MINISTÈRE
DE L'EUROPE
ET DES AFFAIRES
ÉTRANGÈRES
Liberté
Égalité

EXPERTISE
FRANCE

AFRISTAT

Programme de renforcement des systèmes statistiques nationaux des pays du G5 Sahel dans le cadre de la redevabilité de l'Alliance Sahel

(Programme statistique G5-Sahel)

THEORIE
ET PRATIQUE DES SONDAGES

SOMMAIRE

Chapitre 1 : Considérations pratiques dans l'élaboration d'un plan de sondage

1. Introduction

Le sondage statistique est une technique qui consiste à enquêter sur un phénomène auprès d'un échantillon d'individus sélectionné selon certaines règles scientifiques pour représenter toute la population dont ils sont issus. L'exemple le plus connu est le sondage d'opinion développé aux Etats-Unis dès le début du 19 siècle.

On distingue le sondage probabiliste et le sondage non probabiliste. Dans le premier cas, la probabilité de sélection de chaque individu est connue d'avance. Dans le second cas, on ne peut pas calculer la probabilité de sélection ou d'inclusion.

Par ailleurs, la décision d'un statisticien ou d'un chercheur de réaliser une enquête par sondage doit sérieusement être pensée. Cette décision doit tenir compte non seulement des objectifs de l'étude mais aussi d'autres facteurs tels que les contraintes budgétaires, la qualité de la base de sondage, etc.

L'objet de ce chapitre est consacré aux différentes considérations à prendre en compte dans le choix de réaliser une enquête par sondage aléatoire.

2. Contexte et justification d'une enquête

Le contexte doit permettre d'appréhender la situation d'un pays en termes d'études et de production de données statistiques pour des prises de décisions dans le secteur cible. Il doit en outre faire ressortir le besoin à combler en réalisant une collecte de données.

Cette analyse situationnelle devra finalement déboucher sur la justification de réaliser une enquête. par rapport au besoin d'indicateurs recherchés.

3. Formulation des objectifs

L'élaboration de l'énoncé des objectifs est un processus itératif qui engage les producteurs et les utilisateurs des données statistiques. Les étapes du processus :

- les besoins d'information ;
- les utilisateurs et les utilisations des données ;
- les principaux concepts et les définitions opérationnelles ;
- le champ de l'enquête ;
- le plan d'analyse.

Exemple : on a besoin suivre le pouvoir d'achat de la population en calculant la dépense par tête et sa répartition selon les fonctions de consommation. La dépense par tête est obtenue en divisant la dépense totale de la consommation par l'effectif total de la population. L'enquête doit alors mesurer la composition des ménages, les conditions de logement et leurs dépenses de consommation

On distinguera ainsi l'objectif général de l'enquête qui consiste à suivre le pouvoir d'achat de la population et les objectifs spécifiques qui consistent à mesurer :

- Les dépenses monétaires de consommation des ménages ;
- L'autoconsommation ;
- Les transferts reçus en nature ;

- les caractéristiques socioéconomiques et démographiques de la population.

4. Concepts et indicateurs

Les principaux concepts de l'enquête doivent être bien définis :

- Le ménage ;
- La dépense monétaire ;
- L'autoconsommation ;
- Les sources de revenu des ménages ;
- Le cadre de vie ;
- La situation d'emploi de la population.

Les concepts sont généralement abstraits. Pour les opérationnaliser, il est requis la définition :

- des indicateurs de suivi ;
- des variables qui contribueront au calcul des indicateurs ou à l'explication du phénomène qui est suivi.

L'inventaire des indicateurs à calculer et des données à collecter est indispensable pour l'élaboration des questionnaires et des outils de collecte. Le tableau suivant illustre de l'opérationnalisation d'un concept par la mesure d'un indicateur.

Tableau 1 : opérationnalisation d'un concept par la mesure d'un indicateur

Concept	Indicateur	Variables requises
Dépense de consommation par tête : dépense effectuée en moyenne par un citoyen pour satisfaire ses besoins de consommation finale	Dépense de consommation par tête : rapport entre la dépense totale de consommation et l'effectif total de la population	Nom des biens et services consommés
		Dépense de consommation calculée pour l'ensemble des ménages
		Taille des ménages ou unités de consommation

5. Champs géographique et social de l'enquête

Le champ social d'une enquête est l'ensemble des populations cibles de l'enquête. Dans l'exemple précédent, la population cible peut être l'ensemble des ménages toutes catégories confondues. La définition du champ géographique impacte le coût des enquêtes. Il y a lieu de savoir si les résultats attendus seront significatifs au niveau : national ou infranational

La définition du champ géographique et du champ social de l'enquête induit des questions suivantes pour la méthodologie de collecte :

- Qui enquêter ?
- Où enquêter ?
- Comment enquêter ?
- Quoi enquêter ?

6. Etapes de conception d'une enquête

Une enquête comprend plusieurs étapes :

- Contexte et justification
- Définition des objectifs
- Concepts et indicateurs
- Champ de l'enquête
- Plan de sondage
- Conception du questionnaire
- Collecte de données
- Traitement des données
- Analyse et la diffusion des données
- Documentation de l'enquête

7. Détermination de la taille d'échantillon

La plus grande question à laquelle doit répondre un statisticien d'enquête dès le départ est la détermination de la taille de l'échantillon des unités statistiques à enquêter. Dans la détermination de la taille de l'échantillon, il faut tenir compte de trois préoccupations : i) la précision des estimations, ii) les contraintes de mise en œuvre du plan de sondage, iii) l'efficacité du plan de sondage.

Précision des estimations

Le recours à une base de données d'enquête antérieure ayant déjà permis de calculer les indicateurs recherchés est très important pour le statisticien. Il permet de tirer les enseignements pour la réalisation d'une nouvelle enquête. Les enseignements peuvent porter sur :

- la taille de l'échantillon ;
- le type de plan de sondage ;
- les outils utilisés pour la collecte des données ;
- les difficultés rencontrées aussi bien pendant la collecte des données qu'à la phase du traitement et d'analyse des résultats (taux de réponse, erreurs d'observation, …) ;
- les estimations des principaux indicateurs et leur précision.

Toutes ces informations doivent aider le statisticien à effectuer des simulations en vue de la détermination de la taille de l'échantillon requise pour la réalisation de l'enquête en vue. Par ailleurs, la taille de l'échantillon doit être déterminée en fonction de : i) la fréquence d'apparition des évènements du phénomène étudié et ii) du niveau des strates d'analyses.

Toutefois, certaines formules peuvent être utilisées pour le calcul de la taille d'échantillon d'une enquête. Ces formules sont présentées ci-après. Dans le cas d'un sondage aléatoire simple (SAS), la formule suivante est souvent utilisée pour déterminer la taille de l'échantillon en vue d'obtenir un degré de précision d'une moyenne :

$$n = \frac{z^2 \hat{S}^2}{e^2 + \frac{z^2 \hat{S}^2}{N}}$$

- Z : fractile d'ordre 1-α/2 de la loi normale centrée réduite ; habituellement, on considère α=0,05 ; dans ce cas Z=1,96 ou 2
- S est l'estimateur de la variabilité de la variable d'intérêt dans un sondage aléatoire simple ;
- e : marge d'erreur à considérer dans l'estimation de l'indicateur ou de la variable d'intérêt ; habituellement on considère e=0,05 ;
- N : taille de la population
- Dans le cas d'une proportion P, l'équation s'écrit :
- $n = \frac{z^2 \hat{P}(1-\widehat{P})}{e^2 + \frac{z^2 \hat{P}(1-\widehat{P})}{N}}$

Dans le cas où N est suffisamment grand, l'expression devient

$$n = \frac{z^2 \hat{P}(1-\widehat{P})}{e^2}$$

Exemple : si z=1,96; P=0,5 et e=0,05 alors :

$$n = \frac{(1{,}96)^2 * 0{,}5(1-0{,}5)}{(0{,}05)^2}$$

Soit n=384

Tableau 2 : Exemples des tailles d'échantillon pour une proportion P=0,5 (SAS) avec un taux de confiance de 95%

Taille de l'échantillon	Marge d'erreur
50	± 0,139
100	± 0,098
500	± 0,044
1000	± 0,031

Source : Théories et pratiques de sondage, Statcan

Dans le cas où N n'est pas suffisamment grand, on peut ajuster la taille de l'échantillon initial (notée n1) par l'expression :

$$n_2 = n_1 \frac{N}{N + n_1}$$

Exemple : si N=8000 et n1=384 alors :

$$n_2 = 384 \frac{8000}{8000 + 384}$$

Soit n2=366

Dans le cas d'un sondage aléatoire à plusieurs degrés, on ajuste la taille de l'échantillon par l'effet de plan de sondage (Deff) à l'aide de l'expression :

$$n_3 = deff * n_2$$

Exemple : si deff=1,2 et n2=366 alors :

$$n_3 = 366 * 1{,}2$$

Soit n3=439

En faisant l'hypothèse sur le taux de réponse noté r à l'enquête, alors on peut ajuster la taille d'échantillon par la formule :

$$n_4 = \frac{n_3}{r}$$

Exemple : si r=0,8 et n3=439 alors :

$$n_4 = \frac{439}{0{,}8}$$

Soit n4=549

Dans le cas d'un sondage aléatoire simple stratifié :

- Option 1 : marge d'erreur pour l'estimation de la population dans l'ensemble. On calcule n et on procède à sa répartition dans les strates (allocation proportionnelle, allocation optimale de Neyman, etc…)
- Option 2 : marge d'erreur pour l'estimation de la population dans chaque strate
- Le cumul des tailles des échantillons des strates donne l'échantillon total

Contraintes de mise en œuvre du plan de sondage

Plusieurs contraintes peuvent avoir des incidences sur la détermination de la taille de l'échantillon à enquêter. La plus importante est la contrainte budgétaire. Le budget d'une enquête doit être réparti en coûts fixes et coûts variables.

Les coûts fixes sont généralement incompressibles indépendamment de la taille de l'échantillon à enquêter (coordination, fonctionnement, organisation des ateliers de formation des agents, etc.). Les coûts variables sont étroitement liés à la taille de l'échantillon (nombre de questionnaires, déplacement et salaire des agents de terrain et d'exploitation des données, etc.)

Efficacité du plan de sondage

Le plan de sondage efficace est celui qui offre la plus grande précision pour une taille d'échantillon prenant en compte les contraintes de mise en œuvre.

Par exemple, si la population concernée par un phénomène à étudier est de petite taille, il est préférable de réaliser une enquête par sondage aléatoire simple, à condition qu'il y ait une bonne maitrise des coûts de déplacement.

Le statisticien peut aussi décider de réaliser un sondage aléatoire stratifié pour améliorer la précision des résultats, à condition de bien choisir les critères de stratification et de ne pas créer plusieurs strates qui n'apportent pas forcément de gain d'amélioration des estimations.

8. Base de sondage

Une autre étape à franchir est le choix de la base de sondage de bonne qualité : i) une couverture exhaustive, ii) pas de doubles enregistrements, iii) pas des unités « mortes », iv) non basée essentiellement sur des informations anciennes. On distingue :

- Une base de sondage aréolaire constituée des images satellitaires et de la télédétection ;
- Une base de sondage liste.

Le terme base de sondage désigne généralement une liste concrète d'unités ayant un lien avec la population à étudier. Il s'agit de la description d'éléments déjà existants sous forme de listes, cartes, annuaires, à partir desquels on peut constituer des unités et sélectionner un ensemble d'unités à enquêter.

