Jeudi 26 Septembre 2024

Chapitre 01 : CONSIDERATIONS PRATIQUES DANS L'ELABORATION D'UN PLAN DE SONDAGE.

I- Introduction

Le Sondage statistique est une technique qui consiste à enquêter sur un phénomène auprès d'un échantillon d'individu sélectionné selon certaines règles scientifiques pour représenter toute la population dont il sont issus. L'exemple le plus connu est le sondage d'opinion développés au Etats-unies dès le début du XIX^e siècle.

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

Par ailleurs, la décision d'un Statisticien ou d'un chercheur de réaliser une enquête par sondage doit soigneusement être pensée. Cette décision doit tenir compte non seulement de objectifs de l'etude mais aussi d'autres facteurs tel que les contraintes budgetaires, qualité de Sondage.

Ce chapitre est consacré aux différentes considération 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 en terme d'étude et de production de données statistiques pour des prises de décisions dans les ~~domaine~~ secteurs ciblés. Il doit en outre faire ressortir le besoin à combler en réalisant une collecte de donnée. Cette analyse situationnelle devra finalement déboucher sur la justification de réaliser une enquête par rapport aux besoins d'indicateurs recherché

3- Formulation des objectifs

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

→ les besoins d'informations
→ les utilisateurs et les utilisations des données
→ les principaux concepts et définitions opérationnelles
→ le champ de l'enquête
→ le plan d'analyse

Exple : On a besoin de 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 obtenu en divisant la dépense totale de consommation par l'effectif total de la population.

L'enquête doit alors mesurer la composition des ménages, les conditions de logement et leur dépense 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 consiste à mesurer :

a - les dépenses monétaire de consommation des ménages
b - l'aide consommation
c - les transferts reçus en nature
d - les caractéristiques socio-économiques et démographiques de la population

4- Concepts et Indicateurs

Les principaux concepts de l'enquête, doivent être bien définit en tenant compte de l'exple précédent, il faudra

définir : le ménage, la dépense monétaire, l'auto-consommat°, les sources des revenus des ménages, le cadre de vie la situat° d'emploi de la population

Les concepts sont généralement abstraits. Pour les opérationnaliser, il est requis la définit° des indicateurs de suivie et des variables qui contribuent aux calculs des indicateurs et variables ou à l'explicat° du phénomène qui est suivi, l'inventaire des indicateurs à calculer, et des données à collecter est indispensables pour l'élaborat° des questionnaires et des outils de collectes.

Le tableau suivant illustre l'opérationnalisat° d'un concept par la mesure d'un indicateurs.

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

Concepts	Indicateurs	Variables requises
dépense de consommat° par tête : dépense effectuée en moyenne pour un citoyen pour satisfaire ses besoins de consommat° finale	Dépense de consommat° par tête : dépense totale de consommat° de la population	- Noms des biens et services consommés - Dépense de consommat° calculé pour l'ensemble des ménages. - Taille des ménages : Nb unités de consommat°

5- Champs géographique et social de l'enquête

Le champs social d'une enquête est l'ensemble des populat° cible de l'enquête. Dans l'exple précédent, la populat° cible peut être l'ensemble des ménages toute catégories confondus.

La définition du champs géographique impact le coût d'une enquête. Il y'a lieu de savoir, si les résultats attendus seront significatifs au niveau national ou infranational.

La définit° du champs géographiques et du champs social de l'enquête induit des questions suivantes pour la méthodologie de collectes :

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

6- Etape de conception d'une enquête

- Concept et
- → Contexte et justification
- → Définition des objectifs
- → Concept et indicateurs
- → Champs de l'étude
- → Plan de sondage

→ Conception du questionnaire
→ Collecte de donnée
→ Traitement de donnée
→ Analyse et diffusion des données
→ Documentation de l'enquête

02-10-2024

7- Détermination de la taille de l'échantillon.

La plus grande quête à laquelle doit répondre un Statisticien d'enquête, est avant tout, 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 03 préoccupations à savoir :

- la précision des estimateurs
- les contraintes de mise en oeuvre du plan de sondage
- l'efficacité du plan de sondage

a- Précision des estimations

Le recours à une base de données d'enquêtes antérieures ayant déjà permit de calculer les indicateurs recherché est très important. Il permet de tirer les enseignement 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 résultat (taux de réponses, erreurs d'observation etc), les indicateurs estimations des principaux indicateurs et leurs précisions.

Toute 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 prévue. Par ailleurs, la taille de l'échantillon doit être déterminé en fonction de :

→ la fréquence d'apparition des événements des phénomènes étudiés
→ du niveau des strates d'analyse

Toutefois, certaines formules que seront présentées dans les chapitres suivant, peuvent être utilisées pour le calcul de la taille d'échantillon lors d'une enquête

b- Contraintes de mise en oeuvre 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 reparti en coût fixe et coût variable.

Les coûts fixes sont généralement incompressible indépendamment de la taille de l'échantillon à enquêter (coordinateur, fonctionnement, organisateur des ateliers de formateur des agents etc)

Les coûts variables sont étroitement liés à la taille de l'échantillon (nombre de questionnaire, déplacement et salaire des agents de terrain et exploitateur des données etc)

c- 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 oeuvre.

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 que il y'ait une bonne maîtrise des coûts de déplacement.

Le statisticien peut aussi décider de réaliser un sondage stratifié pour améliorer la précision des résultats à condition de bien choisir les critères de stratification et de ne pas avoir trop plusieurs strate qui n'apporte 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 :

→ Couverture exhaustive
→ Pas de double enregistrement
→ Pas d'unité morte
→ Non basé essentiellement sur des informations anciennes type

Comme exple de base de sondage, on peut citer : une base de sondage aréolaire constitué des images satellitaires et de la télédétection.

une base de sondage liste est un exple plus typique (plus utilisé)

Le terme base de sondage désigne généralement une liste ~~contraintes~~ concrète d'unité ayant un lien avec la population à étudiée. Il s'agit de la duplication d'élément déjà existante sous forme de liste carte, annuaire à partir desquelles on peut constituer des unités et sélectionner un ensemble d'unité à enquêter.

La base de sondage doit comprendre toute les informations auxiliaires (mesure de taille, donnée démographique etc) nécessaires pour la mise en oeuvre technique spécial de sondage tel que la stratification ou les deux types de tirage retenu.

