MINISTERE DES ENSEIGNEMENTS SECONDAIRES, SUPERIEURS ET DE LA RECHERCHE SCIENTIFIQUE
==========
UNIVERSITE JOSEPH KI-ZERBO
ISSP
==========

LICENCE PROFESSIONNELLE EN ANALYSE STATISTIQUE

COURS DE SONDAGE

Dr A. NIKIEMA

Février 2025

SOMMAIRE

INTRODUCTION

Dans le processus d'acquisition de l'information statistique ; il arrive, en raison de diverses contraintes (budgétaires, temporelles, spatiales, institutionnelles, ressources humaines, etc) que l'on ne soit pas en mesure d'interroger toute la population étudiée. Notamment au cas où l'enquête requiert la destruction ou la consommation d'un produit (exemple des contrôles de fabrication) ou encore la population concernée est pratiquement inaccessible (évolution du prix de détail, de la pollution de l'eau, …).
L'idée d'interroger une partie de cette population, ***l'échantillon***, afin de produire l'information sur l'ensemble a donné lieu aux techniques de sondages.
La plus connue des techniques de sondage est le sondage d'opinion qui n'est qu'une infime partie des méthodes de sondage. Ainsi, au-delà des sondages d'opinion, il est appliqué dans plusieurs domaines :

- **Démographie** : pour l'évaluation de la population, des types de biens privilégiés par cette population ; pour l'identification de leurs comportements face aux problèmes économiques et sociaux. En outre, l'étude de problèmes spécifiques tels que l'importance de la population active au chômage, le travail à temps partiel, le niveau de salaire, les effectifs des diverses branches d'activité est envisagée par les enquêtes par sondage.
- **Dans le domaine industriel et commercial** : les sujets tels : l'effectif d'employés dans les entreprises, les quantités de matières premières consommées, les capacités de production des branches, l'importance des investissements réalisés ou envisagés, l'état des stocks, les perspectives de marchés, etc sont abordées à partir des enquêtes par sondage.
- **L'agriculture** : les surfaces cultivées, le type de culture et superficies consacrées, importance et état du cheptel, modes d'exploitation, rendements, … sont dans le champ d'application des méthodes de sondage.
- D'autres domaines scientifiques telles les sciences de l'***éducation***, la ***santé***, la ***biologie***, la ***physique***, la ***météorologie***, l'***écologie***, le ***contrôle industriel de qualité*** et ***les études de l'opinion politique*** se retrouvent également dans le champ d'application des méthodes de sondage.

La mise en œuvre de ces techniques pose la question de savoir si ***les informations recueillies sur l'échantillon décrivent exactement ou presque les caractéristiques de la population comme l'aurait donné le recensement***. Plus généralement, l'utilisation des techniques de sondage pose les problèmes suivants :

- Quelle méthode de tirage retenir pour la sélection des individus à enquêter (choix de l'échantillon : échantillonnage) ?
- Une fois l'échantillon constitué, comment agréger les informations recueillies pour produire des résultats sur l'ensemble de la population concernée (estimation) ?
- Jusqu'à quel point peut-on faire confiance aux résultats de l'inférence statistique (calcul de précision) ?
- Qu'est-ce qui influence cette précision en dehors même de la procédure d'agrégation utilisée ?

L'objet de ce cours de sondage est de donner des éléments de réponse à ces préoccupations. Il s'agira donc tout au long de ce cours de présenter les différentes méthodes de sondage depuis les procédés d'échantillonnage jusqu'aux méthodes d'estimation statistiques et d'essayer de faire un lien avec la pratique en s'appuyant sur un certain nombre d'exemples d'enquêtes par sondage réalisées dans des pays en développement. Ainsi, nous nous proposons d'organiser ce cours au tour des aspects suivants :

- Les étapes d'une enquête

- Le rappel de quelques notions essentielles
- les méthodes empiriques
- le sondage aléatoire simple
- Les sondages stratifiés a priori
- Les sondages à plusieurs degrés, par grappes

Des travaux pratiques de simulations de tirages d'échantillons et les méthodes de redressement seront également appliquées.

CHAPITRE I : LES ETAPES D'UNE ENQUETE.

Une enquête statistique, fruit d'un travail d'équipe composée de plusieurs partenaires parmi lesquels le statisticien, est une chaîne de travaux complexes qui part de la préparation administrative et technique à la publication des résultats, en passant par la collecte de données sur le terrain et le traitement et l'analyse des données. Nous présentons ci-dessous quelques unes des phases importantes du déroulement d'une enquête :

I. La définition des objectifs et les contraintes de l'enquête.

On précise ce que l'on veut. Quelles sont les variables d'intérêt ?
Les objectifs de l'enquête sont généralement fixés par les demandeurs. La première chose que le responsable du sondage doit obtenir d'eux est les objectifs précis de l'enquête. Il s'agit d'expliciter ses buts exacts. Dans cet exercice, le statisticien, dans son dialogue avec les demandeurs, doit les amener à fournir les renseignements précis sur les données à collecter :

- La population à étudier : individus, ménages, exploitations agricoles, entreprises industrielles et commerciales, localités, etc.
- Les variables à observer : superficie physique totale, cultivée, le chiffre d'affaires, la valeur ajoutée, etc.
- La définition claire du champ de l'enquête : ensemble du pays, une région particulière ? etc.
- Les principaux concepts et définitions. Par exemple : quelle définition retient-on pour le ménage, la parcelle, etc ?
- La forme sous laquelle les données doivent être présentées.
- Le niveau de signification souhaité (ensemble du pays, région, etc).
- Les ressources financières disponibles.
- Les contraintes

Par exemple la plupart des enquêtes agricoles, en Afrique particulièrement, comprennent quatre éléments principaux : les exploitations, l'exploitant, le ménage de l'exploitant, la production.
Les principaux objectifs que l'on rencontre généralement dans ces enquêtes sont les suivants :

- étude des structures de production agricole : nombre d'exploitations et leur répartition géographique et selon certaines caractéristiques simples des exploitants (sexe et âge, niveau d'instruction, etc)
- étude des exploitations selon certaines caractéristiques (répartition suivant le matériel et outillage, les superficies totale et cultivée, la main-d'œuvre utilisée, etc);
- la démographie de l'exploitation (ménage de l'exploitant)
- estimation de la superficie par culture
- estimation du rendement par culture
- estimation de la production par culture
- estimation du cheptel par espèce

Mais l'on trouve aussi d'autres types d'objectifs comme : les prix agricoles, les revenus des agriculteurs, stocks céréaliers, etc.

II. La détermination de la période d'enquête et de la période de référence.

Sur le plan opérationnel, il est souhaitable de décider de ces deux périodes longtemps à l'avance.

La période d'enquête.

La période d'enquête est celle pendant laquelle la collecte des données sera effectuée sur le terrain. Elle doit garantir un bon contrôle des probabilités de tirage des unités (neutralisation des effets saisonniers en particulier). Pour les enquêtes agricoles, celle-ci correspond généralement à la campagne agricole. Il est toutefois recommandé d'organiser les passages en fonction du déroulement de la campagne.

La période de référence.

La période de référence est celle à laquelle les données se rapportent. Elle dépend des objectifs de l'enquête. Selon les cas, c'est un intervalle (semaine, mois, campagne agricole, etc) ou une date précise. Il faut noter que, pour une même enquête, plusieurs variables peuvent avoir des périodes de référence différentes. Par exemple, dans une enquête sur la production végétale incluant un module sur les prix à la consommation des produits agricoles, la période de référence des prix peut être le mois alors que celle de la production végétale est la campagne agricole.

III. La méthode de collecte de données.

La planification et l'exécution d'une enquête sont largement influencées par la méthode de collecte de données. Après un examen minutieux des objectifs de l'enquête, de la base de sondage, du plan de sondage et du budget, une décision doit être prise pour ce qui concerne la méthode de collecte de données. Il existe plusieurs méthodes de collecte : voie postale, téléphone, enquêteur, une combinaison de deux méthodes. Toutefois, pour les pays en développement en général et d'Afrique au Sud du Sahara en particulier, les deux premières ne sont pas souvent envisageables. On a généralement recours à la troisième ou à une combinaison de méthodes. On peut par exemple décider de recenser les exploitations modernes par correspondance avec un questionnaire spécifique et de couvrir les exploitations traditionnelles par une enquête par sondage.
Un soin particulier doit être apporté à la rédaction des questionnaires. De même, des instructions précises doivent être préparées pour l'ensemble du personnel de terrain. Celui-ci devra recevoir une formation adéquate. Le questionnaire doit être testé sur un échantillon pour s'assurer qu'il réponde aux objectifs fixés et relever certaines insuffisances (formulation de certaines questions, difficultés de répondre, temps d'administration du questionnaire etc.);

IV. Le choix de la base et des unités de sondage.

L'exigence principale d'une enquête par sondage est de disposer d'une base de sondage, c'est-à-dire une liste de l'ensemble des unités pour lesquelles les données doivent être collectées. La base de sondage est un élément clé d'une enquête car c'est elle qui conditionne les procédures de tirage et d'estimation. Son choix dépend donc des objectifs de l'enquête.

V. Le choix du plan de sondage.

C'est cette partie qui fait l'objet central de ce cours et qui sera développée dans la suite.
Après avoir pris en considération les divers paramètres techniques et organisationnels, le choix d'un plan de sondage optimal est une phase capitale. C'est à ce niveau que l'on détermine :

- la taille de l'échantillon ;
- les procédures de tirage ;
- les estimateurs calculés, ainsi que leur précision théorique.

Pour une situation donnée, il existe divers plans de sondage possible. Cependant, il faut toujours prendre en compte la précision des données souhaitée et les ressources disponibles. Dans ce cadre, il est essentiel d'avoir toujours en mémoire le principe d'optimisation, c'est-à-dire de chercher à obtenir :
(i) un degré de précision donné au moindre coût, ou bien
(ii) un maximum de précision pour un coût donné fixé à l'avance.

VI. Le suivi du déroulement du terrain.

Il s'agit de contrôler le travail du personnel de terrain, de l'assister et de l'informer des cas litigieux, de prévoir des stratégies de relance, etc. Un aspect important est le respect des probabilités de tirage lors de l'exécution de l'enquête. **Cela suppose que les unités sélectionnées, non seulement sont celles qui sont effectivement observées, mais le sont toutes une et une seule fois**.