La base de sondage doit comprendre toutes les informations auxiliaires (mesure de taille, données démographiques, …) nécessaires pour la mise en œuvre de techniques spéciales de sondage telles que la stratification ou les types de tirage retenus.

La construction d'une base de sondage doit tenir compte de :

- la nature de la population à étudier : individus, familles, ménages, exploitations agricoles ;
- la répartition géographique de la population cible : est-elle limitée à une localité, une région ou est-elle répartie sur l'ensemble du pays ;
- la nature des opérations de terrain: enquête par téléphone, par correspondance, ou interview directe par des enquêteurs.

Une base de sondage peut convenir pour une enquête donnée et ne pas être adaptée pour un autre type d'enquête. Les unités statistiques dans la base doivent être de taille assez homogène. Le nombre d'unités dans la base doit être connue pour permettre les extrapolations.

Exemples de bases de sondage utilisables pour les enquêtes auprès des ménages : i) listes de concession, ii) zones de dénombrement du recensement de la population.

Zones de dénombrement du recensement de la population (ZD) :

- Découpage du territoire national en petites aires géographiques pour des besoins de recensement de la population.
- Les limites des ZD sont bien connues ;
- La mise à jour de la base de données sur les ZD nécessite des travaux de terrain coûteux ;
- D'où l'intérêt de constituer plus tard un échantillon maitre de ZD, destiné à fournir des sous-échantillons devant servir aux besoins d'enquêtes diverses.

9. Principaux paramètres à estimer par une enquête

Une enquête par sondage peut permettre d'estimer les caractéristiques suivantes d'une variable d'intérêt Y :

- Le total ;
- La moyenne (la proportion est une moyenne calculée pour une variable qualitative) ;
- Le ratio ;
- La variance et le coefficient de variation des estimateurs ;
- L'effet de sondage.

10. Différents types d'erreurs rencontrés dans une enquête par sondage

Erreur d'échantillonnage : elle est due au plan de sondage ; elle est mesurable par le biais, la variance ou l'écart quadratique moyen. L'estimateur de l'erreur d'échantillonnage suit une loi de probabilité.

Erreur d'observation ou erreur de mesure : c'est une erreur liée au dispositif et au support de collecte de données. Elle survient surtout dans la formulation des questions ou lors de la manipulation des instruments de mesure. Elle est très difficile à quantifier, à moins de retourner sur le terrain.

Erreur due au défaut de couverture ou à la non-réponse : elle est liée d'une part à l'utilisation d'une base de sondage incomplète et d'autre part à la non-réponse complète ou partielle à l'enquête par certains individus. Les non-réponses peuvent être redressées ou imputées. Par contre, il est difficile de corriger à posteriori l'erreur due au défaut de couverture.

11. Choix et justification du type de plan de sondage

Les différents types de sondage aléatoires sont :

- le sondage aléatoire simple ;
- le sondage aléatoire à probabilités inégales ;
- le sondage stratifié ;
- le Sondage en grappes ;
- le sondage à plusieurs degrés.

Le choix d'un plan de sondage doit être justifié. Par exemple, lorsqu'il s'agit de réaliser l'étude d'un phénomène sur une population de petite taille et bien cernée, le statisticien peut décider de réaliser un sondage aléatoire simple ou avec stratification (si les variables de stratification existent).

Par contre, la réalisation d'une étude qui concerne une grande population utilisera soit à un plan de sondage soit par grappe ou à plusieurs degrés avec des possibilités de stratification.

Attention : un sondage à deux degrés est plus précis qu'un sondage à trois degrés ou plus.

Exemple

1) Quels sont les critères qui définissent une base de sondage de bonne qualité ?
2) Pourquoi fait-on souvent recours au plan de sondage à deux degrés dans la plupart des enquêtes statistiques en Afrique ?
3) Décrire la méthode d'un tirage systématique. Quelles sont ces limites ?
4) Qu'est-ce qui peut justifier la mise en œuvre d'un sondage aléatoire simple ?
5) Pourquoi a-t-on besoin de constituer un échantillon maître ?

Chapitre 2 : Sondage aléatoire simple

1. Introduction

Sondage aléatoire simple (SAS) : une méthode qui consiste à tirer au hasard avec ou sans remise dans une population de taille N, un échantillon de n individus et ceci en donnant la même chance de sélection à chaque unité. Exemple : tirage de n boules dans une urne qui en contient N.

2. Notations

Notations sur l'univers		**Notations sur l'échantillon**	
N	Population totale	n	Taille de l'échantillon
Y_i	Valeur de la caractéristique Y pour le ième individu de la population	y_i	Valeur de la caractéristique Y pour le ième individu de l'échantillon
$Y = \sum_{i=1}^{N} Y_i$	Total d'une caractéristique dans une population	$y = \sum_{i=1}^{n} y_i$	Total d'une caractéristique sur l'échantillon
$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$	Moyenne d'une caractéristique dans la population	$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	Moyenne d'une caractéristique sur l'échantillon
σ_Y^2	Variance de la caractéristique Y dans la population	σ_Y^2	Variance de la caractéristique Y dans l'échantillon
$S_Y^2 = \frac{N}{N-1}\sigma_Y^2$	Variance corrigée de la caractéristique Y dans la population	$s_Y^2 = \frac{n}{n-1}\sigma_Y^2$	Variance corrigée de la caractéristique Y dans l'échantillon
$\frac{1}{f} = \frac{N}{n}$	Facteur d'extrapolation	$f = \frac{n}{N}$	Taux de sondage
CV(Y)	Coefficient de variation de Y dans la population	cv(Y)	Coefficient de variation de Y dans l'échantillon

3. Probabilité d'inclusion

Dans un sondage aléatoire simple sans remise (SASSR), la probabilité d'inclusion ou de sélection de n unités dans une population finie de taille N est : $P = \frac{n}{N}$.

4. Estimateur du total, estimateur de la moyenne

Soit Y_i la valeur de la caractéristique Y dans la population. Soient N l'effectif de la population et n la taille de l'échantillon. Alors on a $Y = \sum_{i=1}^{N} Y_i$. La moyenne dans la population est $\bar{Y} = \frac{Y}{N}$. Mais la vraie valeur de Y est inconnue. L'estimateur sans biais $\hat{Y}$ de Y s'écrit : $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

On note que dans un SASSR, la probabilité d'inclusion d'un individu i dans l'échantillon est P_i= n/N. D'où : $\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{P_i}$ ou encore $\hat{Y} = N\bar{y}$

L'estimateur sans biais de la moyenne de Y est par définition égale à : $\hat{\bar{Y}} = \bar{y} = \frac{\hat{Y}}{N}$. On note que f = n/N est aussi appelé le taux de sondage.

5. Estimateur de la variance des estimateurs du total et de la moyenne

Par définition, la variance corrigée ou la dispersion de la caractéristique Y dans la population s'écrit : $S_Y^2 = \frac{N}{N-1}\sigma_Y^2$ ou encore $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$

Le développement de la formule donne : $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N} Y_i^2 - \frac{N}{N-1}\bar{Y}^2$

Pour N grand, on obtient la formule simplifiée : $S_Y^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \bar{Y}^2$ qui est l'expression de la variance.

On montre que la variance de l'estimateur de la moyenne s'écrit : $V(\bar{y}) = (1-f)\frac{S_Y^2}{n}$ et celle de l'estimateur du total s'écrit : $V(\hat{Y}) = (1-f)\frac{N^2 S_Y^2}{n}$. L'estimateur de S_Y^2 est s_Y^2. D'où : $\hat{V}(\bar{y}) = (1-f)\frac{s_Y^2}{n}$ et $\hat{V}(\hat{Y}) = (1-f)\frac{N^2 s_Y^2}{n}$ qui désignent respectivement les estimateurs des variances des estimateurs de la moyenne et du total.

Propriété : pour réaliser un sondage aléatoire précis, on peut :

- Augmenter la taille n de l'échantillon
- Augmenter le taux de sondage
- Diminuer la variance corrigée

En pratique, on agit surtout sur la taille n de l'échantillon.

6. Estimation des intervalles de confiance

Soit θ un paramètre à estimer et $\hat{\theta}$ son estimateur sans biais. Alors l'intervalle de confiance de θ s'écrit : $IC = \left[\hat{\theta} \pm 2\sqrt{\hat{V}(\hat{\theta})}\right]$, avec 5% de risque de se tromper.

Si θ est une moyenne de la variable Y, alors on a : $IC = \left[\bar{y} \pm 2\sqrt{\frac{(1-f)s_Y^2}{n}}\right]$

Si θ est un total de la variable Y, alors on a : $IC = \left[\hat{Y} \pm 2N\sqrt{\frac{(1-f)s_Y^2}{n}}\right]$

7. Estimation des intervalles de confiance d'une proportion

Si θ est une proportion p, alors la moyenne estimée vaut p et l'estimateur de la variance de l'estimateur de θ vaut : $\hat{V}(\hat{\theta}) = \frac{p(1-p)(1-f)}{n}$

D'où l'intervalle de confiance $IC = \left[p \pm 2\sqrt{\frac{p(1-p)(1-f)}{n}}\right]$

8. Cas d'un sondage aléatoire simple à probabilités égales avec remise (SASR)

Dans ce cas tous les tirages sont effectués dans les mêmes conditions et les Yi avec i= 1 à n sont des variables aléatoires indépendantes. La probabilité d'inclusion est : $P_i = 1 - \left(1 - \frac{1}{N}\right)^n$, quel que soit i.

Quand N est suffisamment grand, on peut écrire : $P_i \approx \frac{n}{N}$.

L'estimateur sans biais de la moyenne $\overline{Y}$ de Y s'écrit : $\hat{\overline{Y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$. La variance de l'estimateur de la moyenne s'écrit : $V(\hat{\overline{Y}}) = \frac{\sigma^2}{n}$ et l'estimateur de la variance a pour expression : $\hat{V}(\hat{\overline{Y}}) = \frac{s^2}{n}$.

On peut aussi définir l'intervalle de confiance de la moyenne de Y avec un niveau de confiance 1-α :

$IC = \left[\hat{\overline{Y}} \pm u_{\alpha/2}\frac{s}{\sqrt{n}}\right]$ avec $u_{\alpha/2}$,le fractile d'ordre α/2 de la loi normale et $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{\overline{Y}}\right)^2$.

9. Détermination de la taille d'échantillon

Dans le cas d'un sondage aléatoire, la première question à résoudre est la taille de l'échantillon. Une taille importante améliore la précision des résultats mais la question du coût de l'enquête peut se poser.

Forte contrainte de coût : soit T le budget total de l'enquête et c le coût unitaire de remplissage d'un questionnaire. Alors l'expression de la taille n s'écrit n = T/c.

Faible contrainte de coût : dans ce cas, la condition importante est l'amélioration de la précision des résultats. Si on veut estimer la moyenne de Y à L près, alors on écrit $L = 2\sqrt{\hat{V}(\hat{\bar{Y}})}$ avec un niveau de confiance de 95%.

Le développement conduit à $n = \frac{N}{1+\left(\frac{NL^2}{4s^2}\right)}$. L'élément inconnu ici est S^2. On peut l'estimer, soit à partir d'une variable corrélée à la variable d'étude, déterminée dans une autre enquête ou d'une enquête traitant directement de la variable cible.