La construction d'une base de sondage

doit tenir compte de :
→ la nature de la population utilisé exple individu, famille, ménage, exploitation
→ la répartition géographique de la population cible (est elle permis à une localité, une région ou est elle repartit sur l'ensemble du pays)
→ la nature des opérations de terrains (enquête par téléphone, par correspondance ou interview direct par des enquêteurs)
Une base de sondage peut convenir, pour une enquête donnée et ne pas être adapté pour un autre type d'enquête.
Les unités statistiques dans la base doivent être de taille assez homogène. Le nombre d'unité dans la base doit être connu pour permettre les extrapolations

3- Principaux parametres à estimer par une enquête

Les principales caractéristiques d'une variable d'interet y sont souvent estimés lors d'une enquête par sondage :
→ le total
→ la moyenne (la proportion est une moyenne calculé pour une variable qualitative)
→ le ratio
→ la variance et le coefficient de variation des estimateurs
→ l'effet de sondage

10 - Type d'erreur rencontrés dans une enquête par sondage

a- erreur d'échantillonage

Elle est dû au plan de sondage et est mesurable par le biais, la variance ou l'ecart quadratique moyen. L'estimateur de l'erreur de l'échantillonage suit une loi de probabilité.

b- erreur d'observation ou erreur de mesure

C'est une erreur liée aux dispositifs et au support de collecte des 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

c- erreur dû au defaut de couverture ou à la non reponse

Elle est liée d'une part, à l'utilisation d'une base de sondage incomplete et d'autre part à la non reponse complète ou partielle à l'enquête par certains individu. Les non reponses peuvent être négligées ou imputées
Par contre il difficil de corriger à pas

postériorité l'erreur dû au défaut de non couverture

II - Choix et justification du plan de sondage

Les différents type de sondage que nous étudierons sont :
- → le sondage aléatoire simple
- → le sondage à probabilité illegal
- → le sondage stratifié
- → le sondage en grappes
- → le sondage à plusieurs degrés

Le choix du 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 scanné, le statisticien peut décider de réaliser un sondage aléatoire simple ou avec stratification si des variable de stratification existent. Par contre, la réalisation d'une étude qui concerne une grande population pourra utilisé un plan de sondage par grappes ou à plusieurs degrés avec des possibilité de stratification

I2 - Notations

a - Sur l'univers (Population totale)

→ Unité statistique $x = 1, \dots, N$

→ moyenne de la variable Y

$$\mu = \frac{1}{N} \sum_{x=1}^{N} Y_x$$

→ Variance de Y : $V(Y) = \frac{1}{N} \sum_{x=1}^{N} (Y_x - \mu)^2$

On utilise souvent la notation σ^2 pour la variance $V(Y) = \sigma^2$. Par ailleurs, on définit aussi la variance corrigé $S^2 = \frac{1}{N-1} \sum_{x=1}^{N} (Y_x - \mu)^2$

b - Sur l'échantillon

→ Unité statistique $i = 1, \dots, n$

→ $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$ moyenne de la variable Y calculé sur l'échantillon. $\hat{\mu}$ est une variable aléatoire, $E(\hat{\mu})$ est l'espérance de la variable aléatoire $\hat{\mu}$, $V(\hat{\mu})$ est la variance de variable aléatoire $\hat{\mu}$ qui est rappellons le, la variance de l'estimateur $\hat{\mu}$ en non de la variance de la variable Y calculé sur les unités de l'échantillon

c - Taux de sondage

On le note $f = \frac{n}{N}$

Mercredi, 09 Octobre 2024

Chapitre 02 :

I - DEFINITIONS.

Le sondage aléatoire simple est le modèle d'échantillonage en apparence le plus simple que l'on puisse imaginer. Il consiste à considérer que dans une population d'effectif (N), tous les échantillons de taille (n) unités sont possibles avec la même probabilité.

Un plan de sondage est dit avec remise si un même individu peut apparaître plusieurs fois dans l'échantillon et si l'ordre dans lequel apparaissent les individus compte.

Exple : P = {bleu, blanc, rouge, violet, vert}

n = 3

l'échantillon {blanc, blanc, rouge} est différent de l'échantillon {blanc, rouge, blanc}

Dans le cas d'un plan avec remise il y'a N^n échantillon possible.

Un plan de sondage est dit sans remise si un individu ne peut apparaître une seule fois dans l'échantillon.

Dans l'exple précédant, l'échantillon {blanc, blanc, rouge} n'est pas possible. Dans le cas d'un plan sans remise, il y'a

$$C_N^n = \frac{N!}{n!(N-n)!} \text{ échantillon possible}$$

La plus part du temps nous nous intéressons au plan sans remise. Car interroger deux fois la même individu n'apporte pas d'information supplementaire. Cependant il n'est pas inintéressant de considérer les plans avec remise, notamment pour servir d'élément de comparaison et de référence.

Un plan de sondage aléatoire est dit simple ou à probabilités égales si chaque échantillon à la même probabilité qu'un autre d'être tiré au sort.

Exple : Dans le cas simple d'un plan sans remise, un échantillon de taille fixe n a donc une probabilité $\frac{1}{C_N^n} = \frac{n!(N-n)!}{N!}$ d'être tiré au sort, puisqu'il y a C_N^n échantillons possibles.

Si N = 5 et n = 2, cette probabilité est

donc égale à $= \frac{2 \times 3 \times 3}{5 \times 4 \times 3 \times 2} = \frac{1}{10}$

Remarque : les données concernant la population toute entière (X_i, $\bar{X}$, T, P ... etc) sont inconnus et déterministes puisque l'on a pas accès aux informations ~~concernant~~ concernant toute la population. En revanche, les valeurs obtenus à partir de l'échantillon sont connues et aléatoire, elles dépendent en effet du hasard puisqu'elle varie d'un échantillon aléatoire à un autre et elles sont connus puisqu'on dispose des informations nécessaires pour les calculer sur l'échantillon.

II - PLAN SIMPLE SANS REMISE :

1- Plan de sondage et probabilité d'inclusion

Soit un échantillon aléatoire S, la variable aléatoire $1\{k \in S\}$, $k \in U$ nous sera très utile. Notons que c'est bien une variable aléatoire puisque S est aléatoire (Variable indicatrice de Comfield $I_k = (1$ quand $k \in S$ et 0 quand $k \notin S)$

Définition

La probabilité d'inclusion de la k-ème unité noté π_k correspond à la probabilité que cette k-ème unité $= \pi_k = P_r(k \in S)$

$$= \sum_{S \in \mathcal{C}} P(S)\, 1\{k \in S\}$$

$$= \sum_{S \ni k} P(S) \quad k \in U$$

Notons également que par définition, π_k c'est $E(1\{k \in S\})$.

De même nous pouvons définir les probabilités d'ordre supérieur.