VII. Le traitement des données.

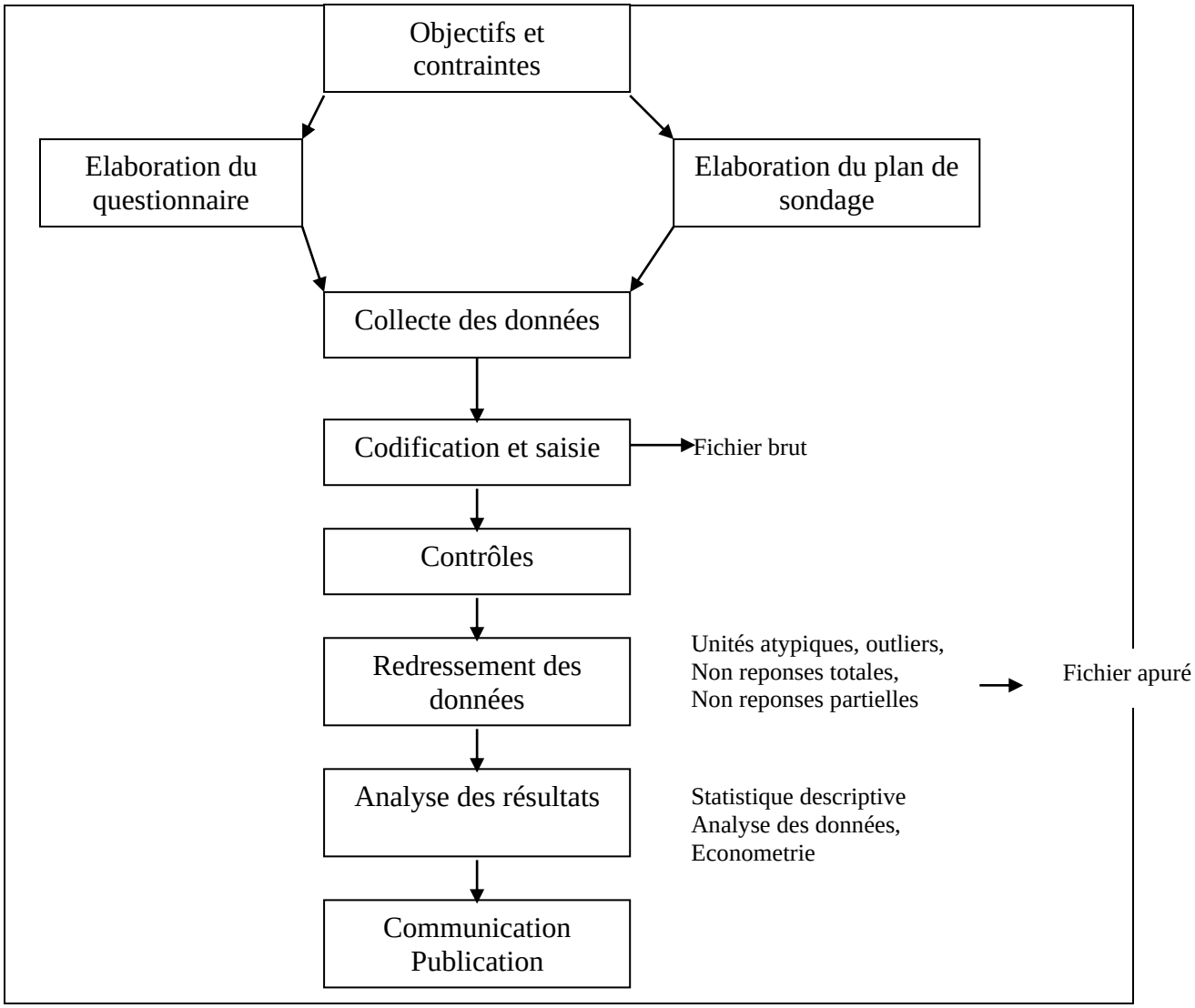
Il s'agit des programmes de saisie, de contrôle de validité, de redressement, de recodification, de tabulation, d'analyse statistique, de diffusion et d'archivage. Au sortir de cette étape, nous disposons d'un fichier brut puit d'un fichier apuré (fichier propre). C'est à ce niveau que l'on redresse l'échantillon, que l'on traite les outliers, les points extrêmes, les non-réponses, … Après cette étape, la base est prête à l'emploi et peut être l'objet de toutes les analyses : statistique descriptive, analyse des sonnées, économétrie,

VIII. La préparation du rapport et la diffusion des résultats.

La dernière étape de l'enquête est la préparation du rapport et sa publication. Un aspect important du rapport est la présentation de la précision des données, qui permettra aux utilisateurs de connaître qu'elle confiance ils peuvent accorder aux données présentées.

Le statisticien intervient pratiquement à toutes les phases de réalisation d'une enquête décrites ci-dessus.
(i) La définition de la population à étudier au travers des unités qui la composent est la responsabilité des demandeurs de l'enquête, mais le sondeur intervient pour restreindre ou accroître l'effectif de la population à enquêter en fonction des besoins exprimés et des ressources disponibles jusqu'à ce qu'un plan de sondage adéquat soit établi. En effet, il y a généralement un dilemme entre le niveau de précision souhaité et les ressources financières disponibles.
(ii) La procédure d'estimation (ou l'estimateur) est liée à la procédure d'échantillonnage ; par exemple, seule la prise en compte simultanée des probabilités de tirage et des pondérations retenues permet de définir des estimateurs sans biais.
(iii) Les erreurs d'échantillonnage sont étroitement liées à la méthode de tirage choisie et font appel à la théorie des sondages.
(iv) Les compétences et le savoir-faire du responsable du sondage sont requises pour résoudre les problèmes d'erreurs d'observation : non-réponses, défauts de couverture du champ de l'enquête, etc, bien que ceux-ci soient liés essentiellement aux opérations de collecte de données sur le terrain qui sortent du domaine de responsabilité du sondeur.

Les différentes étapes sont consignées dans le schéma ci-après :



NB : Il convient de définir clairement le temps imparti à chacune de ces étapes et les optimiser.

CHAPITRE II : RAPPEL DE QUELQUES NOTIONS ESSENTIELLES.

I. Généralités

Le sondage est une technique statistique qui permet de produire de l'information sur un domaine donné à partir de l'observation d'une partie de celui-ci. A ce titre, il se situe en amont de nombreuses études statistiques dans la mesure où il fournit un cadre méthodologique propre à guider le choix de méthode de collecte de l'information. Notamment, il définit les conditions pour lesquelles certaines stratégies sont préférables à d'autres, et permet d'éviter des erreurs. En outre, il fournit au statisticien les rudiments à même de lui permettre de minimiser les marges d'erreur lors de la collecte de données et d'en évaluer l'importance.
Comme technique statistique, les sondages disposent d'un langage propre qu'il convient de s'approprier avant l'analyse des différentes techniques.

II. CONCEPTS ET DEFINITIONS DE BASE.

1. Unités statistiques et population ou univers.

Le contexte général où l'on se situe en sondage est celui d'une collection finie d'unités U = (U1 ,U2 ..., UN), chaque unité Ui étant porteuse d'une valeur numérique Yi. (Par codification, les caractères qualitatifs peuvent être convertis en valeurs numériques, par exemple 1 pour le sexe masculin et 2 pour le sexe féminin). Nous voulons connaître le total Y ou la moyenne des valeurs Yi .
On appelle unités statistiques les éléments (Ui) porteurs des variables à l'étude, et sur lesquels sont mesurées ces variables. Les unités statistiques sont également appelées unités d'observation.
On appelle population la collection (U) des unités statistiques étudiées.
Pour différentes raisons, nous voulons connaître les statistiques Y en ne considérant qu'une partie u de U plutôt qu'en procédant à un recensement (dénombrement complet). On dit que l'on fait un sondage de la population U.
L'univers représente le domaine de l'étude. On parle aussi de population au sens statistique du terme. Il peut représenter une population de personnes aussi bien qu'une population de ménages, de villages, de pays ou même d'événements (naissance, décès, …).
Pour une étude donnée, cet univers doit être défini de manière précise c'est à dire par les éléments qui le composent et par ses limites car ceux-ci conditionnent la pertinence des résultats que l'on tirera du sondage.

2. Sondage et échantillon.

Sonder une population U, c'est en sélectionner une partie seulement u en vue d'en inférer certaines statistiques sur l'ensemble de la population. La partie u est appelée échantillon de U.
Il faut bien noter que tout sondage implique deux opérations intimement liées :

a) la sélection d'une partie u de U ;
b) l'inférence (extrapolation) à partir de u sur U.

Dans l'usage courant de la technique de sondage, le secteur de l'inférence statistique directement impliqué est celui de l'estimation de paramètres. Le test d'hypothèse, au sens mathématique et restrictif du terme, est rarement considéré comme l'objectif principal d'un sondage (ce qui ne l'exclut pas). Bien plus, l'estimation d'un total, objectif majeur d'un sondage, est négligée dans la théorie générale de l'inférence statistique : en effet, cette théorie traite le plus souvent du cas idéal des distributions continues, où la notion de taille de population n'a pas de sens. Enfin, pour mieux cerner le domaine de l'inférence statistique qui

intéresse la technique de sondage, notons que l'estimation d'une moyenne par unité statistique est souvent réalisée au moyen de l'estimation d'un rapport de totaux, ce qui exige de recourir à des quotients de variables aléatoires.

3. Unités d'échantillonnage.

On appelle unité d'échantillonnage ou de sondage les unités directement soumises à une opération de sélection. L'unité d'échantillonnage est souvent constituée d'une grappe d'unités statistiques.
Exemple : supposons que dans une préfecture donnée, on désire mesurer la production vivrière. Pour ce faire, l'unité statistique normale est l'exploitation. Si l'on sélectionne quelques villages et l'on recense les exploitations de chacun, alors le village est l'unité d'échantillonnage, et l'exploitation est l'unité statistique.

4. Estimateurs : définition et choix

4.1. Estimateur : fonction d'intérêt

Après avoir déterminé l'échantillon et observé un certain nombre de caractéristiques sur celui-ci, il convient de procéder à l'estimation pour une ou plusieurs variables dites d'intérêt.
Ainsi un estimateur permet de calculer l'estimation d'une grandeur à partir des données observées sur l'échantillon tiré. Le choix de ces estimateurs dépend de la méthode de sondage utilisée.

4.2. Le biais

Un estimateur A d'une grandeur G est dit sans biais si E(A) = G c.-à-d. si ''en moyenne'' les résultats fournis par cet estimateur sont égaux aux paramètres qu'on cherche à estimer. Dans le cas contraire, on a un estimateur biaisé.
Le biais est donc défini par la différence entre l'espérance en question et le paramètre :

$$B = E(A) - G$$

4.3. Critères de choix d'un estimateur

Il existe deux critères de choix des estimateurs : l'optimalité, et l'admissibilité.
En raison de l'absence de solution par le maximum de vraisemblance, d'autres critères ont été utilisés pour choisir des estimateurs.
Un estimateur $\hat{\theta}$ d'un paramètre θ est meilleur qu'un autre $\tilde{\theta}$ si et seulement si:

Pour tout échantillon x_i, i = 1, …, n $\qquad EQM(\hat{\theta}) \leq EQM(\tilde{\theta})$

L'idéal qui serait de trouver l'estimateur optimal, c'est à dire qui est le meilleur que tous les autres, est incertain (EQM= **erreur quadratique moyenne)**. Toutefois, la recherche d'un tel estimateur s'avère complexe. On se limitera à des classes restreintes d'estimateurs telles :

- La classe des estimateurs linéaires, homogènes ;
- La classe des estimateurs sans biais ;
- La classe des estimateurs linéaires sans biais.

Malheureusement, même en se restreignant à des classes particulières d'estimateurs, l'existence d'un estimateur optimal n'est pas assurée.

5. Sondages et recensements : vocabulaire

L'on peut mettre en œuvre deux stratégies pour collecter l'information auprès d'une population composée d'unités statistiques relativement nombreuses :

- Soit interroger toutes les unités de la population concernée : C'est le principe du recensement.

- Soit, à partir d'un plan judicieusement préétabli, interroger un échantillon d'unités quitte à extrapoler ces résultats à l'ensemble de la population. Tel est le but du sondage.

Ainsi, l'on privilégie dans certains cas le sondage au recensement parce que ce dernier est une opération lourde qui exige des délais et des budgets importants. En outre ; le sondage, bien préparé et bien exécuté, semble de meilleure qualité que le recensement en raison de la réduction des ***erreurs d'observation*** relativement difficile à contrôler dans le cas du recensement. Toutefois, plutôt que de les opposer, il convient d'établir une relation de complémentarité entre ces deux stratégies.