10. Tirage d'un échantillon aléatoire simple

Deux méthodes de tirage :

Tirage systématique

Il est très pratiqué. Les étapes sont :

- Disposer d'une liste finie d'unités de sondage de taille N finie
- Numéroter les unités de sondage de 1 à N
- Calculer un pas de tirage noté P = N/n, où n est la taille de l'échantillon à tirer
- Tirer un nombre aléatoire compris entre 0 et 1
- Multiplier ce nombre aléatoire par le pas P. La valeur obtenue est le numéro de la première unité tirée
- Ajouter à la valeur tirée le pas P calculé pour obtenir le numéro de l'unité suivante à sélectionner
- Poursuivre le processus jusqu'à l'obtention des n unités de l'échantillon.

Précautions : soit P le pas de tirage calculé. Soit α un nombre aléatoire tiré entre 0 et 1. Alors P et α sont des nombres réels. Soit $X_1=P\alpha$. Alors X_1 désigne le numéro de la première unité sélectionnée.

$X_2=X_1+P$, $X_i=X_{i-1}+P$

Soit X_i, tel que $u \leq X_i<u+1$ où u est un entier naturel. On considère alors que l'unité sélectionnée correspond à l'entier immédiatement supérieur, soit u+1.

Une autre façon de procéder à un tirage systématique de n individus sur un total N, consiste à :

- Regrouper les N individus en n groupes
- Tirer de façon systématique un individu dans chaque groupe

Avantages : facile à mettre en œuvre, forts gains de représentativité dès que l'on connaît une variable auxiliaire corrélée avec la variable d'intérêt et que l'on classe les unités en fonction de cette variable, quand le fichier a une tendance linéaire

Inconvénients : une catastrophe si la variable étudiée présenterait des variations périodiques dont la périodicité serait voisine d'un sous multiple du pas de tirage.

Tirage de nombres aléatoires

Il y a plusieurs variantes de la méthode de tirage des nombres aléatoires. Le recours à l'informatique notamment des tableurs, facilite aujourd'hui le choix entre les variantes. Lorsqu'on veut tirer un échantillon de taille n dans une population de N unités, l'une des variantes consiste à :

- Trier d'abord la liste des unités de l'univers dans un sens donné et les numéroter de 1 à N.
- Ranger les numéros dans une seule colonne.
- Générer un nombre aléatoire entre 0 et 1 devant chaque unité.
- Trier les unités dans l'ordre décroissant des nombres aléatoires
- Sélectionner les n premières unités pour constituer l'échantillon.

11. Exercices pratiques

Exercice 1 : soit dans le tableau ci-après, la distribution des notes à une épreuve sur la théorie de sondage dans une classe de 5 étudiants numérotés de 1 à 5.

Numéro de l'élève	1	2	3	4	5
Note reçue (sur 20)	12	15	10	7	8

1) Calculer la moyenne $\bar{Y}$ et la variance corrigée S^2 des notes reçues par les élèves
2) Former tous les échantillons possibles de 3 notes parmi les 5.
3) Calculer la moyenne $\hat{\bar{Y}}$ et la variance corrigée s^2 des notes pour chaque échantillon
4) Calculer la moyenne des moyennes des notes des différents échantillons. Qu'observe-t-on ?

Exercice 2 : un enseignant d'un cours de théorie et pratique de sondage dans une classe de 51 étudiants a corrigé toutes les copies d'un devoir qu'il a donné. Il a décidé d'évaluer le niveau de la classe à partir d'un échantillon de 17 copies.

1) quel est le nombre possible d'échantillons ?
2) quel plan de sondage peut-on lui conseiller ? pourquoi ?
3) quel mode de tirage peut-on lui conseiller? pourquoi ?

L'enseignant a procédé à un tirage aléatoire simple des 17 copies. Les caractéristiques de l'échantillon sont : moyenne des notes $\bar{y} = 14$ et l'écart-type $\sigma_y = 14$

4) calculer les grandeurs suivantes $\hat{V}(\bar{y})$ et $\hat{V}(\hat{Y})$ et les intervalles de confiance respectifs.

Chapitre 3 : Sondage stratifié

1. Introduction

Dans un sondage aléatoire simple, on tire directement l'échantillon des individus dans une population totale et unique. Il se pose généralement un problème de répartition spatiale de l'échantillon tiré pour assurer la représentativité de la population étudiée.

Dans un sondage stratifié, la population étudiée est divisée en plusieurs groupes indépendants les uns des autres. Le processus d'établissement des groupes s'appelle la stratification.

Les strates peuvent répondre à des caractéristiques économiques (revenu, catégorie socio professionnelle, branche d'activité), sociales (pauvres, non pauvres), géographiques (région, département, etc..).

Dans une stratification, on met ensemble des individus qui se ressemblent sur les critères de base (constitution de groupes homogènes d'individus). Ceci a pour avantage de réduire la dispersion des variables étudiées au sein de chaque strate. Par conséquent, l'objet d'une stratification est de réduire la variance totale. Toutefois, les critères de stratification sont satisfaisants si les estimateurs des moyennes et des totaux sont réellement différents d'une strate à l'autre.

2. Probabilité d'inclusion

Un sondage stratifié de base, est un sondage aléatoire simple réalisé indépendamment dans l'ensemble des strates qui forment l'univers. Ainsi, les tirages des échantillons sont indépendants d'une strate à une autre.

Soient N_h et n_h désignant respectivement la population totale et la taille de l'échantillon à tirer dans la strate h. Soit N la population totale de l'univers. La probabilité d'inclusion d'une unité dans l'échantillon d'une strate s'écrit : $P_h = \frac{n_h}{N_h}$.

On note que : $N = \sum_{h=1}^{k} N_h$

3. Estimateur du total, estimateur de la moyenne

Soit h = 1, 2, 3, …. k le nombre total de strates constituant l'univers. L'estimateur du total d'une variable Y sur l'univers est le total des estimateurs de Y sur les strates h. En d'autres termes,

$\hat{Y} = \sum_{h=1}^{k} \hat{Y}_h$ avec $\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{ih}$ ou encore $\hat{Y} = \sum_{h=1}^{k} N_h \hat{\bar{Y}}_h$

L'expression $\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{ih}$ est la moyenne de la variable Y sur la strate h. Mais la moyenne de la variable Y sur l'univers s'écrit : $\hat{\bar{Y}} = \frac{\hat{Y}}{N}$ ou encore $\hat{\bar{Y}} = \sum_{h=1}^{k} \frac{N_h \hat{\bar{Y}}_h}{\sum_{h=1}^{k} N_h}$

L'expression $\hat{\bar{Y}} = \frac{1}{N}\sum_{h=1}^{k}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{ih}$ .La grandeur $W_h = \frac{N_h}{Nn_h}$ est le poids de sondage d'un individu sélectionné dans la strate h.

4. Estimateur de la variance de l'estimateur du total

Par définition l'estimateur de la variance de l'estimateur du total Y s'écrit : $\hat{V}(\hat{Y}) = \hat{V}(\sum_{h=1}^{k} N_h \hat{\bar{Y}}_h)$. Le tirage des échantillons étant indépendant d'une strate à l'autre, cette expression devient : $\hat{V}(\hat{Y}) = \sum_{h=1}^{k} N_h^2 \hat{V}(\hat{\bar{Y}}_h)$.

Dans un sondage aléatoire simple, l'expression de $\hat{V}(\hat{\bar{Y}}_h)$ s'écrit : $\hat{V}(\hat{\bar{Y}}_h) = (1-f_h)\frac{s_{Yh}^2}{n_h}$

Où s_{Yh}^2 et f_h désignent respectivement l'estimateur de la variance corrigée et le taux de sondage dans la strate h. L'expression de l'estimateur de la variance de l'estimateur du total Y s'écrit alors : $\hat{V}(\hat{Y}) = \sum_{h=1}^{k}(1-f_h)\frac{N_h^2 s_{Yh}^2}{n_h}$

5. Estimateur de la variance de l'estimateur de la moyenne

L'estimateur de la variance de l'estimateur de la moyenne s'écrit : $\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{k}\left(\frac{N_h}{N}\right)^2 \hat{V}(\hat{\bar{Y}}_h)$ ou encore

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{k}\left(\frac{N_h}{N}\right)^2 (1-f_h)\frac{s_{Yh}^2}{n_h}$$

6. Détermination de la taille d'échantillon par strate

La question consiste à savoir comment répartir la taille (n) de l'échantillon global entre les différentes strates constituées. La répartition de l'échantillon entre les strates doit tenir compte de deux critères importants. Le premier critère est la commodité. Il consiste à choisir une méthode simple à appliquer telle que l'allocation proportionnelle. Le second critère est la fiabilité des estimations. D'où l'intérêt souvent porté à la méthode d'allocation optimale de Neyman.

6.1 Allocation proportionnelle

Cette méthode consiste à répartir la taille de l'échantillon selon la même structure de répartition de l'effectif de la population totale entre les strates.

Soient N et N_h désignant respectivement la population totale de l'univers et la population totale de la strate h. Soient n et n_h désignant respectivement la taille de l'échantillon global et la taille de l'échantillon affecté à la strate h. Alors on a la relation : $\frac{n_h}{n} = \frac{N_h}{N}$

L'estimateur de la moyenne de la variable Y s'écrit : $\hat{\bar{Y}} = \sum_{h=1}^{k}\frac{n_h \hat{\bar{Y}}_h}{n}$ ou encore $\hat{\bar{Y}} = \frac{1}{n}\sum_{h=1}^{k}\sum_{i=1}^{n_h} y_{ih}$

Ainsi un sondage stratifié à allocation proportionnelle est un sondage auto pondéré.

L'estimateur de la variance de l'estimateur de la moyenne s'écrit : $\hat{V}(\hat{\bar{Y}}) = \frac{1-f}{n}\sum_{h=1}^{k}\frac{N_h s_{Yh}^2}{N}$. On montre que $\sum_{h=1}^{k}\frac{N_h s_{Yh}^2}{N} \leq S_Y^2$. Ou encore $\hat{V}(\hat{\bar{Y}}) \leq \hat{V}_{SAS}(\hat{\bar{Y}})$

Le sondage stratifié à allocation proportionnelle améliore la variance de l'estimateur de la moyenne par rapport au sondage aléatoire simple.

6.2 Allocation optimale

L'objectif visé par cette méthode est la répartition de l'échantillon entre les différentes strates sous la contrainte des coûts de réalisation.

Soit c le coût total de réalisation d'une enquête sur un échantillon de taille n. Alors c_h désigne le coût de réalisation de l'enquête sur un échantillon de taille n_h dans une strate h. On a la relation : $c = \sum_{h=1}^{k} c_h n_h$

. La répartition optimale est donnée par la formule : $\frac{n_h}{n} = \frac{\frac{N_h s_h}{\sqrt{c_h}}}{\sum_{h=1}^{k}\frac{N_h s_h}{\sqrt{c_h}}}$

On cherche alors les valeurs de n_h qui minimisent l'estimateur de la variance obtenu dans le cas de l'allocation proportionnelle sous la contrainte de c.