La probabilité d'inclusion d'ordre 2 est la probabilité que deux (02) unités distinguent appartiennent simultanément à un échantillon c'est-à-dire :

$$\pi_{kl} = P_r(k \in S, l \in S) = \sum_{S \ni k,l} P(S), \quad k, l \in U, k \neq l$$

Notons comme précédemment que :

$\pi_{kl} = E(1\{k \in S\}, 1\{k \in S\})$, on a :

$\text{Var}(1\{k \in S\}) = E(1\{k \in S\}) - E(1\{k \in S\})^2 = \pi_k(1-\pi_k)$

et $\text{Cov}(1\{k \in S\}) = E(1\{k \in S\}\, 1\{k \in S\}) - E(1\{k \in S\})$

$E(1\{k \in S\})$, $k, l \in U$, $k \neq l$

Dans la suite on notera :

$$\Delta_{kl} = \begin{cases} \text{Cov}(1\{k \in S\}, 1\{k \in S\}), & k \neq l \\ \text{Var}(1\{k \in S\}), & k = l \end{cases}$$

Il est souvent utile de connaître les probabilité d'inclusion afin d'établir les T estimateurs et sa variance par exple ces probabilités d'inclusion se calculent com-

me sorte : $\pi_k = \sum_{s \ni k} P(s) = \binom{n-1}{N-1} \cdot \binom{n}{N}^{-1} =$ Nbre d'échantillon contenant k

$= \frac{(N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} = \frac{n}{N}$

$\pi_{kl} = \sum_{s \ni k,l} P(s) = \binom{n-2}{N-2} \cdot \binom{n}{N}^{-1}$ Nbre échan $\in k,l$

$= \frac{(N-2)!}{(n-2)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} = \frac{n(n-1)}{N(N-1)}$

De ces deux (02) expressions, on déduit :

$\Delta_{kl} = \begin{cases} \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)} & , k \neq l \\ \frac{n}{N}\left(1 - \frac{n}{N}\right) = \frac{n(N-n)}{N^2} & , k = l \end{cases}$

2 - Le Π-estimateur

Les estimateurs de Horvitz - Thompson sont des outils essentiels pour les statisticiens, permettant d'obtenir des estimations ~~possibles~~ précises, même dans des contextes de sélection inégale. La maîtrise de leur utilisations peux permettre aux chercheurs d'approfondir la qualité de leurs analyses et de leurs résultats.

Dans de nombreux cas, les unités d'une population n'ont pas toutes la même probabilité d'être sélectionné dans un échantillon.
Par exemple dans un sondage sur les ménages, il peut être plus probable de sélectionner certains types de ménage (comme ceux avec un revenu plus élevé) en raison de la structure du sondage.

Les estimateurs de Horvitz - Thompson corrigent cette inégalité en ajustant les estimations.

* Estimation d'un total et d'une moyenne

Horvitz - Thompson (1952) ont introduit un estimateur linéaire sans biais d'un total t_y pour tout plan de sondage.
$\hat{t}_{y\pi} = \sum_{k \in s} \frac{y_k}{\pi_k}$. Cette estimateur est appelé le π-estimateur, l'estimateur de Horvitz - Thompson ou encore l'estimateur des valeurs dilatées.

Théorème

Si $\pi_k > 0 \ \forall k \in U$, alors $\hat{t}_{y\pi}$ estime t_y sans biais

Nous avons introduit le π-estimateur pour estimer le total t_y, mais nous pouvons également l'utiliser pour estimer la moyenne $\bar{y}_U$ par $\hat{\bar{y}}_{y\pi} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k}$

Notons toute fois que pour utiliser cet estimateurs, il faut que la taille de la population soit connue, ce que n'est pas toujours

seulement pas toujours le cas. Cependant plus que $N = \sum_{k \in U} 1$ on peut estimer à N par horvitz - thompson c-a-d $\hat{N}_\pi = \sum_{k \in s} \frac{1}{\pi_k}$

* Variance du π-estimateur

Il est également possible de connaître la variance du π-estimateur.

Théorème :
Soit $\hat{t}y_\pi$ le π-estimateur d'un total ty.
Si $\pi_k > 0$ pour tout $k \in U$ alors :
$\text{Var}(\hat{t}y_\pi) = \sum_{k,l \in U} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl}$

3 - Variance pour les plans de taille fixe

Dans le cas des plans de taille fixe, on peut réécrire la variance du π-estimateur sous une forme différentes.

Théorème
Soit $\hat{t}y_\pi$ d'un total ty, si le plan est de taille fixe n et que
alors $\text{Var}(\hat{t}y_\pi) = -\frac{1}{2} \sum_{\substack{k,l \in U \\ k \neq l}} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$

4 - Estimation de la Variance de π-estimateur

L'idée de base du π-estimateur, peut être naturellement étendu au contexte des fonctions de deux (ou) variables $f(\cdot,\cdot)$.

Théorème
Soit $f(\cdot,\cdot)$ une fonction de deux variable quelconque.
Si $\pi_{kl} > 0$ pour tous $k, l \in U$, $k \neq l$ alors
$\sum_{\substack{k,l \in U \\ k \neq l}} \frac{f(y_k, y_l)}{\pi_{kl}} \mathbb{1}(k \in s, l \in s)$ est un estimateur sans biais de $\sum_{\substack{k,l \in U \\ k \neq l}} \frac{f(y_k, y_l)}{\pi_{kl}}$

Avec les probabilités d'inclusion calculés plus haut, nous pouvons donner une version plus explicite du π-estimateur. Le π-estimateur d'une moyenne $\bar{y}$ par :
$\hat{\bar{y}}_\pi = \frac{1}{N} \sum_{k \in U} \frac{y_k}{\pi_k} \mathbb{1}_{\{k \in s\}} = \frac{1}{\pi N} \sum_{k \in U} y_k \mathbb{1}_{\{k \in s\}}$
et le π-estimateur du total ty est
$\hat{ty} = N\bar{y}$ avec $\bar{y} = \frac{1}{n} \sum_{k \in U} y_k \mathbb{1}_{\{k \in s\}}$

Puisque le plan est de taille fixe, et que les probabilités d'inclusion des deux premiers ordres sont strictement positives, on peut utiliser la formule de la variance de $\hat{ty}_\pi$ de Sen-Yates-Grundy, c'est-à-dire : $\text{Var}(\hat{ty}_\pi) = -\frac{1}{2N^2} \sum_{\substack{k,l \in U \\ k \neq l}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \times \Delta_{kl}$

$$= \frac{1}{2N^2} \times \frac{N-n}{n(N-1)} \sum_{k \neq \ell \in U} (y_k - y_\ell)^2$$

$$= \frac{N-n}{N} \frac{S_y^2}{n}$$

avec $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2$

$$= \frac{1}{2N(N-1)} \sum_{k \neq \ell \in U} (y_k - y_\ell)^2$$

Remarque :

La variance ou l'estimateur du total peut également s'écrire :

$\hat{Var}(\hat{t}_{y\pi}) = N(N-n) \frac{S_y^2}{n}$

Théorème

Pour un plan de taille fixe n simple et sans remise, la variance corrigée de la population S_y^2 est estimé sans biais par : $S_y^2 = \frac{1}{n-1} \sum_{k \in U} (y_k - \bar{y}_\ell)^2 \mathbb{1}\{k \in s\}$

Au final, on peut estimer sans biais la variance de $\hat{\mu}_{y\pi}$ par ces plans particuliers par $\hat{Var}(\hat{\mu}_{y\pi}) = \frac{(N-n)}{N} \frac{S_y^2}{n}$ et pour le π-estimateur du total t_y par :

$Var(\hat{t}_{y\pi}) = N(N-n) \frac{S_y^2}{n}$

16-10-2024

II - PLANS SIMPLES AVEC REMISE

Le plan de taille fixe n simple et avec remise correspond correspond au cadre de la statistique inférentielle. En effet, le plan de sondage consiste à sélectionner une unité aléatoire avec probabilité égale à $\frac{1}{N}$ et de recommencer l'opération n fois indépendamment. On est donc ramené au cadre des variable aléatoires indépendantes et identiquement distribué de moyenne :

$\mu_y = \frac{1}{N} \sum_{k \in U} y_k$ et de variance

$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \mu_y)^2$

La moyenne sur la population μ_y est estimé sans biais par $\hat{\mu}_y = \frac{1}{n} \sum_{k \in U} y_k \mathbb{1}\{k \in s\} = \bar{y}$.