5.1. Base de sondage

Une ***base de sondage*** est une liste exhaustive et à jour des unités de l'univers sans omission ni double décompte, et telle que l'identification de chaque unité se fasse sans ambiguïté.
Il est intéressant de disposer, dans la base de sondage, d'informations concernant les unités statistiques utilisables pour le sondage. Ces informations renseignées pour toutes les unités de la base de sondage sont appelées ***variables auxiliaires***. Ces informations peuvent être utilisées, soit pour appliquer un plan de sondage plus précis, soit pour déterminer des estimateurs plus précis.
Une base de sondage peut être obtenue à partir :

- Des documents administratifs existants ;
- Du fichier des clients d'une société ou des anciens élèves d'une école ;
- De la liste venant d'une enquête précédente, en particulier d'un recensement ;
- Une liste qui est dressée à l'occasion de l'enquête,
- etc.

5.2. Plan de sondage

Il existe deux types de plan de sondage :

- Les sondages déterministes et
- Les sondages probabilistes ou aléatoires.

Un sondage est dit aléatoire ou probabiliste si chaque unité statistique de la population cible a une probabilité non nulle et connue d'appartenir à l'échantillon. Les techniques de sondage aléatoires sont fondées sur le principe selon lequel : ''l'échantillon doit être déterminé d'une façon objective'' de sorte que tout élément de l'ensemble à étudier ait au moins une chance d'être choisi et que cette chance puisse être déterminée avec certitude.
Le sondage est dit déterministe si cette condition n'est pas satisfaite.
Remarque : l'organisation d'un sondage aléatoire impose la disponibilité d'une base de sondage.

III. La qualité d'une enquête par sondage

L'on peut se placer sur deux angles pour apprécier la qualité d'une enquête par sondage :

- Sur le plan purement statistique ou
- Sur un plan plus général.

1. Définition purement statistique de la qualité d'une enquête

Sur le plan statistique, la qualité d'une enquête est appréciée à deux niveaux :

- La qualité du plan de sondage et
- La qualité des données de l'enquête

1.1. La qualité du plan de sondage

La qualité du plan de sondage est approchée par deux critères : d'une part, la base de sondage, et d'autres part, la notion de la ***''représentativité''*** de l'échantillon.
Au niveau du plan de sondage, on s'interroge sur la pertinence de la technique de sondage utilisée : un sondage aléatoire qui suppose l'existence d'une base de sondage et le choix de la technique utilisée (SAS, …) ou un sondage déterministe. Au niveau de la représentativité de l'échantillon, on se demande si votre échantillon fournit une image réduite mais fidèle de l'ensemble sur lequel il est prélevé, au moins en ce qui concerne les caractéristiques que l'on cherche à apprécier. Nous verrons dans les chapitres suivants qu'il n'est pas nécessaire que l'échantillon soit une marquette ''exacte'' de l'univers : certaines parties de celui-ci peuvent-être surreprésentées dans l'échantillon.

1.2. La qualité des données

La qualité des données de l'enquête est approchée par deux critères : la ***précision des estimateurs*** et le ***biais***. L'objectif étant d'obtenir des estimateurs qui aient une erreur quadratique moyenne la plus petite possible .
Or, nous avons, $EQM = V(X) + B^2$. Il est nécessaire de minimiser les deux quantités qui composent l'erreur quadratique moyenne. Ainsi, la précision est évaluée soit par la variance, soit par le coefficient de variation ou encore par l'intervalle de confiance. On souhaiterait avoir des estimateurs de variance ou de coefficient de variation le plus petit possible. Concernant le biais, il est imputable aux différentes erreurs que l'on peut rencontrer. Il en existe quatre types :
- Les erreurs d'échantillonnage,
- Les erreurs de couverture,
- Les erreurs de mesure ou erreurs d'observation,
- Les erreurs de non-réponse.

2. Définition générale

Une autre approche consiste à dépasser les critères purement statistiques en leur adjoignant cinq critères :
- ***La rapidité*** : qui est l'appréciation du temps qui s'est écoulé entre la période de référence (date de signature du contrat) et la disponibilité des premiers résultats.
- ***La cohérence*** : il existe plusieurs sources d'informations. Vos résultats sont-ils cohérents avec ceux d'une autre source ? Sinon pourquoi ?
- ***L'accessibilité*** : comment l'accès aux résultats, aux données sera-t-il assuré ? Quel est le support : manuscrit, …cd-rom ? Quel est le prix ?
- ***L'interprétabilité*** : vos données sont-elles interprétables facilement par un grand public ? ce critère impose une définition des concepts conforme à d'autres disciplines, …
- ***La pertinence*** : vos résultats permettent-ils de répondre aux préoccupations du commanditaire ou des utilisateurs potentiels ?

Avant de rentrer dans le détail des méthodes de sondage, commençons par **rappeler**, à travers des exercices, quelques notions essentielles pour la suite.

IV. NOTIONS DE STATISTIQUE.

Un paramètre d'une population est une **valeur numérique** qui caractérise un aspect bien particulier de la distribution d'une (ou de plusieurs) variable(s) numérique(s) de cette population. Définissons les paramètres les plus importants pour la suite du cours.

1. Total, moyenne, variance et écart-type, et coefficient de variation d'une variable Y.

Total d'une variable Y dans la population : $Y = \sum_{i=1}^{i=N} Y_i$

Moyenne d'une variable Y, caractéristique de tendance centrale, dans la population est : $\overline{Y} = \frac{\sum_{i=1}^{i=N} Y_i}{N} = \frac{Y}{N}$

Variance d'une variable Y dans la population et écart-type : $\sigma^2 = \frac{\sum_{i=1}^{i=N} \left(Y_i - \overline{Y}\right)^2}{N}$

Nous savons qu'il s'agit d'un paramètre de dispersion par rapport à la valeur centrale qu'est la moyenne. Nous savons que nous pouvons simplifier les calculs comme suit : $\sigma^2 = \frac{\sum_{i=1}^{i=N} Y_i^{\,2}}{N} - \overline{Y}^2$

L'écart-type $\sigma = \sqrt{V(Y)}$.

La variance et l'écart-type sont des caractéristiques de dispersion.

Le coefficient de variation $C.V = \frac{\sigma}{\overline{Y}}$.

Alors que la moyenne, la variance et l'écart-type sont exprimés avec la même unité que la variable étudiée, le coefficient de variation est « sans unité » (rapport de deux nombres exprimés avec la même unité). Elle définit l'importance de l'écart-type par rapport à la moyenne. Il exprime plus ou moins l'importance de la dispersion de la distribution.

2. Covariance de deux variables X et Y dans la population.

Lorsque chaque unité U_i de la population porte deux valeurs numériques X_i et Y_i on peut vouloir mesurer le degré la relation existante entre les deux variables X et Y.

La **covariance** mesure leur degré de relation *linéaire* (c'est-à-dire descriptible par une droite. La covariance se définit comme suit :

$$COV(X,Y) = \frac{\sum_{i=1}^{i=N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{N} = \frac{\sum_{i=1}^{i=N} X_i Y_I}{N} - \overline{X}\,\overline{Y}$$

Une covariance positive indiquera une relation croissante entre X et Y et une covariance négative indiquera une relation décroissante.

3. Coefficient de corrélation de deux variables X et Y.

A certains égards, la covariance est insatisfaisante pour mesurer le degré de relation entre deux variables. Sauf le signe, sa valeur comme telle n'a pas une signification précise,

puisqu'elle dépend de l'échelle utilisée dans l'expression des variables. De plus, la covariance n'a pas de borne inférieure ni supérieure, ce qui permet difficilement de juger à quel point on est éloigné d'une relation linéaire parfaite. Le coefficient de corrélation résout ces difficultés :

$$r(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)}$$

Il est évident que la valeur de r n'est pas modifiée par un changement d'échelle, et conserve le signe de la covariance. De plus :

a) r est toujours compris entre -1 et +1.
b) plus la relation linéaire est forte, plus r s'éloigne de 0 pour se rapprocher de -1 ou +1 ;
c) lorsque la relation linéaire est parfaite, c'est-à-dire lorsque tous les points sont sur une droite, r a la valeur +1 ou -1, et réciproquement.

CHAPITRE III : LES METHODES EMPIRIQUES

Les méthodes empiriques ou encore méthodes à choix raisonné sont utilisées quand on n'est pas en mesure de disposer d'une base de sondage. Leur mise en œuvre n'exige pas d'avoir la liste complète ou un fichier informatique répertoriant les individus à enquêter.
L'utilisation de ces méthodes dans les Pays en Développement reste encore limitée et est l'apanage des entreprises privées ou des structures qui n'ont pas toujours les moyens financiers ou le droit de posséder des bases de sondages incluant tous les individus de la population.
Elles sont le plus souvent utilisées dans les enquêtes d'opinion, dans les études de marché lors de la mise sur pied d'un nouveau produit, … Nous examinons trois méthodes classiques : la méthode des quotas, celle des itinéraires et la méthode des unités types.
Il convient, en outre, d'indiquer que ce sont des méthodes non probabilités car la probabilité d'être enquêtée de chaque unité statistique n'est pas connue à priori. Elle dépend en fait de l'enquêteur. De plus, certaines unités peuvent avoir une probabilité nulle d'être enquêtées.

I. Méthode des quotas

La méthode des quotas est la méthode empirique la plus utilisée.

1. Le principe

Cette méthode consiste à imposer à l'échantillon de respecter une répartition selon certains critères afin de ''représenter'' au mieux l'univers. Pratiquement, on divise la population en un certain nombre de sous-population selon une ou plusieurs variables catégorielles. Ensuite, on demande aux enquêteurs d'interroger un nombre d'individus proportionnel à chacune de ces sous-populations.
Ainsi, les enquêteurs ont le libre choix des personnes à enquêter, et donc, de constituer l'échantillon qui de ce fait est inconnu à priori.
En outre, il faut préciser que l'utilisation de quotas suppose que les variables retenues pour la détermination de ces quotas ont une influence prépondérante pour les caractéristiques que l'on veut estimer.

2. Exemples de quotas

On veut faire une enquête socio-économique sur la population active d'une ville. Un recensement récent a fourni les répartitions globales suivantes d'après trois critères : le sexe, l'âge et le secteur d'activité.

Sexe		Age		Secteur d'activité	
Hommes	48 %	15 – 24 ans	14 %	Cadres/patrons formel	16 %
Femmes	52 %	25 – 44 ans	37 %	Employés/ ouvriers formel	24 %
	100 %	45 – 64 ans	35 %	Cadres/patrons informel	3 %
		65 ans et +	14 %	Travailleurs indépendant	36 %
			100 %	Employés/ aides familiaux, etc.	21 %
					100 %

Source: données fictive

On veut enquêter 1000 personnes avec 10 agents enquêteurs travaillant pendant 10 jours. Et donc chaque agent enquêteur doit enquêter 100 personnes dont le profil doit reproduire les distributions marginales de chaque critère pris séparément.

Pratiquement, l'on remet à chaque agent une fiche de quotas en plus du questionnaire afin de lui permettre de gérer ses quotas. Cette fiche, en général se présente comme suit (Cf. schéma).

Nom de l'enquêteur:
Interviews à réaliser

Critères	**Catégories**		**Effectifs**
Sexe	Hommes		48
	Femmes		52
Age	15 – 24 ans		14
	25 – 44 ans		37
	45 – 64 ans		35
	65 ans et +		14
Secteur d'activité	Cadres/ patrons formel		16
	Employés/ ouvriers formel		24
	Cadres/ patrons informel		3
	Travailleurs indépendant		36
	Employés/ aides familiaux, etc.		21

Et donc au fil des interviews, l'agent enquêteur coche une case dans chaque critère selon les caractéristiques de l'individu interrogé.