On parle d'allocation optimale de Neyman lorsqu'on considère que les coûts des questionnaires sont identiques entre les différentes strates. La relation précédente devient : $\frac{n_h}{n} = \frac{N_h s_h}{\sum_{h=1}^{k} N_h s_h}$

On montre que l'estimateur de la variance de l'estimateur de la moyenne s'écrit : $\hat{V}(\hat{\bar{Y}}) = \frac{1}{n}\left(\sum_{h=1}^{k}\frac{N_h s_h}{N}\right)^2 - \frac{1}{N}\sum_{h=1}^{k}\frac{N_h s_h^2}{N}$

Soient V_{SAS}, V_{prop} et V_{opti} les estimations respectives de la variance de l'estimateur de la moyenne pour un sondage aléatoire simple, un sondage stratifié à allocation proportionnelle et un sondage stratifié à allocation optimale. Alors on a la relation suivante : $V_{opti} \leq V_{prop} \leq V_{SAS}$

Ainsi, lorsqu'on dispose des informations suffisantes sur la taille et la variance corrigée de la variable étudiée dans chaque strate, il est préférable de porter son choix sur le sondage stratifié à allocation optimale.

7. Problèmes de constitution des strates

Procéder à la stratification d'un univers c'est chercher à répondre à trois questions :

- Quelle est la variable de stratification ?
- Combien de strates faut-il créer ?
- Quelles sont les limites des strates ?

7.1 Variable de stratification

La variable de stratification doit être la plus discriminante possible pour permettre de constituer des strates autant homogènes en intra et hétérogènes en extra, vis-à-vis de la variable d'intérêt. Exemple : le milieu de résidence (urbain, rural), les régions administratives d'un pays, les zones agroclimatiques.

La variable de stratification ne doit pas être trop écartée de la variable d'intérêt. Par exemple, est-il pertinent de stratifier l'univers en zones agroclimatiques, lorsqu'on veut mesurer le nombre de clients d'un réseau de la téléphonie mobile dans un pays ?

7.2 Nombre de strates

En théorie, il est toujours souhaitable d'avoir un nombre important de strates. Cependant, plus il y a de strates, faible sera la taille d'échantillon par strate. Par conséquent, il y a risque de détérioration des variances intra-strates. Par conséquent, il peut y avoir également une détérioration de la variance totale.

En outre, plus l'échantillon d'une strate est petit plus le risque d'un taux de non réponse élevé devient grand dans la strate.

En pratique, le nombre de strates doit être choisi non seulement pour la précision des résultats mais aussi en fonction des coûts de gestion de l'enquête.

7.3 Limites des strates

Les limites de certaines strates sont parfois difficiles à appréhender. Dans certains pays par exemple, la notion de ville est définie par rapport à l'effectif de population de la localité, indépendamment de l'état des infrastructures. Du coup, on range dans la même catégorie de milieu dit urbain, des villes importantes qui jouent des rôles de pôles économiques avec des agglomérations jugées importantes en terme démographique mais qui ne disposent réellement pas d'infrastructures urbaines.

En pratique, s'agissant des strates basées sur le découpage administratif, l'on est obligé de s'y conformer.

8. Exposés des cas pratiques

Le G5 Sahel a réalisé une enquête par sondage sur la fécondité dans un pays appelé Gondouana comprenant deux provinces. Le chercheur du G5 Sahel a tiré un échantillon 30 ménages selon un plan de sondage stratifié aléatoire simple. Les deux provinces sont les strates. Le tableau suivant porte sur les caractéristiques des échantillons respectifs :

Strates	Nombre total de ménages N_h	Echantillon de ménages sélectionnés et enquêtés
A	1200	20
B	600	10

Le tableau suivant donne la répartition du nombre de naissances vivantes au cours des 12 derniers mois déclarés par les ménages enquêtés

Province	Numéro des ménages enquêtés																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	0	0	2	0	0	1	1	1	0	0	0	0	0	0	0	3	0	2	1	1
B	0	1	1	2	1	3	1	1	0	0										

1) Quel est le type de sondage proposé ?

1) Calculer le nombre de naissances vivantes dans l'échantillon de chaque strate

2) Estimer le nombre total de naissances vivantes dans chaque strate

3) Estimer le nombre total de naissances vivantes dans le Gondouana

4) Estimer la variance corrigée du nombre de naissances vivantes par strate S_{Yh}^2

5) Calculer l'estimateur de la variance du nombre moyen de naissances vivantes au Gondouana

6) Quels sont les risques d'augmenter indéfiniment le nombre de strates dans un sondage aléatoire ?

Chapitre 4 : Sondage à probabilités inégales

1. Introduction

Par défaut, le sondage aléatoire simple est un sondage dans lequel on accorde la même chance de sélection à toutes les unités. Il s'agit donc d'un sondage à probabilité égale ou encore appelé un sondage équiprobable. Il existe cependant, une forme de sondage dont le dispositif accorde plus de chances de sélection à certaines unités que d'autres.

Exemple 1 :

Un exemple concret est la répartition de la population dans une localité. Elle n'est jamais uniforme. Il existe des zones fortement peuplées et d'autres qui le sont moins. On peut donc choisir de tirer l'échantillon des zones avec probabilités inégales et effectuer directement l'enquête auprès de la population des zones sélectionnées.

Exemple 2 :

On veut réaliser l'étude de marché d'un produit dans les lieux de vente. La base de sondage donne l'information sur le nombre moyen de clients par lieu de vente dans un mois. Alors, on peut décider de tirer les lieux de vente proportionnellement au nombre de clients qui les fréquentent.

Dans un sondage à probabilités inégales, il y a une dissymétrie du tirage des unités. Les plus importantes en taille ont plus de chance d'être sélectionnées. Alors les formules des estimations doivent permettre de leur accorder moins d'importance par rapport aux unités qui ont moins de chance d'être sélectionnées. C'est une façon de chercher à rétablir l'équilibre.

2. Probabilité d'inclusion

Soit P_i la probabilité d'inclusion d'une unité i dans un échantillon. Cette probabilité doit être comprise entre 0 et 1. Pour un échantillon de taille fixe n à tirer dans une population finie de taille N, on a : $n = \sum_{i=1}^{N} P_i$. Dans le cas d'un tirage d'une unité i avec une probabilité proportionnelle à sa taille notée X_i, on écrit : $P_i = n\frac{X_i}{\sum_{i=1}^{N} X_i}$. **Cette relation est vraie si on a :** $\frac{1}{n} \geq \frac{X_i}{\sum_{i=1}^{N} X_i}$ **quelle que soit la valeur de i. Cette condition est très importante à vérifier lorsqu'on porte le choix sur un tirage à probabilité inégale.**

3. Estimateurs du total et de la moyenne

Soit Y une variable dont on cherche à estimer le total et la moyenne. L'estimateur du total T(Y) s'écrit :

$\hat{T}(Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{A_i}$, avec $A_i = \frac{X_i}{\sum_{i=1}^{N} X_i}$

C'est l'estimateur linéaire et sans biais de Y appelé estimateur de Horvitz-Thompson. On a : $E(T(Y)) = \hat{T}(Y) = \sum_{i=1}^{N} Y_i$

La moyenne de Y est estimée par : $\hat{\bar{Y}} = \frac{\hat{T}(Y)}{N} = \frac{1}{nN}\sum_{i=1}^{n}\frac{Y_i}{A_i}$

4. Variance de l'estimateur du total

4.1 Cas de tirage avec remise

Par définition, la variance de l'estimateur du total s'écrit : $V(\hat{T}(Y)) = \frac{1}{n}\sum_{i=1}^{N} A_i\left[\frac{Y_i}{A_i} - \left(\sum_{i=1}^{N} Y_i\right)\right]^2$

L'estimateur de la variance de l'estimateur du total s'écrit : $\hat{V}(\hat{T}(Y)) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{A_i} - \hat{T}(Y)\right)^2$

D'où l'estimateur de la variance de l'estimateur de la moyenne s'écrit : $\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2}\hat{V}(\hat{T}(Y))$

4.2 Cas de tirage sans remise

Les probabilités d'inclusion du premier et du second ordre s'écrivent :

$$\sum_{\alpha=1}^{N}\pi_\alpha = n\ ,\ \sum_{\substack{\beta=1\\ \beta\neq\alpha}}^{N}\pi_{\alpha\beta} = (n-1)\pi_\alpha\ ,\ \sum_{\alpha}\sum_{\substack{\beta\\ \beta\neq\alpha}}\pi_{\alpha\beta} = n(n-1)$$

L'estimateur du total T(Y) s'écrit : $\hat{T}(Y) = \sum_{i=1}^{n}\frac{y_i}{\pi_i}$

L'estimateur de la variance de l'estimateur de Y s'écrit : $\hat{V}(\hat{T}(Y)) = \frac{1}{2}\sum_{i}\sum_{\substack{j\\ j\neq i}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$

En pratique, on se donne un jeu de probabilités d'inclusion et on cherche l'algorithme de tirage qui respecte ces probabilités.

Remarques :

- Choisir les π_i les plus proportionnels possibles aux Y_i
- Avoir si possible : $\pi_i\pi_j - \pi_{ij} \geq 0$ (conditions de Yates-Grundy)
- Pour un échantillon de taille n finie, si les probabilités d'inclusion sont égales, alors il s'agit d'un sondage auto-pondéré qui se dépouille comme un recensement. Cela signifie que quelle que soit la valeur de i, le poids de l'unité i est constant : $W_i = \frac{1}{\pi_i} = W_0$
- D'où : $\hat{T}(Y) = \sum_{i=1}^{n}\frac{y_i}{\pi_i} = W_0\sum_{i=1}^{n} y_i$

5. Conditions optimales d'application (taille des unités primaires, variance intra-UP, etc.)

L'application d'un tirage d'un échantillon à probabilités inégales nécessite une certaine proportionnalité entre la variable d'étude et le critère de taille choisi.

Exemple 1 : Lorsqu'on veut mesurer la production agricole, on peut tirer les exploitations agricoles proportionnellement à leur superficie en supposant que plus la superficie des terres cultivées est grande plus la quantité récoltée est importante. C'est souvent le cas rencontré dans les pays africains où il est souvent pratiqué une culture extensive. Par-contre le choix de ce mode de tirage peut se révéler inefficace, si nous sommes en présence d'une culture intensive à rendement élevé.

Exemple 2 : plusieurs études ont prouvé que les dépenses de consommation finale d'un ménage dans les pays africains (notamment pour la consommation alimentaire) sont généralement proportionnelles au nombre de personnes qui le composent. Il est indiqué dans ce cas, pour l'estimation de cet agrégat, de procéder à un tirage à probabilités inégales des ménages. Toutefois, s'il est vrai que les dépenses sont fonctions des tailles des ménages, elles sont aussi liées à leur niveau de revenu. Le revenu, n'est pas par-contre forcément lié à la taille du ménage.

Dans le cas d'un tirage d'un échantillon de taille n à probabilités inégales avec remise, la condition suivante doit être respectée : $\frac{1}{n} \geq \frac{X_i}{\sum_{i=1}^{N} X_i}$ **quelle que soit la valeur de i et la variable X désigne le critère de taille.**

6. Les algorithmes et logiciels de tirages

6.1 Tirage de Bernoulli

Soit X_i le critère de taille utilisé pour le tirage avec remise d'une unité u_i. Alors on a : $A_i = \frac{X_i}{\sum_{i=1}^{N} X_i}$

Avec $\sum_{i=1}^{n} A_i = 1$. Le dispositif de tirage se présente comme suit :

Unités	Valeurs A_i	Valeurs cumulées de A_i
1	A_1	A_1
2	A_2	$A_1+ A_2$
3	A_3	$A_1+ A_2+ A_3$
.	.	
.	.	
N	A_N	$A_1+ A_2+ A_3+….+A_N=1$

Pour chaque tirage, on choisit un nombre aléatoire u compris entre 0 et 1. On sélectionne l'unité i si :

$A_1+ A_2+ A_3+….+A_{i-1} < u \leq A_1+ A_2+ A_3+….+A_i$

Exemple du tirage de 3 ménages parmi 6 proportionnellement à leur taille.