En effet $E(\hat{\mu}_y) = \frac{1}{n} \sum y_k \frac{n}{N}$

$= \mu_y$

De plus, puisque les y_k de l'échantillon sont sélectionnés indépendamment et sont de même loi, $Var(\hat{\mu}_y) = \frac{\sigma_y^2}{n}$

Théorème :

Pour un plan de taille fixe n simple et sans remise, la variance non corrigée de la population, $\sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \mu_y)^2$ est estimé sans biais par :

$$\hat{\sigma}_y^2 = \frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2$$

Au final, la variance de $\hat{\mu}_y$ pour un plan simple avec remise est estimé sans biais par $\hat{Var}(\hat{\mu}_y) = \frac{\hat{\sigma}_y^2}{n}$

TABLEAU RECAPITULATIF DES RESULTATS

	Sans remise	Avec remise
Estimateur de la moyenne	$\frac{1}{n} \sum_{k \in U} y_k \mathbb{1}_{\{k \in S\}}$	$\frac{1}{n} \sum y_k$
Variance de l'estimateur de la moyenne	$\frac{N-n}{N} \frac{S_y^2}{n}$	$\frac{\sigma_y^2}{n}$
Estimateur de la variance de l'estimateur de la moyenne	$\frac{N-n}{N} \frac{\hat{S}_y^2}{n}$	$\frac{\hat{\sigma}_y^2}{n}$

La plus part du temps, l'on s'intéresse essentiellement au plan sans remise car interroger deux (02) fois le même individu n'apporte pas d'information supplémentaire. De plus, le sondage simple et sans remise est toujours préférable à celui avec remise. En effet, si l'on appelle $\hat{\mu}_{y\pi}$ et $\tilde{\mu}_{y\pi}$, les π-estimateurs de la moyenne avec et sans remise, alors $\forall n \geq 2$

$$\frac{Var(\hat{\mu}_{y\pi})}{Var(\tilde{\mu}_{y\pi})} = \frac{(N-n)}{N} \times \frac{S_y^2}{\sigma_y^2} =$$

$$= \frac{(N-n)}{N} \times \frac{N}{N-1} = \frac{N-n}{N-1} < 1$$

Variance du plan sans remise est plus petit qu'une variance du plan avec remise. Ainsi la précision est meilleur dans un plan sans remise qu'avec remise.

IV - PLANS SIMPLES SANS REMISE ET FONCTION D'INTERET

* Estimation d'une proportion

Il est fréquent qu'une étude porte sur une estimation d'une proportion P. Estimer une proportion revient à compter le nombre d'unité y_k avec $k \in U$ possédant une certaine caractéristique. À partir du caractère y_k on introduit un nouveau caractère $\delta_k = 1$ si y_k possède la caractéristique 0 si non $k \in U$

Estimer une proportion c'est estimer une moyenne. Toute fois pour des proportion, les expressions pour la variance se voit considérablement simplifié du fait que $\delta_k^2 = \delta_k \ \forall k \in U$. Ainsi nous aurons pour un plan simple sans remise nous au sens : $\hat{P} = \frac{1}{n} \sum_{k \in U} \delta_k \mathbb{1}_{\{k \in S\}}$

$$S_\delta^2 = \frac{1}{n-1} \sum_{k \in U} (\delta_k - \hat{P})^2 \mathbb{1}_{\{k \in S\}} = \frac{n}{n-1} P(1-\hat{P})$$

$$Var(\hat{P}) = \frac{N-n}{N} \frac{S_\delta^2}{n} = \frac{N-n}{N-1} \frac{P(1-P)}{n}$$

$$\hat{Var}(\hat{p}) = \frac{N-n}{N} \cdot \frac{\hat{P}(1-\hat{P})}{n-1}$$

V - DETERMINATION DE LA TAILLE DE L'ECHANTILLON

Avant de commencer un sondage, il est toujours souhaitable de se poser la question des incertitudes quant à nos estimations. Généralement, les limites budgetaires fixent la taille de l'échantillon et on se contentera alors de répondre si le budget alloué est suffisante pour précision donnée.

Par précision donnée, nous attendons que le paramètre d'interet θ soit contenu dans un intervalle de confiance centré en $\hat{\theta}$ avec une probabilité d'au moins $1-\alpha$. C-a-d trouver $\ell > 0$ tel que $P_R\{\theta \in [\hat{\theta} - \ell ; \hat{\theta} + \ell]\} \geq 1-\alpha$.

En supposant que notre estimateur θ suit approximativement une loi normale (ce qui sera souvent le cas) on sait que :

$$P_R\left\{\theta \in \left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\theta)}, \hat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\theta)}\right]\right\} = 1-\alpha$$

où $z_{1-\frac{\alpha}{2}}$ est le quantile d'une loi normale centré réduit au non dépassement $1-\frac{\alpha}{2}$.

C-a-d $P_R\left(Z \leq z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$, $Z \sim \mathcal{N}(0,1)$

Remarque : Puis que $\hat{Var}\,\hat{\theta}$ dépend de la taille de l'échantillon n, on cherchera la taille minimale n_0 induisant la précision requise.

Dans le cas de l'estimation de la moyenne μ_y pour un plan simple sans remise, on aura donc :

$$P_R\left\{\mu_y \in \left[\bar{y} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N}\frac{S_y^2}{n}}, \bar{y} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N}\frac{S_y^2}{n}}\right]\right\} = 1-\alpha$$

Exercice d'application :

Soit une caractéristique X définit sur une population $N=4$ Unités.

Individus : 1, 2, 3, 4

Valeur de X : 11, 10, 8, 11

1- Calculez la valeur des paramètres suivant de la population : la moyenne, la variance et la variance corrigée notées respectivement μ, σ^2 et s^2.

2- On tire un échantillon sans remise de taille $n=2$ à probabilité égale.

a- Combien d'échantillon peut on tirer ?

b- pour chaque échantillon possible, calculer pour l'échantillon, la variance et la moyenne (s^2 et $\bar{y}$).

c- Calculer $E(\bar{y})$, $Var(\bar{y})$ et $E(s^2)$.

