SAS ( est remplacé par

est donc biaisé mais ce biais devient très petit lorsque

La précision de cet estimateur, assimilée à son erreur quadratique moyenne est inférieure à la précision de dans un SAS lorsque où est le coefficient de corrélation linéaire entre X et Y dans la population générale, les écarts-types respectifs de X et Y dans la population générale.

On a :

Ainsi, lorsque CV(X) et CV(Y) sont sensiblement voisin, (c’est le cas lorsque ), on a intérêt à utiliser l’estimateur par le quotient lorsque

### 8.5. Estimation par la régression

Elle suppose une relation linéaire du type Y=a+bX entre la variable d’intérêt et la variable auxiliaire X.

On connaît la moyenne dans la population mère.

La méthode consiste à estimer par MCO le modèle à partir de l’échantillon.

On a :

Ainsi, on a :

Cet estimateur est légèrement biaisé. Le biais est d’ordre 1/n

Sa variance approximative est telle que ) avec =coefficient de corrélation linéaire entre X et Y.

### 8.6. Le traitement des non réponses

Au vu des différents taux de refus dans certains types d’enquête (par exemple les enquêtes sensibles), les non réponses constituent un problème majeur pour les enquêtes puisqu’elles biaisent les estimateurs.

On distingue deux types de non réponse : les non réponses totales et les non réponses partielles.

### 8.6.1.- Les non réponses totales

On a une non réponse totale lorsque l’on n’a aucune donnée sur l’unité d’observation. On n’a ainsi aucune réponse aux questions posées.

Une telle situation peut se présenter par exemple lorsque :

* L’individu est absent et n’a pu être interrogé malgré de nombreux passages des enquêteurs ;
* Le questionnaire de l’individu a été perdu ;
* L’unité d’observation (ou d’échantillonnage) est souffrant, décédé, indisponible, etc.
* Les non réponses totales se traitent par la méthode de la repondération.

### 8.6.2.- Les non réponses partielles

On dit qu’il y a non réponse partielle ou valeur manquante lorsque la personne enquêtée répond seulement à certaines questions mais pas à d’autres. On considère aussi qu’on a une valeur manquante lorsqu’à l’évidence la réponse obtenue est à l’évidence erronée (on n’a pas fait assez attention sur le terrain pour avoir une valeur vraisemblable).

### 8.7.- Correction des non réponses

Le plus souvent les gens corrigent les non réponses en les ignorant simplement. Ceci n’est valide que lorsque la non réponse est purement l’objet du hasard. Ceci est rarement le cas. En effet, les répondants ont en général un comportement différent de celui des non répondants. Dans ce cas, les ignorer constitue une faute statistique à ne pas commettre.

Différents moyens existent pour corriger les non réponses.

### 8.7.1.- Correction des non réponses totales

En général, les gens se contentent de remplacer les individus en cas de non réponses totale (sur le terrain on remplace un individu qu’on n’arrive pas à avoir ou qui a été sélectionner à tort). Ceci n’est pas la bonne attitude à avoir. On devrait corriger une non réponse totale par la repondération.

La repondération consiste à probabiliser l’univers des comportements des individus. Soit Ri la probabilité pour un individu de l’échantillon d’accepter de répondre et wi son poids (inverse de sa probabilité d’inclusion). Il s’agit d’estimer les Ri puis de calculer les nouvelles pondérations  ce qui revient à faire une double inférence d’abord de l’échantillon des répondants à l’échantillon totale (grâce aux Ri estimés) puis de l’échantillon à la population totale grâce aux probabilités d’inclusion (ou encore aux poids wi )

NB : pour un SAS, dans l’estimation d’un total.

Plusieurs méthodes existent pour estimer les Ri

1. Estimation des Ri par le taux de sondage empirique

Lorsqu’on pense que chaque individu a la même chance de répondre, Ri=R= où n0= effectif des répondants

On a

Dans le cas par exemple d’un SAS, on a

d’où

, moyenne de Y parmi les répondants. Ce qui équivaut à ignorer les non répondants.