3. Facteur influençant la qualité

La qualité des résultats dépend de deux paramètres :

- Des enquêteurs bien formés, n'introduisant pas de biais de sélection systématique, et habitués à ''gérer'' leurs quotas.
- Des informations sur les quotas fiables, c'est à dire récents et exacts.
- En outre, il est nécessaire de contrôler l'échantillon obtenu au regard de certains critères pour lesquels on connaît la distribution.

En somme, la méthode des quotas présente un intérêt certain en raison de sa facile mise en œuvre (pas de besoin de base, des enquêteurs moins contraints). Toutefois, son application repose d'abord sur la disponibilité de statistiques fiables et le savoir - faire du gestionnaire de l'enquête et des enquêteurs.

II. La méthode des itinéraires

Elle est parfois présentée comme une variante de la méthode des quotas.

1. Principes

On fournit à l'agent enquêteur des indications sur les unités à enquêter en lui imposant un itinéraire fixé sur une carte avec un point de départ et des points d'enquête déterminés le long de celui-ci. Le choix de l'itinéraire se fait de façon aléatoire.

2. Avantages/ inconvénients

Elle est facile à mettre en œuvre mais dans une mesure moindre que les quotas. Car le travail de détermination de tous les itinéraires possibles permettant d'accéder à l'ensemble des habitants de la zone est difficile à boucler de manière parfaite. Cette difficulté à des impacts significatifs sur l'estimation des totaux. En outre, sa mise en œuvre pose de nombreux problèmes en particulier dans les PVD.

III. La méthode des unités types

1. Principes

Elle consiste à considérer des groupes homogènes d'individus à enquêter et à interroger au niveau de chaque groupe un individu désigné comme l'***unité-type*** c.-à-d. celui dont les caractéristiques reproduisent au mieux les caractéristiques du groupe.

2. Avantages/ inconvénients

Cette méthode présente un intérêt certain notamment, lorsque l'on désire, sans grands moyens matériel, obtenir des renseignements pour des sous-populations relativement petits de la population. Toutefois, le choix de l'unité-type peut poser problème dans la mesure où son choix repose sur un jugement à priori qui s'il n'est pas fondé sur des connaissances objectives peut avoir des effets néfastes sur les grandeurs estimées.

IV. Remarques sur les méthodes empiriques

Les méthodes empiriques sont souvent critiquées. Cependant, elles ne sont pas nécessairement plus mauvaises que les méthodes probabilistes. Elles sont utiles dans les cas où les méthodes probabilistes sont impossibles à mettre en œuvre.
Par ailleurs, il faut préciser que les résultats obtenus par ces méthodes sont biaisés et il est très difficile de se faire une idée précise de l'ampleur du biais car la théorie probabiliste des sondages ne peut être appliquée aux méthodes empiriques. La seule méthode légitime d'évaluation de la validité d'une méthode empirique est l'expérimentation : la comparaison des résultats obtenus par recensement à des estimations à partir de l'enquête.

CHAPITRE IV : LE SONDAGE ALEATOIRE SIMPLE

Le sondage aléatoire simple (SAS) ou plan simple est le modèle de référence des méthodes de sondage probabiliste. Quoique rarement utilisé en pratique, il sert de base pour définir la plupart des sondages aléatoires. Il permet de présenter le vocabulaire de base des sondages aléatoires et de poser les problèmes pratiques de choix des estimateurs et de mesure de la précision des estimateurs.

I. Généralités sur les SAS

1. Vocabulaire du sondage aléatoire

1.1. Le taux de sondage

On appelle ***taux de sondage***, le rapport entre la taille de l'échantillon et celle de la population totale. Si nous notons :

- f ce taux
- n la taille de l'échantillon
- N la taille de la population

On a: $f = \frac{n}{N} * 100$

Dans les limites des contraintes liées à l'enquête, ce taux doit être le plus grand possible.

1.2. Les probabilités d'inclusion

Il existe deux types de probabilités d'inclusion : la probabilité d'inclusion d'ordre un (1) et celle d'ordre deux (2).
Le premier désigne la probabilité qu'a une unité, un individu d'appartenir à l'échantillon.
Le second est la probabilité que deux unités distinctes apparaissent conjointement dans l'échantillon.
Soient S l'échantillon d'un univers U

i un individu de cet univers

π_i la probabilité d'inclusion de cet individu.

On a: $\pi_i = P(i \in S)$

On montre que $\pi_i = P(i \in S) = E(I_S(i))$ où $I_S(i)$ est la ***fonction indicatrice*** ou ***fonction de cornefield***.

Soient i et j deux unités distinctes de l'univers $\pi_{ij} = P(i \in S et j \in S)$

Parallèlement, on définit le poids ou coefficient d'extrapolation d'un individu de l'échantillon comme l'inverse de sa probabilité d'inclusion.

Notons P_i ou W_i le poids de l'unité i de l'échantillon. On a : $W_i = \frac{1}{\pi_i}$

Concrètement, le poids d'un individu désigne le nombre d'individus de la population qu'il est censé représenter dans l'échantillon par rapport à la variable d'étude.

Dans le cas particulier d'un sondage aléatoire simple, on a : $\pi_i = f = \frac{n}{N}$ et $W_i = \frac{N}{n}$ pour tout i appartenant à S

Exemple : nous avons une population de 1000 individus. L'on désire en tirer 100 par un SAS.

$$\pi_i = P(i \in S) = \frac{100}{1000} = \frac{1}{10} \qquad W_i = 10$$

2. Notion de π–estimateurs

Deux estimateurs importants en théorie des sondages sont : le π–estimateur et l'estimateur de Hàjek. Nous présentons ici le π–estimateur

2.1. π-estimateur

Horvitz et Thompson (1952) ont présenté un estimateur linéaire sans biais d'un total valable pour tout plan de sondage :

$$\hat{T}(Y)=\sum_{i=1}^{n}\frac{y_i}{\pi_i}=\sum_{i=1}^{n}W_i y_i$$

Cet estimateur est appelé π–estimateur ou estimateur de Horvitz–Thompson ou encore estimateur par les valeurs dilatées.

II. Principe du sondage aléatoire Simple

De l'univers, on extrait un échantillon de taille n, en accordant à chaque unité la même chance d'être tirée. Il peut être tiré avec remise ou sans remise.

1. Le SAS avec remise

Dans un tirage avec remise, un échantillon de taille m est tiré selon la procédure : on sélectionne m unités de l'univers à probabilités égales à 1/N. les tirages des m unités se font de manière indépendante sur la même population.

2. Le SAS sans remise

Une fois une unité tirée, elle n'est plus prise en compte pour le tirage suivant. Ainsi la sélection est faite unité après unité.

III. Calcul des estimateurs dans un sondage aléatoire simple (SAS)

Une fois l'échantillon tiré, on cherche à estimer certaines grandeurs de la population ; les plus naturelles étant la moyenne, le total et la proportion. Comme il existe plusieurs échantillons possibles et que le sondage n'observe que l'un des échantillons, alors ces estimateurs qu'on cherche à calculer seront des variables aléatoires. C'est pourquoi on parlera, par exemple, de variance d'un estimateur de la moyenne.

1. Notations

Rappel : on désignera par Y la variable d'intérêt et on notera : Y_α la valeur de Y pour l'unité statistique appartenant à l'univers et par Y_i la valeur de Y pour l'unité statistique i appartenant à l'échantillon. Les caractéristiques de Y seront notées comme suit :

1.1. Sur la population

- moyenne de Y: $\bar{Y}=\frac{1}{N}\sum_{\alpha=1}^{N}Y_\alpha$
- total de Y: $T(Y)=\sum_{\alpha=1}^{N}Y_\alpha=N\bar{Y}$
- Variance de Y: $\sigma^2=\frac{1}{N}\sum_{\alpha=1}^{N}\left(Y_\alpha-\bar{Y}\right)^2=\frac{N-1}{N}S^2$ où $S^2=\frac{1}{N-1}\sum_{\alpha=1}^{N}\left(Y_\alpha-\bar{Y}\right)^2$

1.2. Sur l'échantillon

- Moyenne: $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$
- Total: $t(Y) = \sum_{i=1}^{n} y_i = n\bar{y}$
- Variance: $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$

2. Estimateur de la moyenne et du total

2.1. Estimation

Pour estimer la moyenne $\bar{Y}$ et le total T (tous les deux inconnus) sur l'univers, on utilise la moyenne empirique calculée sur l'échantillon.

En effet, on montre que $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$ est un estimateur sans biais de $\bar{Y}$ et que $t=N\bar{y}=\sum_{i=1}^{n}\frac{N}{n} y_i$ est un estimateur sans biais du total T.
N/n est le coefficient d'extrapolation.

2.2. Précision des estimateurs

On mesure la précision d'un estimateur soit par sa variance, soit par son coefficient de variation, soit par l'étendue de son intervalle de confiance.

2.2.1. Cas de la moyenne

On montre que :

$$Var(\bar{y})=\begin{cases}\frac{\sigma^2}{n} & \text{Avec remise}\\ \frac{N-n}{N-1}*\frac{\sigma^2}{n} & \text{Sans remise}\end{cases}$$

On remarque que le sondage sans remise est plus précis que le sondage avec remise ; mais lorsque la population est très grande (N grand), les deux tirages deviennent équivalents en précision.

2.2.2. Cas du total

Comme $t = N\bar{y}$ alors $V(T)=N^2V(\bar{y})$ et donc on déduit de la précision de la moyenne que la variance du total s'établit comme suit :
*

$$Var(t) = \begin{cases} N^2 * \frac{\sigma^2}{n} & \text{Avec remise}\\ N^2\frac{N-n}{N-1} * \frac{\sigma^2}{n} & \text{Sans remise}\end{cases}$$

La précision des estimateurs de la moyenne et du total dépend de :

- La taille n de l'échantillon : le sondage devient de plus en plus précis quand on tend vers un recensement.
- La variance de la variable d'intérêt Y : le sondage sera d'autant plus précis que la population est homogène.

Comme $\sigma^2=V(Y)$ est inconnue, il faut chercher à l'estimer. On montre que $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ est un estimateur sans biais de $\frac{N}{N-1}V(Y)$, pour le cas sans remise, et de V(Y) pour le cas avec remise. On en déduit que les estimateurs des précisions sont les suivantes :

Pour la moyenne $\hat{V}(\bar{Y})=\begin{cases}\frac{s^2}{n} & \text{Avec remise}\\ (1-f)\frac{s^2}{n} & \text{Sans remise}\end{cases}$

Pour le total $\hat{V}(t)=\begin{cases}N^2*\frac{s^2}{n} & \text{Avec remise}\\ N^2*(1-f)\frac{s^2}{n} & \text{Sans remise}\end{cases}$

<u>Remarque</u> : En pratique, on utilise aussi le coefficient de variation pour mesurer la précision d'un estimateur. Ainsi on dira qu'un sondage est précis à 2 % près pour l'estimateur d'une moyenne $\bar{Y}$ pour signifier que le $CV=\frac{\sigma(\bar{y})}{\bar{y}}=2\%$.