Corrigé de la fiche de TD

Exercice 1

$(y_i - \bar{y})^2$	1,96	12,96	19,36	6,76	0,16	41,2
	1	2	3	4	5	$n=2$
	8	3	M	4	7	$N=5$

1- Calculons

- la moyenne $\bar{y}$

$$\bar{y} = \frac{1}{N}\sum y_i = \frac{8+13+M+4+7}{5}$$

$$\bar{y} = 6,6$$

- la variance corrigée s_y^2

$$s_y^2 = \frac{1}{N-1}\sum (y_i - \bar{y})^2$$

$$= \frac{41,2}{4} = 10,3$$

2- Listons tous les échantillons possibles

(1,2) (1,3) (1,4) (1,5)
(2,3) (2,4) (2,5)
(3,4) (3,5) (4,5)

3-

Ech	$\frac{1}{y}$	$(\hat{y} - \bar{y})^2$	$V(\frac{1}{y})$
(1,2)	5,5	1,21	6,25
(1,3)	3,5	8,41	2,25
(1,4)	6	0,36	4
(1,5)	7,5	0,81	0,25
(2,3)	7	0,16	16
(2,4)	3,5	9,61	0,25
(2,5)	5	2,56	4
(3,4)	7,5	0,81	12,25
(3,5)	9	5,76	4
(4,5)	5,5	1,21	2,25
		30,13	

$$E(\frac{1}{y}) \frac{1}{10}\sum \hat{y}_k$$

$$= \frac{5,5+3,5+6+7,5+7+3,5+5+7,5+9+5,5}{10}$$

$$= \frac{66}{10} = 6,6 = \bar{y}$$

d'où $\hat{y}$ est un estimateur sans biais de $\bar{y}$

$$V(\hat{y}) = \frac{1}{10}\sum (\hat{y}_k - \bar{y})^2 = \frac{31,1}{10} = 3,11$$

4) $V(\hat{y}) = (1 - \frac{n}{N})\frac{s_y^2}{n}$

$$= (1 - \frac{2}{5})\frac{10,3}{2}$$

$$= \frac{3}{5} \times \frac{10,3}{2}$$

$$= 3,09$$

5) Estimateur $\hat{V}(\bar{Y}) = \frac{1}{n}\sum (Y_k - \frac{1}{\bar{Y}})^2$

Exercice 2

1- $n=2$; $P(\{1,2\}) = \frac{1}{2}$; $P(\{1,3\}) = \frac{1}{4}$

$P(2,3) = \frac{1}{4}$

* Non.

2- $P(1 \in G) = \sum_{S \ni 1} P(S) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$

- $P(2 \in S) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$

- $P(3 \in S) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$

3-

Ech	μ
$\{1,2\}$	1,5
$\{1,3\}$	2
$\{2,3\}$	2,5

4- $E(\hat{\mu}) = \sum p_i \mu_i$

$= \frac{1}{4} \times 1,5 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2,5$

$E(\hat{\mu}) = 1,875$

or $\mu = \frac{1+2+3}{3} = 2$

d'où $\hat{\mu}$ est biaisé.

Exercice 4

Ind	1	2	3	4	Total
Y_k	98,7	102,6	108,5	100,3	430,1
$(Y_k - \bar{Y})^2$	77,250	24,255	0,135	163,203	266,285

1- Calculons

$\bar{Y} = \frac{1}{4}\sum_{k=1}^{4} Y_k = \frac{430,1}{4} = 107,525$

$\sigma_Y^2 = \frac{266,285}{4} = 66,571$

2- On peut extraire $C_4^3 = 4$ échantillons de taille $n=3$

	$\hat{\mu}$	$\hat{\sigma}^2$	P_k
$S_1 = (1,2,3)$	107,2	14,228	1/4
$S_2 = (1,2,4)$	101,2	88,34	1/4
$S_3 = (1,3,4)$	109,166	77,388	1/4
$S_4 = (2,3,4)$	110,1666	54,148	1/4
Total	430,038		

$E(\hat{\mu}) = \frac{1}{n}\sum P_k \hat{\mu}$

$= \frac{430,038}{4} = 107,5245$

$B(\hat{\mu}) = E(\hat{\mu}) - \mu$

$= -0,0005$

$EQM = B^2(\hat{\mu}) + V(\hat{\mu})$

ou

$V(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{\hat{S}^2}{n}$

$= \left(1 - \frac{12}{N}\right) \cdot \frac{1}{n}\left[\frac{N\sigma^2}{N-1}\right]$

A.N

$V(\hat{\mu}) = \left(1 - \frac{3}{4}\right) \times \frac{1}{3} \times \left(\frac{4 \times 66,571}{3}\right)$

$= \frac{1}{4} \times \frac{1}{3} \times \frac{4 \times 66,571}{3}$

$= 7,396$

$EQM = (-0,0005)^2 + 7,396$

$= 7,39600025$

3 - Supposons que la femelle est l'individu 1.

	P_k
a_1	115
a_2	115
a_3	115
a_4	215

$E(a) = \sum P_k a_k$

$= \frac{1}{5}(103,266 + 107,2 + 103,166) + \frac{2}{5} \times 1,10,446$

$= 10,8126$

EXO 6

$\sum_{k=1}^{100} y_k = 2907 \qquad \sum_{k=1}^{100} y_k^2 = 154\,573$

1- L'estimateur

$\hat{\mu} = \bar{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k$

$= \frac{1}{100}\sum_{k=1}^{100} y_k$

$= 29,07$

2-

$IC\hat{\mu} = \hat{\mu} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\mu})}$

$= \hat{\mu} \pm z_{1-\frac{\alpha}{2}} \sqrt{\left(1 - \frac{n}{N}\right)\frac{S^2}{n}}$

ou $\alpha = 0,05 \Rightarrow 1 - \frac{\alpha}{2} = 0,975$

$z_{0,975} = 1,96$

$\hat{\sigma}^2 = \frac{1}{n}\sum y_k^2 - \bar{Y}^2$

$\Rightarrow \hat{S}^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum y_k^2 - \bar{Y}^2\right]$

$z_{1-\frac{\alpha}{2}} \sqrt{V(\hat{\mu})}$

$= 1,96 \times \sqrt{\left(1 - \frac{n}{N}\right) \times \frac{1}{n} \times \frac{n}{n-1}\left[\frac{1}{n}\sum y_k^2 - \bar{Y}^2\right]}$

$= 1,96 \times \sqrt{\left(1 - \frac{100}{2000}\right) \times \frac{1}{99} \times (5455,33 - 29,07^2)}$

$\frac{S^2}{n} \quad \frac{1}{n} S^2$

Mercredi, 30 Octobre 2024.

Chapitre 03 : SONDAGE A PROBABILITES INEGALES.

I- CARACTERE AUXILIAIRE ET PROBABILITES D'INCLUSION.

Soit X_k, $k \in U$ les Valeurs prises par le Caractère auxiliaire. Notons que cela implique sa connaissance sur toute la population. Notre étude porte toujours sur une fonction d'intérêt telle que la moyenne ou le total d'un caractère y. Le principe d'un plan à probabilité inégales, consiste à définir les probabilités d'inclusion du 1er ordre proportionnel aux X_k.