Ménage	Nombre de personnes dans le ménage (X_i)	A_i	Cumul de A_i
1	6	0,14	0,14
2	8	0,19	0,33
3	5	0,12	0,44
4	10	0,23	0,67
5	5	0,12	0,79
6	9	0,21	1,00
Total	43	1	

On choisit 3 nombres aléatoires entre 0 et 1 : 0,5639, 0,0344 et 0,8003. Les numéros de ménages correspondants sont : 4, 1 et 6 sont choisis.

6.2 Tirage systématique

Une autre procédure est le tirage systématique :

Etape 1 : cumuler la taille des ménages. Le total est 43.

Ménage	Nombre de personnes dans le ménage (X_i)	Cumul de X_i
1	6	6
2	8	14
3	5	19
4	10	29
5	5	34
6	9	43
Total	43	

Etape 2 : déterminer le pas de tirage P en divisant le total obtenu par le nombre 3 (nombre d'unités à sélectionner).Ici, le pas P vaut 14,333.

Etape 3 : tirer un nombre aléatoire entre 0 et 1 (exemple, le nombre tiré est 0,1977).

Etape 4 : multiplier le pas de tirage par le nombre aléatoire tiré. Dans notre cas, le produit donne $F_1 = 14,2598$.

Etape 5 : choisir l'unité i telle que : $X_1+ X_2+ X_3+….+X_{i-1}<F_1\leq X_1+ X_2+ X_3+….+X_i$. Dans notre cas, F1 est compris entre 14 et 19, soit les valeurs cumulées respectivement devant les ménages 2 et 3. On sélectionne alors l'unité 3.

Etape 6 : ajouter le pas de tirage à la valeur de F1 pour repérer la deuxième unité tirée. Dans notre cas, on obtient F2 = 28,5931, compris entre 19 et 29. C'est l'unité 4 qui est choisie. La troisième unité est obtenue en calculant F3 =42,9265. C'est l'unité 6 qui est choisie.

Attention : les deux modes de tirage ne doivent pas aboutir obligatoirement à un même échantillon en termes de composition.

Plusieurs logiciels tels que SPSS, R, STATA disposent d'un module qui permet de faire le tirage d'un échantillon à probabilités inégales.

7. Exposés des cas pratiques

Un Démographe étudie en 2015, la fécondité dans une localité appelée «Vital ». La base de sondage est le RGPH de 2010 et comprend 20 ménages présentés dans le tableau ci-après :

Numéro du ménage	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Effectif de femmes en âge de procréer (15 à 49 ans) dans les ménages	1	3	4	1	2	1	1	1	2	5	1	1	1	2	1	3	2	2	2	2

Il tire 8 ménages proportionnellement à leur taille exprimée en nombre de femmes en âge de procréer. En supposant qu'il a procédé à un tirage systématique et qu'il a tiré le nombre aléatoire compris entre 0 et 1 qui vaut 0,05374651.

Questions :

1. Déterminer l'échantillon des ménages
2. Calculer la probabilité d'inclusion des ménages sélectionnés

Les ménages sélectionnés sont renumérotés de 1 à 8 pendant l'enquête. Le tableau suivant donne le nombre de naissances vivantes observées dans les ménages de l'échantillon.

Numéro du ménage	1	2	3	4	5	6	7	8
Nombre de naissances vivantes	2	2	6	0	2	0	3	1

3. Estimer le nombre total de naissances vivantes dans la localité « Vital».
4. Calculer l'estimateur de la variance du nombre de naissances vivantes.

Chapitre 5 : Sondage à plusieurs degrés

1. Introduction

Il est généralement attribué au sondage aléatoire simple deux principaux inconvénients :

- Le coût prohibitif de la collecte des données lorsque les unités sélectionnées sont suffisamment dispersées
- La mise à jour au préalable de la base de sondage comprenant tous les individus de l'univers étudié

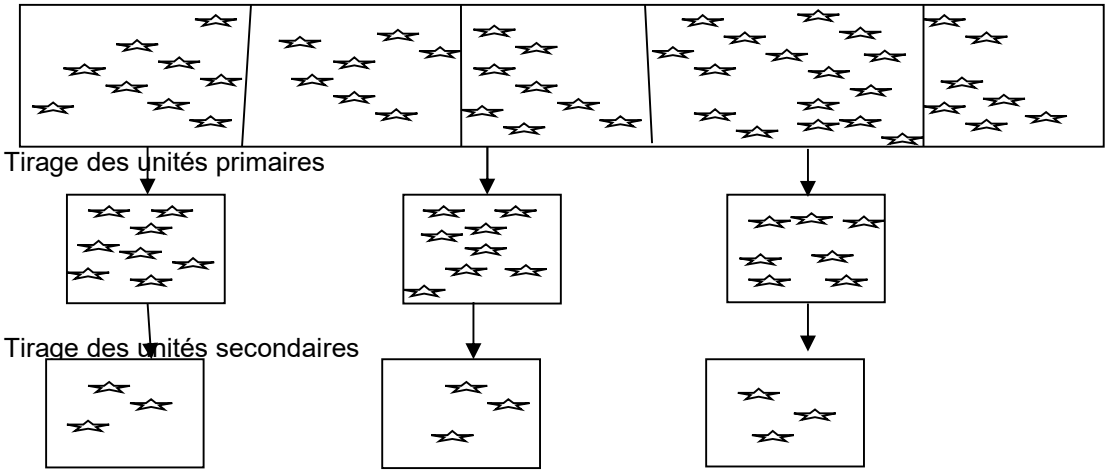
Exemple 1 : il est difficile de réaliser une enquête sur la prévalence d'une maladie par sondage aléatoire simple dans un pays. Il est d'abord difficile d'inventorier tous les individus affectés par la maladie. Même si c'est le cas, le coût de déplacement pour la collecte des données sera prohibitif pour des raisons de dispersion de l'échantillon.

Exemple 2 : Il est difficile d'estimer la production nationale d'un produit agricole à l'aide d'une enquête agricole par sondage aléatoire simple. L'opération coûtera très chère.

Pour résoudre les difficultés du sondage aléatoire simple, en pratique, il est plutôt fait recours au plan de sondage à plusieurs degrés dont le principe est le suivant :

1) On procède au découpage de l'univers en des zones ou groupes d'individus disjoints deux à deux.
2) On tire au premier degré, un nombre fini de zones ou de groupes d'índividus. Chaque zone ou groupe tiré constitue une unité primaire notée UP.
3) On tire au second degré, un nombre finis d'individus dans chaque UP. Chaque individu sélectionné à ce stage constitue une unité secondaire, notée US.
4) L'unité secondaire peut aussi être un sous groupe d'individus dans lequel il pourrait être tiré au troisième degré, des individus.

Ci-après un schéma qui illustre un tirage aréolaire à deux degrés.



La nécessité de mettre en place un sondage à deux ou plusieurs degrés est due souvent à la non connaissance a priori de l'effectif total de la population dans l'univers (noté N). En d'autres termes, N est aléatoire et il faut plus tard l'estimer.

Avantages : Coût moins onéreux pour la mise à jour de la base de sondage. Seule la connaissance des unités primaires est nécessaire pour procéder à leur tirage. Budget de collecte de données moins coûteux (les déplacements sont réduits)

Inconvénients : risque d'homogénéité à l'intérieur des unités primaires (effet de grappe). En général, la grande partie de la variance est expliquée par le tirage du premier degré.

2. Sondage équiprobable à deux degrés

On parle de sondage équiprobable, lorsqu'il y a tirage des unités avec probabilité égale. Dans le cas d'espèce, il s'agit d'un tirage aléatoire à deux degrés avec probabilité égale à chaque niveau.

2.1 Probabilité d'inclusion

De façon générale, dans un sondage à deux degrés, la probabilité totale d'inclusion d'une unité dans l'échantillon s'écrit : $\pi = \pi_1 \times \pi_{2/1}$

Avec π_1 et $\pi_{2/1}$ désignant respectivement la probabilité de tirage d'une unité primaire au premier degré et d'une unité secondaire dans une unité primaire sélectionnée.

Soient :

M : le nombre total d'unités primaires dans l'univers (UP)

m : le nombre d'UP à tirer

N_i : le nombre total d'unités secondaires dénombrées dans l'UP i sélectionnée au premier degré (i= 1 à m)

n_i : le nombre d'unités secondaires tirées dans l'UPI i.

y_{ij} = valeur de la variable Y observée ou calculée pour l'unité secondaire j située dans l'unité primaire i.

La probabilité d'inclusion s'écrit : $\pi = \frac{m}{M} \times \frac{n_i}{N_i}$

L'inverse de la probabilité d'inclusion est le coefficient de pondération des unités enquêtées.

2.2 Estimateur du total et de la moyenne

Soit T(Y) un total à estimer. La formule de Horvitz Thompson s'écrit : $\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^{m} \hat{T}_i(Y)$

Avec $\hat{T}_i(Y) = \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$. D'où $\hat{T}(Y) = \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$

L'estimateur du total est un estimateur sans biais, car l'estimateur de chaque degré est un estimateur sans biais.

L'estimateur de la moyenne s'écrit : $\hat{\bar{Y}} = \frac{\hat{T}(Y)}{N}$

Cas particulier : si on a la relation : $\frac{N_i}{n_i}$ est une constante quel que soit i, alors on a : $\hat{T}(Y) = \frac{M}{m} \times \frac{N_i}{n_i} \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}$ Il s'agit d'un sondage autopondéré.

2.3 Variance de l'estimateur du total

La variance de l'estimateur du total s'écrit :

$$V(\hat{T}(Y)) = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m} + \frac{M}{m} \times \sum_{i=1}^{M} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{S_{2,i}^2}{n_i}$$

On note : $A = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{S_1^2}{m}$

Et $B = \frac{M}{m} \times \sum_{i=1}^{M} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{S_{2,i}^2}{n_i}$

Le terme A fait intervenir seulement les paramètres liés au tirage du premier de degré. L'expression ressemble à la variance de l'estimateur du total dans le cas d'un SAS. Il s'agit en réalité d'une variance inter UP.

Le terme B par-contre caractérise la dispersion de la variable Y au sein d'une UP i quelconque.

On note : $S_1^2 = \frac{1}{M-1} \times \sum_{i=1}^{M} \left(T_i - \bar{T}\right)^2$ et $S_{2,i}^2 = \frac{1}{N_i - 1} \times \sum_{j=1}^{N_i} \left(Y_{i,j} - \bar{Y}_i\right)^2$

$$\bar{T} = \frac{1}{M} \times \sum_{i=1}^{M} T_i \quad T_i = \sum_{j=1}^{N_i} Y_{ij} \quad \bar{Y}_i = \frac{1}{N_i} \times \sum_{i=1}^{N_i} Y_{ij}$$

L'estimation de la variance de l'estimateur du total s'écrit :

$$\hat{V}(\hat{T}(Y)) = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{s_1^2}{m} + \frac{M}{m} \times \sum_{i=1}^{m} N_i^2 \times \left(1 - \frac{n_i}{N_i}\right) \times \frac{s_{2,i}^2}{n_i}$$

On note : $s_1^2 = \frac{1}{m-1} \times \sum_{i=1}^{m} \left(\hat{T}_i - \hat{\bar{T}}\right)^2$ et $s_{2,i}^2 = \frac{1}{n_i - 1} \times \sum_{j=1}^{n_i} \left(y_{i,j} - \hat{\bar{Y}}_i\right)^2$

$$\hat{\bar{T}} = \frac{1}{M} \times \sum_{i=1}^{m} \hat{T}_i$$

$$\hat{T}_i = \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

$$\hat{\bar{Y}}_i = \frac{1}{n_i} \times \sum_{i=1}^{n_i} y_{ij}$$

Cas particulier de sondage autopondéré (n_i/N_i est une constante) :

$$\hat{V}(\hat{T}(Y)) = M\left[\frac{(1-f_1)}{f_1} \times s_1^2 + \frac{(1-f_2)}{f_2} \times \frac{1}{m} \times \sum_{i=1}^{m} N_i \times s_{2,i}^2\right]$$

Avec $f_1 = m/M$ et $f_2=n_i/N_i$

Si on augmente m et sans modifier n_i, les deux termes A et B diminuent ensemble, entraînant une diminution significative de la variance. Par contre l'augmentation de n_i sans toucher à m ne fait que baisser le terme B. Il est alors préférable d'augmenter m ou m et n_i pour améliorer l'estimateur de la variance.