3. <u>Estimateur d'une proportion</u>

Il est fréquent dans un sondage que l'on s'intéresse à une proportion de la population présentant un intérêt particulier. Par exemple, si l'on réalise un sondage auprès des ménages de la capitale, on peut chercher à estimer la proportion des ménages ayant accès à l'eau courante dans l'ensemble des ménages.
On montre que l'estimation d'une proportion se ramène à celle d'une moyenne. En effet, notons N_D le nombre d'individus présentant l'intérêt particulier en question et on cherche à estimer la proportion $P_D=\frac{N_D}{N}$.
Cette proportion peut être considérée comme la moyenne d'une variable aléatoire de Bernoulli. En effet, soit la variable Z définie par $Z_\alpha=1$ si $\alpha\in D$; 0 sinon. On a : $P_D=\bar{Z}=\frac{1}{N}\sum_{i=1}^{N}Z_i$. Et donc, d'après l'estimateur de la moyenne, P_D peut-être estimée sans biais par la proportion dans l'échantillon ie par $p_D=\frac{n_D}{n}$ où n_D est le nombre d'unités de l'échantillon qui appartiennent au domaine d'intérêt D.
De la même façon, on montre que l'estimateur de la précision du sondage dans le calcul d'une proportion est :

$\hat{V}(p_D)=\begin{cases}\frac{p(1-p)}{n-1} & \text{Avec remise}\\ (1-f)\frac{p(1-p)}{n-1} & \text{Sans remise}\end{cases}$

Si n est grand et f négligeable (f<< 1), alors $\hat{V}=\frac{p(1-p)}{n}$.

4. Estimation par intervalle de confiance

4.1. Estimation

On sait d'après le théorème Central limite que lorsque n est grand (n > 29), la distribution de $\bar{Y}$ s'ajuste à une loi normale de moyenne $\bar{y}$ et de variance $V(\bar{y})$. D'autre part, on sait qu'une distribution normale N (m, σ²) a 95 % de ses valeurs dans l'intervalle [m-1,96σ, m+1,96σ]. On en déduit que la moyenne $\bar{Y}$ a 95 % de chance de se retrouver dans l'intervalle $\left[\bar{y}-1{,}96\sqrt{V(\bar{y})};\bar{y}+1{,}96\sqrt{V(\bar{y})}\right]$, ce qui peut encore s'écrire. $Prob\left(\bar{Y}\in IC\right)=0.95$ est appelé ***intervalle de confiance ou fourchette d'estimation à 95 %***.
L'intervalle de confiance pour la proportion est : $IC(p_D)=[p_D\pm 1{,}96\hat{\sigma}(p_D)]$

4.2. La précision

On déduit de cet intervalle, deux mesures de précision :

- la ***précision absolue*** $\varepsilon=1{,}96\sqrt{V(\bar{y})}$ et
- la ***précision relative*** $k=\frac{\varepsilon}{\bar{y}}=1{,}96.\frac{\sigma(\bar{y})}{\bar{y}}$.

Exemple : Sur un échantillon de taille n = 100 ménages, on observe que n_D = 10 ménages ont accès à l'eau courante.

On a donc : p = 10 % et $\varepsilon=2*\sqrt{\frac{0.1*0.9}{100}}=0.06$ et $IC=[0.1\pm 0.06]=[4\%;16\%]$

La précision absolue vaut donc 0.06. Ce qui s'interprète ainsi : ''d'après l'échantillon, l'estimation de p se fait à plus ou moins 6 points''.
La précision relative vaut 60 %. Ce qui signifie que la marge d'incertitude est de 60 % de la quantité estimée (ici la proportion).

Remarque :

- L'intervalle de confiance est ici trop large, l'estimation est donc peu précise. Cela peut-être dû à la taille de l'échantillon qui est petite.
- L'estimation par intervalle de confiance est une méthode d'estimation doublement prudente. D'abord, elle ne donne pas une valeur ponctuelle de l'estimateur, mais un intervalle de valeurs possibles. Ensuite, elle indique qu'il y a risque (en général de 5 %) que la vraie valeur soit en dehors de l'intervalle.

On peut utiliser les intervalles de confiance pour déterminer la taille de l'échantillon à observer.

IV. Détermination de la taille de l'échantillon dans un SAS

Une des questions délicates qui se pose au statisticien avant même de commencer le sondage est le choix de la taille de l'échantillon : Combien d'unités faut-il enquêter pour répondre valablement aux objectifs de l'enquête ? Bien entendu, il est clair que pour obtenir de bons estimateurs, il suffit de choisir une taille de l'échantillon beaucoup plus grande. Mais jusqu'où peut-on aller dans ce choix ?
Dans l'absolu, il n'existe pas de réponse directe quant à la taille ''optimale'' de l'échantillon à retenir. Dans la pratique, on peut avoir à trancher entre deux contraintes : une contrainte budgétaire et une contrainte de précision.

1. Contrainte budgétaire

La méthode budgétaire consiste à déterminer la taille de l'échantillon en tenant compte du budget total de l'enquête. Ainsi, si C désigne ce budget total et c le coût unitaire total par individu enquêté (prime, per diem, reproduction des questionnaires, codification, saisie …), alors on a n = C/c.
Cette taille peut ne pas donner de ''bons'' résultats.

2. Contrainte de précision

Si la contrainte budgétaire est faible, on peut se fixer un niveau de précision et de crédibilité à atteindre, on calcule ensuite la taille ''optimale'' de l'échantillon n et on définit le budget permettant de garantir cette précision.

2.1. Détermination de la taille de l'échantillon à partir d'une moyenne

Soit Y la variable d'intérêt et ε la précision absolue à ne pas dépasser. La crédibilité est mesurée par le niveau de confiance 1- α (=95% en général).
On sait, d'après l'estimation par l'intervalle de confiance, que: $prob\left(\bar{Y}\in\left[\bar{y}-2\sqrt{V(\bar{y})};\bar{y}+2\sqrt{V(\bar{y})}\right]\right)=0.95$.
Lorsque le tirage de l'échantillon peut être assimilé à un tirage avec remise (ce qui est le cas lorsque le taux de sondage est inférieur à 10%) alors on peut utiliser l'approximation $V(\bar{y})=\frac{\sigma^2}{n}$.

Dès lors, on détermine *n* à partir de la contrainte $2\sqrt{\frac{\sigma^2}{n}}\leq\varepsilon$, soit $n\geq\frac{4\sigma^2}{\varepsilon^2}$.

Si la précision est donnée en *valeur relative(k),* on a $\frac{2}{\bar{Y}}\sqrt{\frac{\sigma^2}{n}}\leq k$ soit $n\geq\frac{4(CV)^2}{k^2}$.

La difficulté ici est qu'il faut connaître, même approximativement, l'écart-type ou le coefficient de variation CV de la grandeur à estimer. Or, ces indicateurs ne sont pas connus avant l'enquête. Il faut alors disposer des informations à priori sur leurs valeurs.
Dans la pratique, on lève cette difficulté en utilisant des estimations à partir d'enquêtes similaires réalisées dans un passé pas trop éloigné. On peut aussi réaliser une pré-enquête sur un échantillon restreint, à partir de laquelle on évaluera approximativement l'ordre de grandeur des paramètres. Si on dispose des informations sur la variable X corrélée avec la variable d'intérêt, on peut aussi déterminer l'échantillon comme si cette variable X était la variable d'intérêt dont on cherche à estimer la moyenne $\bar{X}$; l'idée étant que ce qui est bon pour l'un doit l'être pour l'autre.

Un exemple
Dans une population, la dépense moyenne consacrée à un produit donné est de l'ordre de 0.4 fois son écart type. Quelle taille d'échantillon permet d'apprécier à 10% près la valeur de la dépense moyenne, au seuil de probabilité 0.95 ?
On a $CV=\frac{\sigma}{\bar{Y}}=2.5$ et k=0.10 $n\geq\frac{4*2.5^2}{0.1^2}=2500$.
On conclut donc qu'un échantillon de 2500 ménages est suffisant.

2.2. Détermination de la taille de l'échantillon à partir des proportions.

La détermination de la taille de l'échantillon devient facile à résoudre quand on s'intéresse à une proportion.
En effet, lorsque les tirages peuvent être considérés comme indépendants, on a $IC(p_D)=\left[p_{Dd}\pm2\hat{\sigma}(p_D)\right]$ avec $\hat{\sigma}(p_D)=\sqrt{\frac{p(1-p)}{n}}$.

Si la précision k est donnée en terme absolue, il faut choisir n tel que : $1{,}96\sqrt{\frac{p(1-p)}{n}} \leq k$, soit

$n \geq \frac{1{,}96^2}{k^2} p(1-p)$

Pour que la précision soit au moins égale à k% de p (précision fixée en valeur relative), il faut choisir n tel que : $1{,}96\sqrt{\frac{p(1-p)}{n}} \leq kp$, soit $n \geq \frac{1{,}96^2}{k^2}\frac{1-p}{p}$.

On voit que la taille de l'échantillon est une fonction décroissante p. Cela signifie que l'échantillon est d'autant plus grand que le phénomène étudié est rare.
Ici encore, il faut avoir une idée sur la proportion que l'on cherche à estimer. Si on ne connaît aucune information à priori sur le phénomène on se place dans le cas plus incertain où p=0.5. On sait que dans ce cas, on majore le risque d'erreur ainsi que la taille de l'échantillon.

Un exemple

On s'intéresse à l'estimation de la proportion de personnes atteintes d'une maladie M dans une localité de 1500 habitants. On sait, par ailleurs, que 3 personnes sur 10 sont ordinairement touchées par cette maladie dans une localité de la même région. On se propose de sélectionner un échantillon au moyen d'un Sondage Aléatoire Simple.

On a ici p=$\frac{3}{10}$=0.3. Pour un risque d'erreur fixé à k=10%, $n = \frac{1{,}96^2}{0.1^2} 0.7 * 0{,}3 = 81$

V. Comment tirer l'échantillon ?

Avant de présenter les méthodes de tirage, il convient d'inventorier les voix classiques d'extraction de nombres aléatoires. A ce sujet, trois principales voies se dégagent :

- les nombres aléatoires sur calculatrice scientifique
- les tables aléatoires et
- les nombres aléatoires sous Microsoft Excel

1. Notion de nombres aléatoires ou pseudo aléatoire

Une série de nombres est dite aléatoire lorsque tous les nombres ont la même chance d'être tiré au sort. Ainsi, à chaque tirage, l'on n'est pas en mesure, à priori de prédire le nombre qui apparaîtra. Ces nombres sont importants dans les procédures de tirage aléatoires.
L'expérience à montrer que les nombres choisis au hasard par un humain ne sont aléatoires car influencés par sa culture, son entourage, … et donc l'idée s'est imposée au statisticien de concevoir des procédures mécaniques de générations de nombres aléatoires ou pseudo-aléatoires dont nous présentons trois types.

1.1. Nombre aléatoire sur une calculatrice scientifique

Certaines calculatrices (scientifiques notamment) possède une touche, voire fonction, random (RND) qui permet d'obtenir des nombre aléatoires selon une loi uniforme sur l'intervalle [0;1[.

1.2. Les tables aléatoires

On désigne sous l'appellation de table de nombres aléatoires, des tables de nombres tirés par des statisticiens qui permettent d'obtenir des nombres aléatoires en leur appliquant un sens de lecture donné. Les plus connus sont :

- La table de Tipett,
- La table de Fischer et Yates,

- La table de Kendall et Babington Smith,
- La table de Burke Horton,
- La table de la Rand corporation.

On les trouve souvent en annexe des documents de probabilités ou de statistiques.

1.3. Nombre aléatoire sous Microsoft Excel

L'obtention de nombre aléatoire, sous Microsoft Excel, se fait selon la procédure décrite dans l'encadré ci-après. La syntaxe ***Alea*** permet d'obtenir des nombres aléatoires selon une loi uniforme sur [0;1[. En la modifiant un peu, l'on peut tirer des nombres aléatoires selon une loi uniforme de n'importe quel intervalle [a;b[.