Rappelons que pour un plan de taille n, la Variance du Π-Estimateur, du total t_y est :

$$Var(\hat{t}_y) = \frac{1}{2} \sum_{k \neq l \in U} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl} \quad (1)$$

Si nous souhaitons minimiser (1), en jouant seulement sur les probabilités d'inclusion de 1er ordre π_k, prendre $\pi_k = \frac{y_k}{\sum_{l \in U} y_l}$ (proportionnel à) y_k, $k \in U$, est un choix judicieux puisque $Var(\hat{t}_y)$ est alors nulle. Bien évidemment, cette approche est impossible puis qu'elle suppose que l'on connaît les valeurs prises par le caractère y sur toute la population U (il serait alors inutile de faire un sondage).

En revanche, si nous disposons d'un caractère auxiliaire x connu sur toute la population et dont on pense qu'il est approximativement proportionnel à y, alors on gagnera à définir les probabilités d'inclusions du 1er ordre proportionnellement aux x_k.

Remarque : si au contraire, le caractère x n'est pas du tout proportionnel à y, le plan de sondage selon alors cet axe serait inexploitable et il sera préférable de prendre un plan simple.

Puisque, pour un plan de taille fixe n, $\sum_{k \in U} \pi_k = n$ (2) on aura obtenir les probabilités d'inclusion proportionnel aux x_k c-a-d $\pi_k = c x_k$ avec $c = \frac{n}{\sum_{k \in U} x_k} = \frac{n}{t_x}$

Toutefois, il y a aucune garanti que les $\pi_k \in [0,1]$, il sera fréquent que

certaines probabilités ou d'inclusions soient > 1. Pour que de telles situations, on sélectionnera d'office les unités correspondantes, c-à-d, $\pi_k = 1$, et l'on recommencera la procédure avec les unités restantes en prenant soin de diminuer la taille de l'échantillon n dans (2).

Exemple : Considérons la population $U = \{1, 2, 3, 4, 5, 6\}$ avec une variable auxiliaire x telle que $x_1 = 1$, $x_2 = 3$, $x_3 = 10$, $x_4 = 70$, $x_5 = 90$, $x_6 = 120$.

On a donc $t_x = 300$. Si l'on souhaite obtenir un plan de taille fixe $n = 3$ alors les probabilités d'inclusion temporaires

$\pi_1 = \frac{3 \times 1}{300} = 0,01$ $\quad$ $\pi_4 = \frac{3 \times 70}{300} = 0,7$

$\pi_2 = \frac{3 \times 3}{300} = 0,03$ $\quad$ $\pi_5 = \frac{3 \times 90}{300} = 0,9$

$\pi_3 = \frac{3 \times 10}{300} = 0,01$ $\quad$ $\pi_6 = \frac{3 \times 120}{300} = 1,2 > 1$

L'unité 6 est alors sélectionnée d'office. Le total sur la 6ème unité est $\sum_{k \in U} x_k - x_6 = 120 = 180$ et les probabilités d'inclusions deviennent :

$\pi_1 = \frac{2 \times 1}{180} = 0,011$ $\quad$ $\pi_3 = \frac{2 \times 10}{180} =$

$\pi_2 = \frac{2 \times 3}{180} =$ $\quad$ $\pi_4 = \frac{2 \times 70}{180}$

duepl

$\pi_5 = \frac{2 \times 90}{180} = 1$

On arrête ici

$\pi_1 = \frac{1}{90}$; $\pi_2 = \frac{1}{10}$; $\pi_3 = \frac{1}{9}$; $\pi_4 = \frac{7}{9}$

$\pi_5 = \pi_6 = 1$ les unités 5 et 6 sont donc sélectionnées d'office et il restera à choisir une unité entre 1, 2, 3 et 4

Notons que $\sum_{k=1}^{6} \pi_k = \frac{1+3+10+70}{90} + 2 = 3$

comme souhaité.

Rappelons qu'un plan de sondage est défini par les $P(s)$ et non par les π_k. Pour avoir un plan à probabilité inégale, il faut donc définir un plan de sondage $P(s)$, telle que $\forall k \in U \quad \sum_{s \subset U, s \ni k} P(s) = \pi_k$

$\mathcal{P}_n = \{s \subset U \text{ tel que } |s| = n\}$

Remarque : Il existe une infinité de plan de sondage vérifiant ces conditions. Nous allons donc introduire quelques plans de sondage à probabilité inégale à taille fixe n couramment utilisés.

I - PLAN DE POISSON

Pour mettre en œuvre le plan de sondage aléatoire de Poisson :

→ On considère n probabilités $\pi_1 - \pi_N$

→ On génère grand N nombre $u_1 - u_N$ (indépendamment des uns des autres) suivant la loi uniforme $\mathcal{U}([0,1])$.

→ $\forall i \in 1,\dots N$, l'individu k_i est sélectionné si $u_i < \pi_i$

→ les individus sélectionnés constitués l'échantillon.

Le Plan de Poisson a de très bonne qualité mais également un gros défaut, la taille de l'échantillon ne pouvant être sue qu'à l'avance.

Puisque les unités sont sélectionnés indépendamment, on a : $\pi_{kl} = pr(k\in s, l\in s) = Pr(k\in s) Pr(l\in s) = \pi_k \pi_l$ et donc $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l = 0 \quad \forall k \neq l$

Le plan de sondage est donc donné $\forall s \subset U$, $P(s) = \prod_{k\in s} \pi_k \times \prod_{k\in U} (1-\pi_k)$ avec $k\in s$,

Proba de sélectionner les unités choisis — Proba de ne pas sélectionner les unités non choisis

Puis que $\Delta_{kl} = 0 \quad \forall k \neq l$, la variance du $\bar{Y}$ estimateur du total $\hat{t}_y$ est $Var(\hat{t}_y) = \sum_{k,l\in U} \frac{y_k y_l}{\pi_k \pi_l} \Delta_{kl} = \sum_{k\in U} \frac{y_k^2}{\pi_k} \pi_k(1-\pi_k)$

$= \sum_{k\in U} \frac{y_k^2 (1-\pi_k)}{\pi_k}$ ne peut qu'être estimé par $\hat{Var}(\hat{t}_y) = \sum_{k\in U} \frac{y_k^2(1-\pi_k)}{\pi_k^2} \mathbb{1}\{k\in s\}$

Le plan de Poisson est intéressant car il est simple.

Nous introduisons maintenant une mesure du désordre

Définition :

On appelle entropie d'un plan $P(\cdot)$ la quantité $I(P) = -\sum_{s\subset U} P(s) \ln P(s)$ avec la convention que $0 \ln 0 = 0$

L'entropie est toujours positive. De plus, comme mesure du désordre, plus $I(P)$ sera grand, plus le plan $P(\cdot)$ sera aléatoire

Pour des probabilités d'inclusion fixés, on cherchera donc le plan le plus aléatoire désordonné, il est celui maximisant l'entropie

Solution : $\sum_{s\subset U, k\in U} \prod \pi_k = \prod_{k\in U} (1+\pi_k)$

B - Sondage systématique à probabilité inégale.