1. Regroupement des répondants en catégories

En pratique, pour estimer les Ri, on regroupe l’échantillon en catégories homogènes dans lesquelles on suppose que les individus ont des comportements globaux similaires. Le taux de réponse dans chaque catégorie estime la probabilité de répondre pour chaque individu du groupe. Soit H catégories et nhr le nombre de répondants dans le groupe h et la moyenne de Y pour les répondants de la catégorie h. Le taux de réponse dans le groupe h est donc

Si on réalise un SAS, on a :

1. Utilisation d’un modèle

Il est également possible d’estimer Ri individu par individu en modélisant le comportement individuel.

Xi étant un vecteur de variables auxiliaires, connus pour chaque individu, on postule que Ri = h(Xi).

Par exemple,



Le modélisateur aura à estimer les paramètres de cette loi et en déduire les estimations individuelles des Ri

NB : on peut également utiliser la post stratification pour corriger le problème de non réponse totale

### 8.7.2.- Correction des non réponses partielles

Elle consiste à estimer les valeurs manquantes. La valeur manquante yi est ainsi remplacée par une autre valeur y’i la plus pertinente possible au regard des valeurs connues sur l’individu i, ses proches, d’une manière plus générale, sur l’ensemble de la population et l’environnement. Cette opération s’appelle l’imputation.

Il faut mieux ne pas y avoir à recours car le remède peut s’avérer plus nocif que le mal. On recommande d’ailleurs de ne pas la tenter lorsque le taux de non réponse est élevé, (en général supérieur à 40 %).

L’idée de l’imputation est l’existence d’une relation stable entre la variable d’intérêt Y et un vecteur X de variables X connues sur les répondants et les non répondants.

### 8.7.2.1- La méthode économétrique

Elle consiste à estimer directement la valeur inconnue de la non réponse pour chaque individu de l’échantillon des non répondants en utilisant un modèle liant la variable non répondue à un vecteur de variables auxiliaires Xi connu. Le modèle est basé sur une hypothèse de comportement. On l’estime en utilisant les données relatives aux répondants. Le modèle construit est ensuite utilisé pour l’imputation des valeurs inconnues se rapportant aux non répondants en utilisant les vecteurs connus Xi. sur les répondants.

On pourrait par exemple utiliser un modèle linéaire pour estimer les loyers (manquants) de propriétaires lors d’une enquête sur les logements. Ce modèle linéaire est ensuite appliqué aux non répondants (propriétaires) pour imputer une valeur pour leur loyer fictif.

### 8.7.2.2- La méthode déductive

Elle consiste à arrêter une règle d’imputation et à l’appliquer systématiquement.

Exemple : Lorsque l’âge est inférieur à 10 ans, l’état matrimonial est « célibataire ».

### 8.7.2.3- La méthode du cold deck

Le principe du « cold-deck » consiste à imputer une valeur à la non réponse sur la base d’une information extérieure à l’échantillon mais crédible.

Exemple : utiliser la valeur de la variable manquante à une époque antérieure si cette valeur est connue. Ou la modifier légèrement pour anticiper son évolution.

### 8.7.2.4- La méthode du hot deck

La technique du « hot-deck » consiste à choisir de façon aléatoire une unité j dans l’échantillon qui soit proche de l’individu i et à affecter sa valeur à la donnée manquante de i. Il faudrait bien entendu avoir une population très homogène car l’hypothèse implicite est l’indépendance entre la non réponse et la variable entachée de non réponse.

L’imputation par le plus proche voisin consiste à définir un voisinage (basé sur une distance donnée). On impute la valeur du voisin le plus immédiat à la valeur manquante.

On peut recourir à une classification des unités grâce à une analyse factorielle.

### 8.7.2.5- La méthode d’imputation par prédicteur

On impute la moyenne ou la médiane de tous les répondants ou des individus proches de i à la valeur manquante.