Encadré 1: procédure de génération de nombre aléatoire sous Microsoft Excel
ALEA Renvoie un nombre aléatoire supérieur ou égal à 0 et inférieur à 1. Un nouveau nombre aléatoire est renvoyé chaque fois que la feuille de calcul est recalculée. **Syntaxe** **ALEA**( ) **Notes** Pour générer un nombre réel aléatoire compris entre a et b, utilisez : ALEA()*(b-a)+a Si vous voulez utiliser ALEA pour générer un nombre aléatoire qui ne change pas chaque fois que la cellule est recalculée, vous pouvez taper =ALEA() dans la barre de formule, puis appuyer sur F9 pour transformer la formule en nombre aléatoire. **Exemple** Pour générer un nombre aléatoire supérieur ou égal à 0 mais inférieur à 100: ALEA()*100

Plusieurs des procédures de tirage de l'échantillon se font par génération de nombres aléatoires d'où l'intérêt de ces nombres aléatoires. Le sondeur pourra choisir, selon la disponibilité, l'un ou l'autre des moyens présentés pour générer des nombres aléatoires.
Pour la suite, nous supposons disposer d'une population de N individus rangés de façon séquentielle de 1 à N. Et que l'on souhaite extraire de cette population un échantillon de taille n de façon aléatoire.

2. Tirage systématique à probabilités égales

Deux types de procédures permettent de tirer des individus à probabilités égales :

- Le tirage systématique et,
- Les algorithmes de tirage.

Pour le second, nous ne présentons que ceux qui présentent de bonnes propriétés. Toutefois, il faut remarquer qu'il en existe plusieurs autres.
Le tirage systématique est sans doute le mode de tirage le plus célèbre à cause de sa facilité d'exécution.

2.1. Procédure

Il consiste à calculer d'abord un pas de tirage, ensuite à tirer au hasard un premier individu en début de liste, et partant de celui-ci à descendre dans la base en faisant les sauts. Formellement, les étapes du tirage systématique se présentent de la façon suivante :

- Calculer le pas $p=\frac{N}{n}$
- Tirer au hasard un nombre entier entre 1 et p. Soit x ce nombre. On retient dans l'échantillon l'unité de rang x.
- Ajouter à x le pas. On retient l'unité de rang x_2 = Arr (x+p)
- De façon générale, à l'étape i. on sélectionne l'unité de rang x_i = Arr [x + (i-1)p]
- Continuer jusqu'à obtenir les n unités-échantillon.

2.2. Avantage

Il est Facile à mettre en œuvre. En outre, il donne une meilleure répartition de l'échantillon sur la population et une plus grande précision, surtout si le classement dans la base est lié à un critère corrélé avec la variable d'intérêt.
Par exemple si on classe les ménages selon leur taille, le tirage systématique assurera d'avoir à la fois des ménages de faible taille et des ménages de grande taille.

2.3. Contre-indication

Si les unités sont rangées de façon à présenter, pour la variable d'intérêt, une certaine périodicité, on risque de sélectionner des unités très particulières, surtout si le pas est égal ou multiple de la périodicité. Dans de tels cas, mieux vaut recourir à un algorithme de tirage.

2.4. Les algorithmes de tirage

Les algorithmes de tirage sont la transcription, sous forme algorithmique, d'un plan de sondage. La construction de tels algorithmes doit obéir à des conditions :

- Donner les mêmes chances aux individus d'être tirés,
- Etre facile à programmer,
- Pouvoir s'appliquer ci-possible en une seule lecture du fichier.

Conclusion

Le sondage aléatoire simple reste un modèle de référence en théorie de sondages. Elle est facile à mettre en œuvre et parfois elle peut donner une répartition « satisfaisante » de l'échantillon. En pratique cependant, cette méthode n'est jamais appliquée toute seule. Elle est utilisée en combinaison avec d'autres méthodes de sondage. Par exemple lorsque les unités statistiques de la base de sondage présentent une forte hétérogénéité vis-à-vis de la variable d'intérêt, un échantillonnage par sondage aléatoire simple peut conduire à ne retenir que des unités statistiques de mêmes catégories. Pour corriger ce déséquilibre et améliorer la qualité des estimateurs, on peut repartir l'échantillon entre les différentes catégories de la population et procéder à un sondage aléatoire simple au sein de chaque catégorie. Ce principe est à la base du sondage stratifié.

CHAPITRE V : SONDAGES STRATIFIES A PRIORI

I. Justification, objectifs et principe de base du sondage stratifié

Nous avons vu que la précision des estimateurs du SAS dépend de la dispersion de la variable d'intérêt Y dans la population totale. Si la population est homogène vis-à-vis de Y (V(Y) faible) alors les estimateurs seront d'une bonne précision. Par contre, si elle est hétérogène, les estimateurs seront mauvais en précision quoique sans biais.
On peut donc rendre les estimateurs du SAS beaucoup plus précis en constituant d'abord des groupes ou strates homogènes, c'est-à-dire des groupes à l'intérieur desquels les unités présentent une certaine ressemblance vis-à-vis de Y, ensuite, procéder de façon indépendante au tirage des unités à l'intérieur de chaque groupe. Tel est l'objectif principal du sondage stratifié. En outre, il permet d'assurer une précision suffisante pour les estimations relatives aux sous-ensembles pris comme des strates et de parer aux non-réponses éventuelles.
Au delà des trois questions classiques que soulèvent les méthodes d'échantillonnage déjà rencontrés à savoir :

- Comment extrapoler les résultats obtenus sur l'échantillon à la population totale,
- Quelle est la précision des estimateurs,
- Comment fixer la taille de l'échantillon,

L'échantillonnage stratifié pose deux autres problèmes :

- Comment constituer les strates,
- Comment repartir l'échantillon dans les strates,

Pour répondre à ces différentes questions, nous allons supposer, dans un premier temps, que les strates sont constituées et que l'échantillon a déjà été reparti afin de résoudre le problème de l'extrapolation et de la précision. Ce n'est qu'ensuite que nous discuterons le problème de la constitution des strates et de la répartition de l'échantillon.

II. Description de la population et notations

La population est de taille N et divisée en H strates (H>= 2) qui forment des sous-populations (P_1, P_2, …, P_H) d'effectifs respectifs (N_1, N_2, …, N_H).
Y désigne toujours la variable d'intérêt.

Moyenne de Y sur toute la population : $\bar{Y}=\frac{1}{N}\sum_{h=1}^{H}\sum_{\alpha=1}^{N_h}Y_{h\alpha}=\frac{1}{N}\sum_{h=1}^{H}N_h\bar{Y}_h$ où $\bar{Y}_h=\frac{1}{N_h}\sum_{\alpha=1}^{N_h}Y_{h\alpha}$ la moyenne dans la strate h

Total : $T=\sum_{h=1}^{H}\sum_{\alpha=1}^{N_h}Y_{h\alpha}=\sum_{h=1}^{H}T_h$ où $T_h=\sum_{\alpha=1}^{N_h}Y_{h\alpha}$ total dans la strate h

Dispersion totale :

$$S^2=\frac{1}{N-1}\sum_{h=1}^{H}\sum_{\alpha=1}^{N_h}(Y_{h\alpha}-\bar{Y})^2=\frac{1}{N-1}\sum_{h=1}^{H}(N_h-1)S^2_h+\frac{1}{N-1}\sum_{h=1}^{H}N_h(\bar{Y}_h-\bar{Y})^2$$

$$\text{où } S^2_h=\frac{1}{N_h-1}\sum_{\alpha=1}^{N_h}(Y_{h\alpha}-\bar{Y}_h)$$

La présente description fournit les notations utiles pour mener à bien l'estimation et le calcul de la précision de ces derniers.

III. Estimation et calcul de précision

L'échantillon à enquêter est sélectionné parmi toutes les strates. Le tirage à l'intérieur de chaque strate se fait selon la procédure d'un SAS. On choisit dans la strate h, n_h individus. On a : $n=\sum_{h=1}^{H} n_h$

1. Estimation de la moyenne, de la proportion et du total

Dans la strate h, on a :

Moyenne : $\bar{y}_h=\frac{1}{n}\sum_{i=1}^{n_h} y_{hi}$ est un estimateur sans biais de $\bar{Y}_h$

Proportion : $p_h=\frac{n_{hD}}{n_h}$ est un estimateur sans biais de $P_h=\frac{N_{hD}}{N_h}$

$t_h = N_h\bar{y}_h = \sum_{i=1}^{n_h}\frac{N_h}{n_h} y_{hi}$ est un estimateur sans biais du total T_h de la strate h

Au niveau de la population, on a :

Moyenne : $\bar{y}_{str}=\frac{1}{N}\sum_{h=1}^{H} N_h\bar{y}_h$ est un estimateur sans biais de $\bar{Y}=\frac{1}{N}\sum_{h=1}^{H} N_h\bar{Y}_h$

Proportion : $p=\frac{1}{N}\sum_{h=1}^{H} N_h p_h$ est un estimateur sans biais de $P=\frac{N_D}{N}$

Total : $t_{str}=\sum_{h=1}^{H} N_h\bar{y}_h$ est un estimateur sans biais de T

2. Calcul de la précision

On montre que pour le sondage stratifié,

$$V(\bar{y}_{str})=\begin{cases}\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{S^2_h}{n_h} & \text{(avec remise)}\\ \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2(1-f_h)\frac{S^2_h}{n_h} & \text{(sans remise)}\end{cases}$$

avec

$$S^2_h=\frac{1}{N_h-1}\sum_{\alpha=1}^{N_h}\left(Y_{h\alpha}-\bar{Y}_h\right)$$

$f_h=\frac{n_h}{N_h}$: taux de sondage dans la strate h.

On obtient un estimateur de la précision en remplaçant S^2_h par s^2_h.

IV. Constitution des strates et Répartition de l'échantillon entre elles

Si les H strates sont déjà constituées, il se pose tout naturellement la question de la répartition des n unités suivant les différentes strates. Lorsque la stratification poursuit un objectif de précision locale, le problème de la répartition de l'échantillon ne se pose pas. En fonction de la précision que l'on s'est fixée, pour chaque strate, on détermine séparément la taille nécessaire. La taille totale sera la somme des différentes tailles.

On s'intéresse ici au cas où l'on poursuit un objectif de précision globale. Dans ce cas, on distingue deux types de répartition : l'allocation proportionnelle et l'allocation optimale.

1. Répartition de l'échantillon entre les strates

1.1. L'allocation proportionnelle

L'allocation proportionnelle correspond au cas où l'échantillon présente la même structure que la population totale en termes d'effectifs. Elle consiste à fixer un taux de sondage identique pour toutes les strates, soit $f_h = \frac{n_h}{N_h} = C_{te} = \frac{n}{N}$.

L'allocation proportionnelle est généralement utilisée lorsque la seule donnée disponible est le nombre d'unités par strate.

L'estimateur de la moyenne devient dans ce cas :

$$\bar{y}_{str} = \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} y_{hi} \text{ et sa variance } \hat{V}(\bar{y}_{str}) = \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} s_h^2$$

On en déduit trois propriétés fondamentales importantes du sondage proportionnel :

1. Tous les individus de la base ont les mêmes probabilités d'être choisi, ces probabilités sont égales au taux de sondage f = n/N.
2. L'estimateur de la moyenne (ou de la proportion) est la moyenne (ou la proportion) empirique calculée sur l'échantillon. Pour cela, on dit que le sondage proportionnel est un sondage ***auto-pondérée*** ou encore qu'il se dépouille comme un recensement.