Ce plan de sondage a été introduit vers 1950 et est toujours largement utilisé puisqu'il a le mérite d'être simple et exacte. Contrairement au plan de Poisson, il est également de taille fixe.

On donne tirés des échantillons dont les probabilités d'inclusion d'ordre 1 sont fixées à priori et telle que $0 < \pi_k < 1$

$k \in U$ et $\sum_{k \in U} \pi_k = n$.

Définissons les probabilités d'inclusions cumulées $C_k = \sum_{l=1}^{k} \pi_l$ $k \in U$, $C_0 = 0$

Comment

L'approche à générer u et de sélectionner des unités à partir de cette unique réalisation. La première unité sélectionnée notée k_1 sera telle que $C_{k_1 - 1} \leq u \leq C_{k_1}$

La deuxième unité sélectionnée notée k_2 sera cette fois ci telle que $C_{k_2 - 1} \leq 1 + u \leq C_{k_2}$

Et ainsi de suite ...

De manière générale, la j-ième unité sélectionnée noté k_j sera alors $C_{k_j - 1} \leq j - 1 + u \leq C_{k_j}$

Exemple : Prenons la situation où $N = 6$ $n = 3$ $\pi_1 = 0,2$; $\pi_2 = 0,7$; $\pi_3 = 0,8$; $\pi_4 = 0,5$; $\pi_5 = \pi_6 = 0,4$.

Déterminer l'échantillon sélectionné sachant que $u = 0,3658$

$C_0 = 0$

$C_1 = \pi_1 + \pi_2$

$C_0 \leq u < C_1$

1° : $0 \leq 0,3658 < 0,2$ X

2° : $C_1 \leq 0,3658 < 0,7$ V

3° : $C_2 \leq 1 + 0,3658 < 1,7$

$0,9 \leq 1,3658 < 1,7$ V

4° : $C_3 \leq 2,3658 < 2,2$ X

$1,3 \leq 2,3658 \leq 2,2$ X

$1,3 \leq 2,3658 \leq 2,2$ X

5^e : $2,4 \leq 2,3658 \leq 2,6$

$2,2 \leq 2,3658 \leq 2,6$ V

6^e : $2,6 \leq 3,3658 \leq 3$ X

Mercredi, 06 Novembre 2024

Chapitre 04 : SONDAGE STRATIFIÉ

I - PRINCIPES ET JUSTIFICATIONS

Dans un sondage aléatoire simple, tout les échantillons de population de taille N sont possibles avec la même probabilité. On imagine que certains d'entre eux puissent satisfaire à priori intéressable. Prenons par exemple le cas où nous disposons de cinq (5) jetons : -1, 2, 4, 10 et 20 dont nous voulons évaluer la moyenne ($\mu = ?$) à l'aide d'un échantillon de taille 2. Parmi les échantillons à deux (02) unités, on trouve les cas extrêmes $\{-1, 2\}$ et $\{10, 20\}$, qui sont particulièrement mauvais.

Plus concrètement, dans l'étude du lancement d'un nouveau produit financier, on peut supposer des différences de comportement entre les petits et les gros clients de la banque. Il serait malencontreux que les hasards de l'échantillon conduisent à n'interroger que des clients appartenant à une seule de ces catégories, ou simplement que l'échantillon soit trop déséquilibré en faveur de l'une d'elles. S'il existe dans

de sondage une information auxiliaire permettant de distinguer, à priori, les catégories de petits et gros clients, on aura tout à gagner à utiliser cette information pour répartir l'échantillon dans chaque sous population. C'est le principe de la stratification : découper la population en sous-ensembles appelés strates et réaliser un sondage dans chacune d'elles.

L'intérêt de cette méthode en comparaison des plans simples, est qu'elle permet d'améliorer la précision des estimateurs. Elle nécessite l'utilisation d'une information auxiliaire connue pour l'ensemble de la population.

Exple : Nous souhaitons estimer l'age moyen de toutes les personnes évoluant sur le site UJI. La base de sondage est composée de l'ensemble des personnes de l'UJI. Supposons que nous disposons de la répartition des éléments de la base suivant les catégories :
- Etudiants ;
- Enseignants ;
- Personnel administratif.

Autrement dit, nous connaissons la répartition des personnes de l'UJI suivant ces trois (3) catégories. Il y a fort à parier que la variable age ne se comporte pas de la même manière dans ces trois classes (en moyenne) on peut penser en effet que la population enseignante ou personnel administratif est plus âgée que la population étudiante) Il paraît dès lors pertinent que l'on prenne en compte cette information dans le plan de sondage.

La répartition des personnes de l'UJI fournit une information auxiliaire en lien avec notre problématique.

L'objectif principal consiste donc à mettre à profit cette information pour obtenir des résultats précis. L'information auxiliaire peut être utilisée à deux moments :
- à l'étape de la conception du plan de sondage.
- à l'étape de l'estimation des paramètres.

Dans ce chapitre, nous utiliserons cette information pour bâtir le plan de sondage.

I - POPULATION ET STRATES

Supposons que la population U soit partitionnée en H sous ensembles $U_1, \dots, U_H$ appelés

petite strate et telle que l'union des U_h $\bigcup_{h=1}^{H} U_h = U$, $U_h \cap U_h = \emptyset$ $\forall l \neq h$

Chaque strate U_h admet une taille N_h et l'on a $\sum_{h=1}^{H} N_h = N$ où N est la taille de la population U

Remarque :

Les tailles de strates N_h sont ici supposés connues et constituent l'information auxiliaires.

Notre but étant toujours d'estimer un total ou une moyenne, remarquons que le total (respectivement la moyenne) s'écrit à l'aide des strates $t_y = \sum_{k \in U} y_k = \sum_{h=1}^{H} \sum_{k \in U_h} y_k$

$= \sum_{h=1}^{H} t_{y,h}$

où $t_{y,h}$ est le total des valeurs prises par le caractère y sur la strate U_h

c-a-d $t_{y,h} = \sum_{k \in U_h} y_k$

De même la moyenne sous la population s'écrit $\mu_y = \frac{1}{N} \sum_{k \in U} y_k$

$= \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \mu_{y,h}$

où $\mu_{y,h}$ est la moyenne des valeurs prises par le caractère y sur la strate U_h

c-a-d $\sigma^2_{y,h} = \frac{1}{N_h} \sum_{k \in U_h} (y_k - \mu_{y,h})^2$

$\mu_{y,h} = \frac{1}{N} \sum_{k \in U_h} y_k$

$S^2_{y,h} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \mu_{y,k})^2$

Remarque :

La variance sur la population (total), σ^2_y s'écrit :

$$\sigma^2_y = \frac{1}{N} \sum_{k \in U}$$

$$= \frac{1}{N} \sum_{h=1}^{H} \sum_{k \in U_h} [(y_k - \mu_{y,h}) + (\mu_{y,h} - \mu_y)]^2$$

$$= \frac{1}{N} \sum_{h=1}^{H} \left\{ \sum_{k \in U_h} (y_k - \mu_{y,h})^2 + 2(\mu_{y,h} - \mu_y) \underbrace{\sum_{k \in U_h} (y_k - \mu_{y,h})}_{0} + N_h (\mu_{y,h} - \mu_y)^2 \right\}$$

$$= \frac{1}{N} \sum_{h=1}^{H} N_h \sigma^2_{y,h} + \frac{1}{N} \sum_{h=1}^{H} N_h (\mu_{y,h} - \mu_y)^2$$

$= \sigma^2_{y\ intra} + \sigma^2_{y\ inter}$

où $\sigma^2_{y\ intra}$ est la variance intra strate (moyenne des variances des strates) et $\sigma^2_{y\ inter}$ la variance inter strate (due aux différences entre strates).