1. Sondage à probabilités inégales au premier degré

3.1 Probabilité d'inclusion

On aborde ici le cas de tirage avec remise. La probabilité d'inclusion s'écrit : $\pi = \pi_1 \times \pi_{2/1}$.

Soit π_{i1} la probabilité d'inclusion d'une unité primaire i dans l'échantillon du premier degré. Cette probabilité doit être comprise entre 0 et 1. Pour un échantillon de taille fixe m à tirer dans une population finie de taille M, on a : $m = \sum_{i=1}^{M} \pi_{i1}$. Dans le cas d'un tirage d'une unité i avec une probabilité proportionnelle à sa taille notée X_i, on écrit : $\pi_{i1} = m\frac{X_i}{\sum_{i=1}^{M} X_i}$. **Cette relation est vraie si on a :** $\frac{1}{m} \geq \frac{X_i}{\sum_{i=1}^{M} X_i}$ **quelle que soit la valeur de i. Cette condition est très importante à vérifier lorsqu'on porte le choix sur un tirage à probabilité inégale au premier degré.**

Soient :

N_i : le nombre total d'unités secondaires dénombrées dans l'UP i sélectionnée au premier degré (i= 1 à m)

n_i : le nombre d'unités secondaires tirées dans l'UPI i.

y_{ij} = valeur de la variable Y observée ou calculée pour l'unité secondaire j située dans l'unité primaire i.

Si le tirage des unités secondaires est équiprobable, alors la probabilité finale d'inclusion d'une unité secondaire dans l'échantillon s'écrit : $\pi_i = m \times \frac{X_i}{\sum_{i=1}^{M} X_i} \times \frac{n_i}{N_i}$

L'inverse de la probabilité d'inclusion est le coefficient de pondération des unités enquêtées.

En pratique, les X_i désignent généralement la taille des UP dans la base de sondage avant leur tirage. Tandis que les N_i représentent les effectifs des unités secondaires dénombrées dans les unités primaires qui ont été sélectionnées. Les X_i et les N_i ne sont pas souvent égaux. Ils peuvent cependant

s'égaliser si la base de sondage des UP vient nouvellement d'être constituée. Dans ce cas, le dénombre des UP sélectionnées ne semblent plus nécessaires.

On a alors : $\pi_i = m \times \frac{n_i}{\sum_{i=1}^{M} N_i}$

3.2 Estimateur du total

Soit T(Y) un total à estimer. La formule de Horvitz Thompson s'écrit : $\hat{T}(Y) = \frac{1}{m} \sum_{i=1}^{m} \frac{\hat{T}_i(Y)}{A_i}$

Avec $A_i = \frac{X_i}{\sum_{i=1}^{M} X_i}$ et $\hat{T}_i(Y) = \frac{N_i}{n_i} \times \sum_{j=1}^{n_i} y_{ij}$

Notons $K = \sum_{i=1}^{M} X_i$. K est une constante. Si les n_i sont constants ($n_i = n_0$), alors on a :

$$\hat{T}(Y) = \frac{K}{mn_0} \sum_{i=1}^{m} \sum_{j=1}^{n_0} \frac{N_i}{X_i} \times y_{ij}$$

Conseils pratiques : Il faut s'assurer que X_i et N_i s'équivalent à priori. Si les rapports $\frac{N_i}{X_i}$ se situent dans la fourchette 1 à 3, l'on peut tirer sans crainte un nombre constant n_0 d'unités secondaires dans chaque UP. Dans ce cas, il s'agit d'un sondage auto-pondéré et on a : $\hat{T}(Y) = \frac{N}{mn_0} \sum_{i=1}^{m} \sum_{j=1}^{n_0} y_{ij}$. Les données sont dépouillées comme dans une opération de recensement.

Dans le cas contraire, on tire les unités secondaires à probabilité proportionnelle à la taille. En d'autres termes, on peut imposer un taux de sondage constant $\frac{n_i}{N_i} = \frac{n}{N} = cste$

3.3 Variance de l'estimateur du total

$$V(\hat{T}(Y)) = \frac{1}{m} \sum_{i=1}^{M} A_i \left(\frac{T_i(Y)}{A_i} - T(Y) \right)^2 + \frac{1}{m} \sum_{i=1}^{M} \frac{Z_i}{A_i}$$

Z_i désigne la variance de l'estimateur de $T_i(Y)$ tenant compte du plan de sondage du deuxième degré.

Estimateur de la variance de l'estimateur

$$\hat{V}(\hat{T}(Y)) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{\hat{T}_i(Y)}{A_i} - \hat{T}(Y) \right)^2$$

Chapitre 6 : Sondage en grappe

1. Introduction

Le sondage en grappe est une forme de sondage à deux degrés. Les unités primaires sont d'abord sélectionnées au hasard. Toutes les unités secondaires dans les UP sélectionnées sont alors enquêtées. On parle de ratissage.

On peut assimiler le sondage en grappe à un sondage aléatoire simple ou l'unité d'échantillonnage serait la grappe.

2. Probabilité d'inclusion

Notation :

M : nombre total de grappes

m : nombre de grappes sélectionnées

Y : variable d'intérêt

La probabilité d'inclusion s'écrit : m/M

3. Estimation du total (Horvitz-Thompson)

$$\hat{T}(Y) = M \sum_{i=1}^{m} \frac{T_i(Y)}{m}$$

Avec $T_i(Y)$ désignant le vrai total dans la grappe i. cette fois-ci $T_i(Y)$ est connu et n'a plus à être estimé.

4. Variance de l'estimateur du total

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

Où $S_1^2 = \frac{1}{M-1} \sum_{i=1}^{M} (T_i - \overline{T})^2$ et $\overline{T} = \frac{1}{M} \sum_{i=1}^{M} T_i$

L'estimateur sans biais de la variance s'écrit :

$$\hat{V}(\hat{T}(Y)) = M^2 \times \left(1 - \frac{m}{M}\right) \times \frac{s_1^2}{m}$$

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

Chapitre 7 : Sondage empirique

1. Introduction

Pour plusieurs raisons (existence de base de sondage à jour, coût des enquêtes), la possibilité n'est pas toujours donnée de tirer un échantillon aléatoire d'une population constituant l'univers. Dans ce cas, on fait recours à l'échantillonnage raisonné. Le principal inconvénient est l'impossibilité d'extrapoler les résultats issus d'une enquête à choix raisonné. Il est aussi impossible de pouvoir calculer l'erreur d'échantillonnage.

Cette méthode consiste à identifier dans la population, quelques critères de répartition significatifs puis d'essayer de respecter cette répartition dans l'échantillon d'unités sélectionnées. La règle ici est que l'échantillon à retenir doit avoir la même composition que la population mère par rapport à une ou plusieurs caractéristiques (sexe, âge, catégorie socioprofessionnelle, etc.).

Exemple : si l'on veut étudier la dégradation des routes d'un pays donné en choisissant quatre (4) régions sur les dix (10) que comptent le pays, on peut le faire par choix raisonné en considérant la qualité du réseau routier (dense, moyen, faible). Ainsi, on peut choisir quatre régions voisines réparties en : une à réseau dense, deux à réseau moyen et une à réseau faible.

Les principales méthodes de sondage non aléatoire sont : i) la méthode des quotas, ii) la méthode des itinéraires, iii) la boule de neige.

2. Méthode des quotas

Les caractéristiques retenues sont appelées variables de contrôle. Une variable de contrôle peut être quantitative ou qualitative et doit présenter les atouts suivants :

- Sa répartition statistique dans la population est connue ;
- Son observation est facile pour un enquêteur pour réduire suffisamment les risques d'erreur.

A titre d'exemple, la tranche d'âge, le sexe, la CSP, la région et la taille des ménages peuvent servir à établir les quotas d'individus à enquêter sur les pratiques contraceptives.

Exemple : enquête sur les pratiques contraceptives auprès de la population sexuellement active dans un pays. Les caractéristiques de population se présentent comme suit :

Caractéristiques de la population sexuellement active

Région	**Effectif absolu**	**%**	**Catégories socio professionnelles**	**Effectif absolu**	**%**
Région A	100	10	Salarié	140	14
Région B	550	55	Non salarié	770	77
Région C	350	35	Chômeur et inactif	90	9
Total	1000	100	Total	1000	100

L'échantillon qui correspondra à ces caractéristiques devra comprendre 9 catégories d'individus constituées par les trois régions et les trois catégories socioprofessionnelles. Pour tirer et enquêter un échantillon de 600 personnes, l'enquêteur sélectionnera : 600 x 0,10 x 0,77 = 46 personnes résidant dans la région A et travailleurs non salariés.

Le tableau ci-après fait l'inventaire de toutes les catégories d'individus à retenir dans l'échantillon à choix raisonné par la méthode des quotas.

Construction de l'échantillon empirique pour l'enquête sur les pratiques contraceptives

N°	Catégories d'individus dans l'échantillon	Effectif d'individus à enquêter	N°	Catégories d'individus dans l'échantillon	Effectif d'individus à enquêter
1	Région A; Salarié	8	7	Région C; Salarié	29
2	Région A; Non salarié	46	8	Région C; Non salarié	162
3	Région A; Chômeur ou inactif	5	9	Région C; Chômeur ou inactif	20
4	Région B; Salarié	46		**Total**	**211**
5	Région B; Non salarié	254			
6	Région B; Chômeur ou inactif	30			
	Total	**389**			

3. Autres exemples de sondages par choix raisonné

Echantillon de convenance

Pratique et bon marché, cette méthode est surtout utilisée pour des études exploratoires. Il s'agit par exemple des enquêtes des personnes en sortie de lieux fermés tels que les caisses des supermarchés, les restaurants qui servent la soupe populaire aux personnes sans domicile fixe, etc...

Méthode des itinéraires

Le principe de cette méthode consiste à imposer à l'enquêteur un itinéraire qu'il doit suivre et sélectionner un certain nombre d'individus à enquêter. La méthode évite le fait pour un agent enquêteur d'être tenté d'aller où il veut ou en enquêtant trop d'individus au même endroit, entraînant ainsi un biais de sélection. La méthode peut être combinée à la méthode des quotas.

Méthode de la boule de neige

Cette méthode est aussi connue par le terme de sondage déterminé par les répondants. Pour enquêter sur une population spécifique (exemple les amis du chat, les amis du chien, les professionnels de la musique Jazz, etc.), on peut se servir des amis des enquêtés pour pouvoir les atteindre.