1.2. L'allocation optimale de Neyman

La répartition proportionnelle présente une certaine faiblesse. Elle suppose en effet que les strates sont uniformes vis-à-vis de la variable d'intérêt, ce qui n'est pas toujours le cas. L'allocation optimale de Neyman corrige ce défaut. Elle est basée à la fois sur la taille et la variance des strates.

L'allocation de Neyman minimise donc la variance globale en tenant compte de la contrainte de budget. On l'appelle pour cela l'***allocation à variance minimale***.

Si on désigne par c_h le coût moyen pour enquêter un individu appartenant à la strate h, on a :

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

Il y a plusieurs solutions (n_1, n_2, …, h_H) possibles à cette équation.

L'allocation optimale sera celle qui donne la plus grande précision sous la contrainte de coût et donc celle qui est solution du programme : $\begin{cases} MinV(\bar{y}_{str}) \\ \sum_{h=1}^{H} n_h c_h \le C \end{cases} \Leftrightarrow \begin{cases} Min \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h) \frac{S^2_h}{n_h} \\ \sum_{h=1}^{H} n_h c_h \le C \\ \sum_{h=1}^{H} n_h = n \end{cases}$

Ce programme est équivalent à : $\begin{cases} Min \sum_{h=1}^{H} \frac{N^2_h S^2_h}{n_h} \\ \sum_{h=1}^{H} n_h c_h \le C \\ \sum_{h=1}^{H} n_h = n \end{cases}$

On obtient : $n_h = \frac{N_h S_h}{\sqrt{c_h}} \frac{C}{\sum_{h=1}^{H} N_h S_h \sqrt{c_h}}$ l'allocation de Neyman

Remarque :

- S'il existe des strates pour lesquelles $n_h > N_h$, alors on prend dans ces strates $n_h = N_h$ et on reprend le calcul pour les autres strates en raisonnant avec le reste de l'échantillon à pourvoir.

- De même, comme les solutions sont décimales, on prendra en pratique les arrondis des valeurs trouvées.

Cas particulier où $c_h = 1$ qui correspond au cas où le coût unitaire est le même dans toutes les strates.

On obtient $n_h = n\dfrac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$

Et $V(\bar{y}_{str}) = \sum_{h=1}^{H}\left(\dfrac{N_h}{N}\right)^2(1-f_h)\dfrac{S_h^2}{n_h} = \sum_{h=1}^{H}\left(\dfrac{N_h}{N}\right)^2\dfrac{S_h^2}{n_h} - \sum_{h=1}^{H} f_h \dfrac{S^2_h}{n_h}$

La difficulté à utiliser la répartition optimale de Neyman est qu'il faut avoir une idée sur l'ordre de grandeur de la variance des strates. En pratique, on se réfère à des études antérieures ou à une enquête pilote à laquelle on assigne cet objectif.
Plus la variance de la strate est élevée par rapport à celle des autres, plus la taille de l'échantillon doit être grande dans cette strate.

2. Constitution des strates

La constitution des strates pose trois questions : la question du mode de définition des strates, celle du choix des critères de stratification et celle du choix du nombre de strate.

2.1. Mode de définition des strates

Il existe deux modes de définitions des strates : la définition en extension et la définition par une variable auxiliaire.
La première est faite en général pour assurer la représentativité de l'échantillon dans des sous-groupes de la population. Par exemple, pour les enquêtes sur le niveau de vie en Côte d'Ivoire, l'on stratifie selon cinq (5) strates : Abidjan, autres villes, forêt ouest, savane, forêt est. Ces strates sont définies par la liste exhaustive des sous-préfectures qui les composent.
Pour la seconde, l'appartenance des individus à une strate donnée est définie par la réalisation d'une variable. Dans cette seconde option, on utilise une matrice d'informations auxiliaires X qu'on pense être corrélé avec la variable d'intérêt Y. Ces variables auxiliaires X sont appelées variables ou critères de stratification. Par exemple, l'on peut décider de stratifier :

- Les entreprises d'après le nombre de salariés qu'elles emploient ;
- Les exploitations agricoles d'après leur superficie ;
- Les ménages d'après leur revenu présumé ;
- Etc.

Cette optique impose le choix du critère de stratification.

2.2. Choix des critères de stratification

Le critère de stratification doit être corrélé avec la ou les variables d'intérêt afin d'assurer une meilleure efficacité à la stratification. Cette variable doit –être disponible dans la base de sondage afin de permettre de classer les individus, sans ambigüité dans les classes formées. En effet, une stratification optimale pour l'estimation du volume de la production peut s'avérer inefficace pour l'estimation du mode de faire valoir, de la participation à une coopérative, etc. Ainsi, la stratification dans une enquête multithématique pose d'énorme problème parce qu'il faudra trouver une ou des variables corrélées aux variables d'intérêt. Toutefois, très souvent, la taille est un bon critère qui est corrélé à beaucoup de variables économiques d'où l'idée de la choisir systématiquement.
En outre, une fois le critère choisi, s'il est quantitatif, il se pose le problème du découpage de la variable en strate qui a été largement traité mathématiquement par Dalenius.

2.3. Choix du nombre de strates

Deux phénomènes sont à intégrer lors du choix du nombre de strate : d'une part, l'augmentation du nombre de strate augmente la précision des estimateurs et, d'autres part, l'augmentation du nombre de strate implique plus de travail en plus de réduire le nombre d'individus dans les strates ce qui pourraient aboutir à une situation où l'on ne puisse pas calculer la variance de l'estimateur faute de données suffisantes.
En stratifiant les strates, on aboutit très vite à un rendement décroissance qui commande donc de s'arrêter. En pratique, si la stratification est basée sur des variables auxiliaires, il est conseiller de ne pas dépasser plus de quatre (4) critères car au delà le nombre de strates devient très élevé.

V. Comparaison avec le SAS

Le sondage stratifié à allocation proportionnelle est plus précis que le sondage aléatoire simple. En effet:

$$V(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{S_h^2}{n_h} = \frac{1-f}{n}\left[S^2 - \sum_{h=1}^{H}\frac{N_h}{N}(\bar{Y}_h - \bar{Y})^2\right] = \frac{1-f}{n}S^2 - A \text{ car}$$

$$S^2 = \frac{N}{N-1}\left[\sum_{h=1}^{H}\left(\frac{N_h - 1}{N_h}\right)\frac{N_h}{N}S^2{}_h + \sum_{h=1}^{H}\frac{N_h}{N}(\bar{Y}_h - \bar{Y})^2\right]$$

En supposant N_h grand pour tout h (N_h >=10), on a: $S^2 = \sum_{h=1}^{H}\frac{N_h}{N}S^2{}_h + \sum_{h=1}^{H}\frac{N_h}{N}(\bar{Y}_h - \bar{Y})^2$ soit $V(\bar{y}_{str}) \le V(\bar{y}_{SAS})$

Et $V(\bar{y}_{str}) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\frac{S_h^2}{n_h} - \sum_{h=1}^{H} f_h\frac{S^2{}_h}{n_h} \approx \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{N}S_h^2$

$$V_{opti}(\bar{y}) \le V_{prop}(\bar{y}) \le V_{SAS}(\bar{y})$$

Pour conclure

Par principe, la stratification améliore toujours la précision des résultats car elle concentre les valeurs observées autour des valeurs moyennes dans chaque strate. Le gain dû à la stratification sera d'autant plus élevé que les variables de stratification sont plus corrélées avec la variable d'intérêt. Il n'y a donc pas de contre-indication à la stratification. On peut même stratifier les strates tout en restant dans les limites d'une taille raisonnable par strate.

CHAPITRE VI : SONDAGES A PLUSIEURS DEGRES, PAR GRAPPES

I. Justification et principe

Pour l'étude de certains phénomènes (consommation, revenu, emploi du temps des ménages, production agricole, …), la disponibilité d'une base de sondage exhaustive et mise à jour n'est pas garantie. Par contre, on peut disposer de listes exhaustives sur des groupes d'unités statistiques (îlots, quartier, district de recensement, …). Au lieu de tirer directement les unités, on peut procéder en escalier :

- On tire d'abord un certain nombre de groupes d'unités et ;
- Dans chaque groupe, on tire les unités à enquêter.

On dit qu'on fait un **sondage à deux degrés**.

Plus généralement, le principe du sondage à plusieurs degrés consiste à :

- Partitionner la population en M groupes, appelés ***unités primaires*** ;
- Suivant une procédure de tirage aléatoire, on tire m groupes : ceci correspond au premier degré du tirage ;
- Dans chaque groupe h tiré (h = 1, …, m), on établit la liste de toutes les unités et on tire un nombre n_h, appelés ***unités secondaires*** : c'est le deuxième degré du tirage ;
- On tire, dans chaque unité secondaire tirée, un nombre d'unités appelées ***unités tertiaires***,
- etc.

Cette méthode présente un certain nombre d'avantages : Elle ne génère pas une dispersion géographique des unités échantillons dans la population. Elle réduit donc le coût des déplacements et nécessite, par conséquent, moins de budget que le sondage aléatoire simple.

On n'a d'ailleurs pas besoin d'avoir une base de sondage complète des unités d'observation. Enfin, on verra qu'on n'a pas nécessairement besoin de connaître l'effectif de la population N pour estimer les paramètres.

En revanche, cette méthode a l'inconvénient d'être moins précise, que le SAS pour une même taille d'échantillon. La perte de précision par rapport au SAS est appelée effet de sondage.

II. Description de la population et de l'échantillon : estimation

On se limite dans la suite au sondage à deux degrés.

On considère que la population de taille N est repartie en M groupes ou ***Unités Primaires*** (UP) de taille moyenne $\overline{N}$.

Soit m le nombre d'UP à tirer et N_h le nombre d'unités secondaires de l'UP_h, h=1,…, M.

N et N_h peuvent-être connus ou inconnu. Soit n la taille de l'échantillon en US et n_h le nombre d'US à tirer dans l'UP_h.

$N=\sum_{h=1}^{M} N_h = M\overline{N}$ et $n=\sum_{h=1}^{m} n_h$

$f_1=\frac{m}{M}$ est le taux de sondage au premier degré.

$f_{2h}=\frac{n_h}{N_h}$ est le taux de sondage au second degré dans l'UP_h.

$f=\frac{n}{N}$ est le taux de sondage global.

1. Description de la population

La population peut-être décrite par les caractéristiques suivantes :

$T_h=\sum_{h=1}^{N_h} Y_{h\alpha}$ total de Y dans UP_h

$\bar{Y}_h = \frac{1}{N_h}\sum_{h=1}^{N_h} Y_{h\alpha}$ Moyenne de Y dans l'UP_h

$S_1^2 = \frac{1}{M-1}\sum_{h=1}^{M}\left(T_h - \bar{T}\right)^2$ la dispersion des totaux

$S_{2h}^2 = \frac{1}{N_h - 1}\sum_{\alpha=1}^{N_h}\left(Y_{h\alpha} - \bar{Y}_h\right)^2$ la dispersion de Y dans l'UP_h

$\bar{Y} = \frac{T}{N} = \frac{1}{N}\sum_{h=1}^{M} N_h \bar{Y}_h$ la moyenne de Y dans la population

$s_1^2 = \frac{1}{m-1}\sum_{h=1}^{m}\left(t_h - \frac{t}{m}\right)^2$ dispersion empirique des totaux au sein de l'échantillon

$s_{2h}^2 = \frac{1}{n_h - 1}\sum_{h=1}^{N_h}\left(y_{hi} - \bar{y}_h\right)^2$ dispersion empirique de Y dans l'échantillon.