Chapitre 8 : Calcul de précisions des estimateurs complexes

1. Introduction

Il est important de s'assurer de la fiabilité des estimateurs qui sont calculés à partir des données d'une enquête. En effet, il y a dans une enquête par sondage, deux types d'erreurs : i) les erreurs d'observation et ii) les erreurs aléatoires. Les erreurs d'observation sont généralement difficiles à quantifier. Il est alors important de prendre toutes les dispositions pour les réduire au maximum.

L'élaboration des supports de collecte appropriés, la bonne formation des agents de terrain, la sensibilisation de la population et le bon encadrement des opérations sur le terrain sont autant d'éléments qui permettent de réduire les erreurs d'observation. A cela s'ajoute l'efficacité de l'atelier de saisie et d'exploitation des données.

Par contre, les erreurs aléatoires sont dues au plan de sondage mis en œuvre pour l'enquête. Elles sont quantifiables. Le calcul de précision des estimateurs se base ainsi sur la quantification des erreurs aléatoires. En dehors d'un sondage aléatoire simple, l'application directe des formules théoriques exprimant la variance d'un estimateur est rare. En pratique, on procède souvent à des calculs approchés.

2. Indicateurs de précision

2.1 Coefficient de variation

Le coefficient de variation (CV) est le rapport de l'écart-type à la moyenne. Il est aussi appelé l'écart-type relatif et est généralement exprimé en pourcentage. Plus sa valeur est élevée, plus la dispersion autour de la moyenne est grande. C'est un indicateur de comparaison de distributions de valeurs dont les échelles de mesure ne sont pas comparables.

En sondage, le CV rapporte l'écart-type de l'estimation du paramètre θ à la valeur de cette estimation, qu'il s'agisse de la moyenne, d'un total ou d'une proportion. Plus la valeur du coefficient de variation est faible, plus l'estimation est précise.

Son expression est : $CV(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\hat{\theta}}$

D'après Statistique Canada, les appréciations suivantes sont données au coefficient de variation (voir http://www.statcan.gc.ca/pub/88-001-x/2011004/userinfo-usagerinfo-fra.htm) :

- 0% à 4,9%, excellent
- 5,0% à 9,9%, très bon
- 10,0% à 14,9, bon
- 15,0% à 24,9%, acceptable ;
- Plus de 25% précision faible.

2.2 Intervalle de confiance

Une autre façon de calculer la précision d'un estimateur est de déterminer son intervalle de confiance (IC). En sondage, un intervalle de confiance permet de définir une marge d'erreur entre l'estimateur d'un paramètre calculé sur un échantillon aléatoire et sa vraie valeur inconnue mais qui concerne toute la population.

Un intervalle de confiance (IC) à 95% est un intervalle dans lequel il y a 95% de chance de trouver la vraie valeur recherchée d'un paramètre. L'intervalle de confiance est donc l'ensemble des valeurs raisonnablement compatibles avec le résultat observé (l'estimation ponctuelle). Il donne une visualisation de l'incertitude de l'estimation.

Des intervalles de confiance à 99% ou à 90% sont parfois utilisés. La probabilité (degré de confiance) de ces intervalles de contenir la vraie valeur est respectivement de 99% et 90%.

En règle générale, l'intervalle de confiance s'écrit : $IC(\theta) = \left[\hat{\theta} - \alpha\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + \alpha\sqrt{\hat{V}(\hat{\theta})} \right]$ ou encore :

$IC(\theta) = \left[\hat{\theta}(1 - \alpha CV(\hat{\theta})), \hat{\theta}(1 + \alpha CV(\hat{\theta})) \right]$

- Pour α=1, IC(θ) est défini à 68%.
- Pour α=2, IC(θ) est défini à 95% (c'est la pratique courante).
- Pour α=3, IC(θ) est défini à 99%.

2.3 Effet de sondage (DEFF) et effet de grappe (ρ)

L'effet de grappe survient dans le cas d'un sondage à plusieurs degrés. Il s'agit d'une perte de précision sur les estimations, due à l'homogénéité des unités secondaires à l'intérieur des unités primaires. Chaque degré de sondage entraîne un effet de grappe. L'effet de grappe est mesuré par le coefficient de corrélation intra-grappe noté RHO (**ρ).**

A partir d'une valeur de **ρ** donnée, on peut calculer la précision d'un tirage à deux degrés. Dans le cas d'un tirage équiprobable à chaque degré, on peut écrire :

$$V(\hat{T}(Y)) = N^2 \times \frac{S^2}{m.\bar{n}} \times (1 + \rho(\bar{n} - 1))$$

Notations:

S^2 : dispersion vraie dans l'ensemble de la population

N : effectif total de la population

m : nombre d'unités primaires (UP) tirées

$\bar{n}$: nombre moyen d'unités secondaires tirées par UP

ρ : effet de grappe

On note que $m.\bar{n}$ est la taille finale de l'échantillon des unités secondaires.

L'expression de l'effet de sondage (noté DEFF ou design effect en anglais) s'écrit :

$$DEFF = \frac{V(estimation\ par sondage\ à\ plusieurs\ \deg rés)}{V(Estimation\ par\ sondage\ aléatoire\ simple\ sans\ remise)}$$

On montre que quand les UP sont de taille équivalente n : $DEFF = 1 + \rho(\bar{n} - 1)$. Les effets de grappe jusqu'à 0,2 sont tolérables dans un sondage à deux degrés ou plus.

3. Méthodes de calcul des variances

Les différentes expressions de calcul de précision développées ci-haut ont prouvé la présence incontournable de variance dans les formules. Deux familles de méthodes de calcul de variance sont exposées dans ce chapitre. Il s'agit des méthodes analytiques ou de linéarisation et des méthodes de réplication.

3.1 Méthode analytique ou de linéarisation

L'approche analytique consiste à calculer la variance des paramètres par des formules mathématiques, soit exactement ou par approximation. Cette approche nécessite une analyse complète du plan de sondage et une distinction précise de la formulation des estimations à chaque étape de tirage.

Les estimations et les estimateurs des variances des totaux sont développés dans les chapitres précédents. Cette partie sera consacrée au calcul de variance des indicateurs complexes que sont les moyennes et les proportions.

Cas des moyennes ou ratios

La moyenne d'une variable quantitative est le rapport entre le cumul des valeurs prises par l'ensemble des individus enquêtés sur l'effectif de ces individus. Exemple : la taille moyenne d'un ménage est le cumul de l'effectif des individus vivant dans les ménages sur l'effectif total des ménages. La moyenne s'exprime alors comme un ratio de deux variables : $R=X_1/X_2$.

L'estimateur du ratio R est le ratio des estimateurs de X_1 et de X_2 : $\hat{R} = \frac{\hat{X}_1}{\hat{X}_2}$. Par contre, la variance de l'estimateur de R n'est pas un simple rapport des variances des estimateurs de X_1 et de X_2. Selon la méthode analytique, il faut d'abord procéder à une linéarisation de l'expression de R à l'aide d'un développement limité d'ordre 1.

$\hat{R}$ est un rapport de deux estimateurs de totaux, soit $\hat{R} = F(\hat{X}_1, \hat{X}_2)$. On crée la variable Z au niveau de chaque enregistrement indicé par j :

$Z_j = \frac{dF}{d\hat{X}_1} X_{1j} + \frac{dF}{d\hat{X}_2} X_{2j}$. Dans notre cas, la variable Z se définit comme suit :

$Z_j = \frac{1}{\hat{X}_2} X_{1j} - \frac{\hat{X}_1}{\hat{X}_2^2} X_{2j}$ ou encore $Z_j = \frac{1}{\hat{X}_2}(X_{1j} - \hat{R}X_{2j})$

La variance de la variable Z est calculée en utilisant la formule appropriée selon le plan de sondage appliqué. Soit la fonction Z qui s'écrit : $Z(X_1, X_2) = \frac{1}{\hat{X}_2}(X_1 - \hat{R}X_2)$

Cas d'une proportion

La proportion est un cas particulier d'un ratio qui a à son numérateur, une constante non nulle et au dénominateur une variable Y. On écrit R = k/Y. L'estimateur de R est : $\hat{R} = \frac{k}{\hat{Y}}$

On crée la variable Z au niveau de chaque enregistrement indicé par j :

$Z_j = \frac{dF}{d\hat{Y}} Y_j$. Dans notre cas, la variable Z se définit comme suit : $Z_j = -\frac{k}{\hat{Y}^2} Y_j$ ou encore $Z_j = -\frac{kY_j}{\hat{Y}^2}$. La variance de la variable Z est calculée en utilisant la formule appropriée selon le plan de sondage appliqué.

3.2 Méthode de réplication d'échantillon

Les méthodes de réplication sont souvent utilisées pour le calcul des variances des estimateurs complexes tels que le ratio, le coefficient de corrélation linéaire, le coefficient de régression, etc.. La mise en œuvre de ces méthodes requiert des outils informatiques adéquats. Deux méthodes de réplication sont présentées dans cette partie.

a) Le bootstrap

Notons s l'échantillon de base, de taille n issu d'une population de taille N.

Etape 1 : on considère que l'échantillon s est représentatif de la population. On décide de substituer l'échantillon s à la population de départ, étant donné qu'il lui est similaire.

Etape 2 : on tire selon un plan de sondage aléatoire simple et avec remise dans l'échantillon s, des échantillons s_i^* de taille n, avec i =1, 2, …..k.

Etape 3 : Soit θ un paramètre estimé sur l'échantillon s par $\hat{\theta}(s)$. Sa variance de l'estimateur sur l'échantillon s est notée $V(\hat{\theta}(s))$. Par parallélisme, on note respectivement $\hat{\theta}(s_i^*)$ et $V^*(\hat{\theta}(s_i^*))$ l'estimateur du paramètre θ et la variance de l'estimateur sur chacun des échantillons s_i^*.

Selon la loi des grands nombres, on estime que : $V(\hat{\theta}(s)) \approx \frac{1}{k}\sum_{h=1}^{k}\left[\hat{\theta}(s_h^*) - \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}(s_i^*)\right]^2$

Enfin, il est possible de définir l'intervalle de confiance du paramètre θ, connaissant $\hat{\theta}(s)$ et $V(\hat{\theta}(s))$.

En résumé, pour mettre en œuvre la méthode de bootstrap, les données suivantes sont requises :

- la taille n de l'échantillon s ;
- le paramètre θ à calculer ;
- l'estimateur $\hat{\theta}(s)$ sur l'échantillon ;
- le nombre k de réplications d'échantillons s_i^* de taille n selon un tirage aléatoire simple avec remise ;
- les estimations $\hat{\theta}(s_i^*)$ du paramètre θ sur les échantillons s_i^*.
- La variance de l'estimateur $\hat{\theta}(s)$ calculé selon la loi des grands nombres.

b) Le Jackknife

Le jackknife est une méthode de ré-échantillonnage qui permet d'estimer la variance d'un estimateur complexe. À partir des années 70, cette méthode de rééchantillonnage a été "remplacée" par une méthode plus sophistiquée, le bootstrap.

Soit un échantillon s de taille n permettant d'estimer un paramètre θ. Le principe du Jackknife consiste à estimer la même grandeur selon le même modèle avec (n-1) observations de l'échantillon. On répète l'expérience n fois. Puis on calcule la dispersion des différentes valeurs estimées de θ. De façon pratique, les étapes de calcul sont les suivantes :

Etape 1 : Considérer le paramètre θ et son estimateur $\hat{\theta}$ sur l'échantillon s. Alors on peut écrire :

$\hat{\theta} = f(Y_1, Y_2, ..., Y_n)$ où i=1 à n sont les identifiants des n individus de l'échantillon s.