2. Estimation du total et de la moyenne : cas général

On suppose que les tirages des US sont indépendants d'une UP à une autre. On note π_h la probabilité de tirer l'UP_h et $\pi_{i/h}$ la probabilité de tirer l'unité secondaire i sachant que l'UP_h à laquelle elle appartient a été tirée au premier degré.
D'après les formules du sondage à probabilités inégales, on a :

$t_h = \sum_{i=1}^{n_h}\frac{y_{hi}}{\pi_{i/h}}$ est un estimateur sans biais du total T_h et $t = \sum_{i=1}^{n_h}\frac{t_h}{\pi_h}$ est un estimateur sans biais du total T.

- Si N est connu alors $\bar{y} = \frac{t}{N}$ est un estimateur sans biais de la moyenne générale $\bar{Y}$.
- Si N est inconnu, cet estimateur est biaisé. On estime N par $\hat{N} = M.\hat{\bar{N}}$ avec $\hat{\bar{N}} = \frac{1}{m}\sum_{h=1}^{M} N_h$ nombre moyen d'US par UP.

Les formules des estimateurs des variances de ces estimateurs étant compliquées, nous donnerons leurs expressions dans le cas particulier qui seront exposés dans la section suivante.

III. Modalités pratiques du tirage d'un échantillon et calcul des estimateurs

En pratique, il existe deux méthodes pour tirer les unités primaires et secondaires. La première consiste à tirer les unités primaires avec des probabilités égales ; la seconde, avec des probabilités inégales proportionnelles à leur taille en nombre d'unités secondaires. Dans les deux cas, les unités secondaires sont tirées avec des probabilités égales et sans remise.

1. Tirage des UP à probabilités égales sans remise : estimation du total et de sa précision

Ce mode de tirage consiste à appliquer le tirage systématique au tirage des UP et des US dans chaque UP tirée. Il accorde en définitive la même probabilité à chacun des individus de la population. Ici, $\pi_h = \frac{m}{M}$ et $\pi_{i/h} = \frac{n_h}{N_h}$.

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ est un estimateur sans biais de $\bar{Y}_h$ soit $t_h = N_h \bar{y}_h$ est un estimateur sans biais de T_h.

On en déduit que $t = M\sum_{h=1}^{m}\frac{t_h}{m} = \sum_{h=1}^{m}\sum_{i=1}^{n_h}\frac{M}{m}\frac{N_h}{n_h} y_{hi}$ est un estimateur sans biais du total T

$\bar{y}=\frac{t}{N}$ Est un estimateur sans biais de la moyenne générale $\bar{Y}$.

Remarques : Contrairement au sondage aléatoire simple, on n'a pas besoin de connaître N pour estimer le total.

Cas particulier :

Si les taux de sondage au second degré sont constants $\frac{n_1}{N_1}=f_1$ et que toutes les UP ont même taille ($N_1=\bar{N}$), alors $t=M\bar{N}\bar{y}=N\bar{y}$.

L'échantillon se dépouille comme un recensement : la moyenne échantillon est l'estimateur sans biais de la moyenne globale.

L'estimateur de la variance s'écrit : $\hat{V}(t)=M^2(1-f_1)\frac{s_1^2}{m}+\frac{M}{m}\sum_{h=1}^{M}N_h(1-f_{2h})\frac{s_{2h}^2}{n_h}=\frac{A}{m}+\frac{B}{m\bar{N}}$. Il indique que :

- La qualité de l'estimateur dépend :
  - Du nombre d'unités primaires échantillon : m ;
  - Du nombre moyen d'unités secondaires : $N=m\bar{N}$:
  - Des modalités du tirage au premier et au deuxième degré.
- L'estimateur sera plus précis lorsque les s^2_1 et s^2_{2h} sont faibles. Pour cela, on doit constituer les UP de telle sorte qu'elles aient des totaux proches. Comme un total est un produit d'effectif et de moyenne ($T_h=N_h\bar{y}_h$), il faut alors choisir les UP de petites tailles et de moyennes semblables ; cela s'obtient en constituant des UP les plus hétérogènes possibles. Toutefois, dans la pratique, on est souvent amené à se contenter des îlots, des DR ou ZD, des villages comme unités primaires.

  Les modalités de tirage étant fixées, la variance apparaît comme une fonction de m et $\bar{N}$ de la forme : $\hat{V}(t)=\frac{A}{m}+\frac{B}{m\bar{N}}$ où

  A est une mesure de la variabilité entre UP, et
  B une mesure de la variabilité moyenne entre US d'une même UP.
- Une augmentation du nombre d'unité primaire à un impact plus prononcé sur la précision que ne l'est l'augmentation du nombre d'unités secondaires.

2. Tirage des UP à probabilités inégales avec remise

Les Up sont tirées avec des probabilités proportionnelles à leurs tailles en US. Au second degré, on tire sans remise dans chaque UP tiré au premier degré, un même nombre n_0 d'unités secondaires.

Dans ces conditions, $\pi_h=m\frac{N_h}{N}$ et $\pi_{i/h}=\frac{n_h}{N_h}$.

On établit que $t=\sum_{h=1}^{m}\frac{N}{m}\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi}$ est estimateur sans biais du total général T et que $\bar{y}=\frac{1}{m}\sum_{h=1}^{M}\bar{y}_h$ est un estimateur sans biais de $\bar{Y}$.

$$\hat{V}(\bar{y})=\frac{1}{m(m-1)}\sum_{h=1}^{m}(\bar{y}_h-\bar{y})^2$$

Cas particuliers

En pratique, on prend presque toujours le même nombre d'unités secondaires par UP: $\forall$ h, $n_h=n_0$. Dans ce cas, $n=mn_0$ est la taille de l'échantillon et l'estimateur du total devient $t=\sum_{h=1}^{m}\frac{N}{n}\sum_{i=1}^{n_0}y_{hi}$.

Par conséquent l'échantillon se dépouille comme un recensement, la moyenne est estimée directement par la moyenne empirique dans l'échantillon.

IV. Sondages par grappes

Le sondage par grappes est un sondage à plusieurs degrés où le dernier degré est un recensement.

Par exemple après avoir tiré m UP (appelées ici grappes) on enquête toutes les unités de ces grappes. Ici, la taille de l'échantillon est une variable aléatoire de moyenne $E(n)=m\overline{N}$, $\overline{N}$ étant l'effectif moyen par grappe.

L'avantage du sondage par grappes par rapport au sondage à plusieurs degrés est qu'il réduit considérablement les frais de déplacement. En revanche, son inconvénient majeur est qu'il souffre de l'effet de grappes : les individus appartenant à une même grappe ont tendance à se ressembler.

Comme dans le sondage à plusieurs degrés, le tirage des grappes peut se faire à probabilités égales ou à probabilités inégales.

1. Tirage des grappes à probabilités égales sans remise

Estimateur du total : $t=\frac{M}{m}\sum_{h=1}^{m}\sum_{i=1}^{n_h}y_{hi}$

Estimateur de la moyenne : $\bar{y}=\frac{1}{\overline{N}m}\sum_{h=1}^{m}\sum_{i=1}^{n_h}y_{hi}$, $\overline{N}=\frac{N}{M}$ est la taille moyenne des grappes dans la population.

Cas particulier :

Où les grappes sont de tailles égales ($\forall$ h, $N_h=\overline{N}$). Dans ce cas, l'échantillon est de taille fixe, et l'estimateur de la moyenne devient $\bar{y}=\frac{1}{n}\sum_{h=1}^{m}\sum_{i=1}^{n_h}y_{hi}$: l'échantillon se dépouille alors comme un recensement.

2. Tirage des grappes à probabilités inégales

Les probabilités de tirage des grappes sont proportionnelles à leurs tailles. On suppose que les tailles des grappes sont connues.

Estimateur du total : $t=\frac{N}{m}\sum_{h=1}^{m}\frac{1}{N_h}\sum_{i=1}^{N_h}y_{hi}=\frac{N}{m}\sum_{h=1}^{m}\bar{y}_h$

Estimateur de la moyenne : $\bar{y}=\frac{1}{N}\sum_{h=1}^{m}\sum_{i=1}^{N_h}y_{hi}$

Remarques : Si les tailles des grappes sont inconnues, alors il faut chercher à estimer le total N.

3. Effet de grappe

A l'intérieur d'une même unité primaire, les individus ont souvent tendance à se ressembler. Par exemple, les ménages d'un même îlot ont sensiblement le même niveau de vie ; les

exploitations agricoles d'une même région ont tendance à se ressembler. Ainsi, le sondage à deux degrés induit souvent une perte de précision par rapport au SAS. Cette perte de précision est appréhendée par :

$$DEFF = 1+\delta(\bar{n}-1) = 1+ROH(\bar{n}-1)^{1} \text{ avec } \delta = \frac{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N_\alpha}\sum_{\gamma=1}^{N_\alpha}(Y_{\alpha\phi}-\bar{Y})(Y_{\alpha\gamma}-\bar{Y})}{\sum_{\alpha=1}^{M}\sum_{\beta=1}^{N_\alpha}(Y_{\alpha\beta}-\bar{Y})^2} * \frac{1}{\bar{N}-1}^{2}$$

puisque l'on montre que : $V(\hat{t}_y) = (1+\delta(\bar{n}-1))V_{SAS}(\hat{t}_y)$. Ainsi, un sondage par grappe est plus précis que le SAS si DEFF>1 sinon il l'est moins. Or le fait que DEFF soit supérieur à 1 ou pas dépend du signe de δ.

En somme, le sondage par grappe (ou à plusieurs degrés) est plus précis que le SAS que si les individus d'une même grappe (Unité primaire) ont tendance à ne pas se ressembler. Si non, il y a une perte de précision dont il faut tenir compte.

Conclusion

Le sondage à plusieurs degrés notamment celui à deux degrés est une méthode de sondage qui présente des avantages certains en matière de réduction de coût d'enquête et de loin moins exigeante, en matière d'information, que les plans aléatoires présentés jusqu'ici. Ainsi, en dépit de la perte de précision qui lui est rattaché, il est très utile dans plusieurs situations. Ainsi, afin de surmonter son principal défaut plusieurs procédés ont été conçus.

- Stratification des unités primaires
- Sous stratification des US

[1] DEFF : Design Effect pour designer l'effet de grappe.

[2] Ce coefficient peut s'interprêter comme un coefficient de corrélation qui est positif lorsque les individus d'une grappe donnée ont tendance à se ressembler et negatif si la tendance est contraire. Il est appelé ROH dans certains manuels. En ce coefficient est compris entre 0 et 0.2.

REFERENCES BIBLIOGRAPHIQUES

Ardilly P., 1994, *les techniques de sondage*, TECHNIP
Tille Y., 1998, *Théorie des sondages*, ENSAI
Gourieroux C., 1981, *Théorie des sondages*, ECONOMICA
Grosbras J. M., 1986, *Méthodes statistiques des sondages*, ECONOMICA
Droesbeke J. J., Fichet B. et Tassi P.,1987, *les sondages*, ECONOMICA
Grais B., 1996, *les Méthodes statistiques*, DUNOD
Dussaix A. M. et Grosbras J. M.,1992, *Exercices de sondages*, ECONOMICA
Clairin R. et Brion P., 1997, *Manuel de sondage : applications aux pays en développement*, Manuel du Ceped
Statéco N° 95-96-97/ 2000, 2000, *Méthodes statistiques et économiques pour le développement et la transition*, INSEE
Statéco N° 78, Juin 1994, *Enquête 1 – 2 – 3 sur l'emploi et le secteur informel à Yaoundé*, INSEE