Etape 2 : Calculer n pseudo-estimateurs sur le même modèle que $\hat{\theta}$, en supprimant à chaque fois un individu parmi n tirés. On note $\hat{\theta}(j)$ l'estimateur du paramètre θ obtenu en supprimant l'observation j.

Etape 3 : Calculer pour tout j = 1 à n, $\hat{\theta}^*(j) = n\hat{\theta} - (n-1)\hat{\theta}(j)$

Etape 4 : Calculer l'estimateur jackknife obtenu par l'expression $\hat{\hat{\theta}}^* = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}^*(j)$

Etape 5 : Estimer la variance de $\hat{\theta}$ par l'expression : $V(\hat{\theta}) \approx \frac{1}{n(n-1)}\sum_{j=1}^{n}\left[\hat{\theta}^*(j) - \hat{\hat{\theta}}^*\right]^2$

On peut ensuite définir l'intervalle de confiance du paramètre θ sur l'échantillon s.

4. Exposé des cas pratiques

Exemple 1 :

Dans le cadre de l'étude de la prévalence du paludisme dans la République de Kangaré en 2013, le pays a plutôt décidé de réaliser un sondage stratifié à deux degrés. Le pays est découpé en trois strates notées A, B, C. La base de sondage est le RGPH de 2012. Le pays compte 1200 grappes ou zones de dénombrement (ZD) réparties sur les trois strates.

L'INS tire au premier degré en tout 100 ZD proportionnellement à leur taille exprimée en nombre de ménages. Les effectifs de ménages issus du RGPH dans les ZD sont notés N_{ihR} selon les strates.

Il a été effectué une opération de dénombrement des ménages dans les ZD retenues avant le tirage du second degré. Les effectifs des ménages issus de l'opération de dénombrement sont notés N_{ihD} selon les strates. Au second degré il est tiré un nombre constant n_0 = 20 ménages par ZD avec probabilité égale.

Les caractéristiques de la base de sondage et des échantillons sont précisées dans le tableau suivant :

Strate	Effectif de la population au RGPH 2012 (en nombre d'habitants)	Effectif total de ZD	Effectif de ménages par strate au RGPH	Nombre de ZD tirés au premier degré	Nombre de ménages tirés	Taux de prévalence de paludisme estimé (%)	Nombre d'individus estimés par l'enquête
A	375 000	450	75 000	38	760	48,8	380 000
B	245 000	350	40 800	30	600	70,0	270 000
C	380 000	400	69 100	32	640	33,6	400 000
Ensemble	1 000 000	1 200	184 400	100	2 000		1 050 000

Questions

1) Calculer la probabilité finale de sélection des ménages selon les strates
2) Que se passe-t-il si $N_{ihR} = N_{ihD}$ pour tout i ?
3) Donner l'expression du taux de prévalence du paludisme
4) Donner l'expression de l'estimateur du taux de prévalence par strate
5) Estimer le taux de prévalence pour l'ensemble du pays
6) Donner l'expression de l'estimateur de la variance de l'estimateur du taux de prévalence du paludisme par strate selon la méthode analytique
7) Donner l'expression du coefficient de variation du taux de prévalence du paludisme.

Chapitre 9 : Traitement des non réponses totales

1. Introduction

Il est très rare de réaliser une enquête par sondage où tous les individus sélectionnés répondent à toutes les questions. Lorsqu'un individu marque un refus catégorique de participer à l'enquête ou qu'il est considéré absent durant toute la période de la collecte des données, il est alors classé dans la catégorie des non réponses totales. Par contre, un individu ayant accepté de participer à une enquête, peut ne pas répondre à certaines questions ou fournir des réponses aberrantes qui sont rejetées lors de la phase d'exploitation des données. Il s'agit ici des individus classés dans la catégorie des non réponses partielles.

Dans ce manuel, il sera abordé seulement le traitement des non réponses totales.

2. Méthode de repondération

L'objectif visé est de se rapprocher au maximum de l'estimateur idéal en considérant uniquement les individus répondants et en modifiant leur pondération initiale.

Notations :

- Y_i : la valeur de Y associée à l'individu i de l'échantillon
- P_i : la probabilité d'inclusion de l'individu i de l'échantillon
- R_i : la probabilité de réponse de l'individu i de l'échantillon à l'enquête
- r la liste des individus de l'échantillon ayant répondu à l'enquête.

L'estimateur du total de la variable Y s'écrit : $\hat{Y} = \sum_{i \in r} \frac{Y_i}{P_i \bullet R_i}$.

Le problème est d'estimer la probabilité de réponse par un taux de réponse empirique. Dans le cas d'un sondage aléatoire auprès d'un échantillon de taille n, si n' individus ont répondu à l'enquête, alors on a, quel que soit i : $R_i = \frac{n'}{n}$ D'où : $\hat{Y} = \frac{n}{n'} \sum_{i \in r} \frac{Y_i}{P_i}$ ou encore $\hat{Y} = \frac{N}{n'} \sum_{i \in r} Y_i$ et $\hat{\bar{Y}} = \frac{1}{n'} \sum_{i \in r} Y_i$.

Il en résulte : $\hat{\bar{Y}} = \frac{1}{n'} \sum_{i \in r} Y_i$. En fait, tout se passe comme si l'ensemble des répondants est un sous-échantillon obtenu par tirage aléatoire simple parmi les individus initialement sélectionnés. **On montre que $\hat{\bar{Y}}$ est un estimateur sans biais de $\bar{Y}$.**

Dans le cas d'un sondage à plusieurs degrés, on peut calculer les taux de réponse par unité primaire (UP) définis par l'expression suivante : $R_i = \frac{n_i'}{n_i}$ où n_i et n_i' désignent respectivement le nombre d'individus sélectionnés et le nombre de répondants dans l'unité primaire i. D'où $\hat{Y} = \sum_{i \in r} \frac{n_i}{n_i'} \bullet \frac{Y_i}{P_i}$. Ainsi les $P_i' = \frac{n_i'}{n_i} \bullet P_i$ constituent le nouveau jeu de pondération des individus répondants dans l'UP i.

La propriété de $\hat{\bar{Y}}$ comme un estimateur sans biais de $\bar{Y}$ est aussi acceptable dans le cas d'un sondage complexe à probabilités égales et de taille fixe.

Des démonstrations ont prouvé que l'estimateur $\hat{\bar{Y}}$ comporte généralement des biais. Une autre méthode de repondération est le mécanisme de réponse homogène. Il consiste à regrouper les individus répondants selon leur catégorie et calculer les taux de réponse respectifs.

Exemple : une enquête a été réalisée auprès d'un échantillon de 1 000 personnes. Le tableau suivant donne la répartition de l'échantillon initial et du nombre des répondants selon le milieu de résidence.

Milieu de résidence	Echantillon initial	Probabilité d'inclusion P_i	Nombre de répondants
Urbain (U)	700	0,01	600
Rural (R)	300	0,02	250
Total	1 000		850

Soit P_i, la probabilité d'inclusion de l'individu i dans l'échantillon. Alors l'estimateur du total de la variable Y pour l'ensemble de la population peut s'écrire : $\hat{Y} = \sum_{i\in U} \frac{n_i}{n_i^{'}} \bullet \frac{Y_i}{P_i} + \sum_{i\in R} \frac{n_i}{n_i^{'}} \bullet \frac{Y_i}{P_i}$ ou encore :

$\hat{Y} = \sum_{i\in U} \frac{700}{600} \bullet \frac{Y_i}{0,01} + \sum_{i\in R} \frac{300}{250} \bullet \frac{Y_i}{0,02}$ soit encore : $\hat{Y} = 0,012\sum_{i\in U} Y_i + 0,024\sum_{i\in R} Y_i$

Dans le cas d'un sondage stratifié avec tirage équiprobable, on peut généraliser l'expression de l'estimateur du total de Y puis de la moyenne. Soit l'univers découpé en k strates. On note : N, n_h et $n_h^{'}$ respectivement le nombre total d'individus dans l'univers, la taille initiale de l'échantillon de la strate h et le nombre de répondants de l'échantillon de la strate h. On démontre que :

On peut écrire : $\hat{\bar{Y}} = \sum_{h=1}^{K} \frac{n_h}{n} \bullet \hat{\bar{Y}}_h$ avec $\hat{\bar{Y}}_h$ étant l'estimateur de la moyenne de Y calculé uniquement avec les répondants de la strate h : $\hat{\bar{Y}}_h = \frac{1}{n_h^{'}} \sum_{i=1}^{n_h^{'}} y_{ih}$

3. Exposé des cas pratiques

Exemple 1

On réalise une enquête sur les dépenses de consommation auprès d'un échantillon de n ménages tirés sur un effectif total de N ménages. On suppose qu'il s'agit d'un tirage à probabilités égales sans remise. Les informations suivantes ont été collectées dans les ménages enquêtés :

- le sexe du chef de ménage (S) ;
- la dépense effectuée par le ménage sur une période de référence fixe (D).

1) Rappeler l'expression de la dépense moyenne des ménages

2) Définir l'estimateur de la dépense moyenne des ménages

3) Définir l'estimateur de la variance de l'estimateur de la dépense moyenne par ménage

On dispose des informations auxiliaires N_1 et N_2 qui sont respectivement le nombre de chefs de ménages de sexe masculin et le nombre de chefs de ménages qui sont de sexe féminin.

4) Proposer un estimateur de la dépense moyenne des ménages qui prend en compte ces informations auxiliaires.

5) Calculer un nouvel estimateur de la variance de l'estimateur de la dépense moyenne par ménage.

Exemple 2

On reprend l'exemple 1 avec cette fois-ci l'information selon laquelle n' est le nombre de répondants à l'enquête.

1) Calculer l'estimateur de la moyenne des dépenses des ménages par la méthode de repondération. On pose la relation : $Y_i = RX_i + U_i$ où X est une variable continue qui mesure le revenu des ménages. Cette variable est collectée sur toutes les unités enquêtées (y compris les non répondants aux questions sur les dépenses). Mais on dispose également de cette information par une source extérieure. U est un petit terme qui désigne les résidus de la relation avec la condition : $\sum_{i=1}^{N} U_i = 0$.

2) Calculer de nouveau l'estimateur de la moyenne des dépenses des ménages en tenant compte de cette information.

Bibliographie

P. Ardilly (1994), les techniques de sondage, Editions Technip, Paris

Rémi Clairin et al. (1996), « manuel de sondages, applications aux pays en développement», documents et manuels de CEPED n°3, Paris, février 1996

FAO (1996), « Enquêtes agricoles à base de sondages multiples », volume 1, enquêtes courantes fondées simultanément sur des méthodes de sondages aréolaires et de sondages par listes d'exploitation

Didier Blaizeau et al. (1997), « techniques de sondage », documents de travail pour l'atelier destiné aux statisticiens des pays en développement francophones, Centre de formation de l'INSEE à Libourne (CEFIL), du 9 au 17 juin 1997

Ousman Koriko (2011), «Analyse comparative de la qualité des plans de sondage dans les enquêtes sur les dépenses des ménages en Afrique de l'Ouest : cas des pays de l'UEMOA », dans Pratiques et méthodes de sondage, sous la direction de Marie-Eve Tremblay et al., édition Dunod 2011, pp 108-112