13-11-2024

II - ECHANTILLONS, PROBABILITÉ D'INCLUSION ET ESTIMATION

Un sondage est dit stratifié, si pour chaque strate, on tire un échantillon selon un sondage aléatoire simple sans remise de taille fixé n_h et que les tirages au sein de chaque strate sont mutuellement indépendants.

Soit S_h l'échantillon aléatoire tiré dans la strate U_h à l'aide d'un plan de sondage $P_h(.)$. L'échantillon aléatoire S obtenu au final est donc $S = \bigcup_{h=1}^{H} S_h$.

Le plan de sondage associé n'est rien d'autre que $P(S) = \prod_{h=1}^{H} P_h(S_h)$, $S = \bigcup_{h=1}^{H} S_h$

et la taille de l'échantillon S est $n = \sum_{h=1}^{H} n_h$

Le calcul des probabilités d'inclusion pour un sondage stratifié n'est pas difficile mais il faut tout de même faire attention. Pour les probabilités d'inclusion d'ordre 1 et si l'unité $k \in U_h$ alors $\pi_k = \frac{n_h}{N_h}$ puisqu'on a effectué un plan simple sans remise de taille n_h pour cette strate.

Pour les probabilités d'inclusion d'ordre 2, c'est un peu plus difficile et le résultat dépend du fait que les unités $k, \ell \in U$ de la même strate ou non.

- si k et ℓ $\in$ à la même strate alors $\pi_{k\ell} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$

- si k, ℓ $\in$ à deux strates différentes U_{h_1} et U_{h_2} alors (par indépendance entre les strates) $\pi_{k\ell} = \pi_k \pi_\ell = \frac{n_{h_1} n_{h_2}}{N_{h_1} N_{h_2}}$

En conséquence on a :

$$\Delta_{k\ell} = \begin{cases} \frac{n_h}{N_h}\left(1 - \frac{n_h}{N_h}\right), & k = \ell, \ k \in U_h \\ -\frac{n_h(N_h - n_h)}{N_h^2(N_h-1)}, & k \neq \ell, \ k, \ell \in U_h \\ 0, & k \in U_{h_1}, \ \ell \notin U_{h_1} \end{cases}$$

Sous coup, les π-estimateurs du total t_y et la moyenne $\bar{y}$ sont :

$$\hat{t}_{y,\text{strat}} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{h=1}^{H} \sum_{k \in S} \frac{n_h y_k}{n_h} = \sum_{h=1}^{H} t_{y,h}$$

$$\hat{\bar{y}}_{y,\text{strate}} = \frac{1}{N} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in S_h} y_k = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h$$

où $\hat{t}_y$ est l'estimateur du total pour la strate h (c-a-d $\hat{t}_y = \frac{N_h}{n_h} \sum_{k=1}^{n_h} y_k$)

et $\bar{y}_h$ est la moyenne de l'échantillon prélevé à la strate h (c-a-d $\bar{y}_h = \frac{1}{n_h} \sum_{k \in s_h} y_k$).

Exemple : les résultats du sondage sont donnés dans le tableau suivant.

Strate	1	2	1	3	1	1	2	3	2	1
Ages	26	30	25	42	23	22	35	44	38	26

1- Calculer la moyenne des ages des individus de l'échantillon par Strate

2- Déterminer la moyenne μ, calculer la variance de la moyenne, calculer la variance du total et calculer la variance estimée $\hat{S}^2$

Solution.

1* Calcul de la moyenne des ages de l'échantillon par strate.

$\bar{y}_h = \frac{1}{n_h} \sum_{k \in s_h} y_k$

Pour $h=1$, $\bar{y}_1 = \frac{1}{n_1} \sum_{k \in s_1} y_k$ avec $n_1 = 5$ $\bar{y}_1 = \frac{1}{5}(26+25+23+22+26) = 23,2$

Pour $h=2$ $\bar{y}_2 = \frac{1}{n_2} \sum_{k \in s_2} y_k$

$\bar{y}_2 = \frac{1}{3}(30+35+38) = 41$

Pour $h=3$ $\bar{y}_3 = \frac{1}{n} \sum_{k \in s_h} y_k$

$\bar{y}_3 = \frac{1}{2}(42+44) = 43$

2- Calcul de la moyenne μ

$\bar{y}_{h,\text{strate}} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_k$

$\Rightarrow \mu = \frac{1}{10}(5 \times 23,2 + 3 \times 41 + 2 \times 43) = 36,5$

$\text{Var}(\hat{\mu}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} S_h^2$

$\text{Var}(\hat{t}) = \sum_{h=1}^{H} \text{Var}(\hat{t}_{y_h})$

$\text{Var}(\hat{t}) = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{\hat{S}_h^2}{n_h}$

$\hat{S}_{y,h}^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2$, $h = 1, 2, \ldots, H$

IV - PLAN STRATIFIÉ ET ALLOCATION PROPORTIONNELLE

1 - Définition

Un plan stratifié est dit à allocation proportionnelle si $\frac{n_h}{N_h} = \frac{n}{N} \quad \forall h = 1, \dots, H$

C'est à dire que, les strates de tailles importantes auront plus d'unités dans l'échantillon que celles de tailles plus petites

Remarque : Généralement, la taille de l'échantillon pour chaque strate $n_h = \frac{n N_h}{N}$ ne sera pas entière. Mais afin de simplifier les développement théoriques qui viennent, nous allons tout de même le supposer.

Les E.T-estimateurs du total et de la moyenne sont alors

$$\hat{t}_{y,\text{strat prop}} = \sum_{h=1}^{H} \hat{t}_{y,h}$$

$$= \frac{N}{n} \sum_{k \in s} y_k$$

$$\bar{y}_{y,\text{strat prop}} = \frac{1}{n} \sum_{k \in s} y_k$$

la variance de la moyenne est alors :

$$\text{Var}(\bar{y}_{y,\text{strat prop}}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^{H} N_h S^2_{y,h}$$

$$\text{Var}(\hat{t}_{y,\text{strat prop}}) = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S^2_{y,h}}{n_h}$$

$$= \sum_{h=1}^{H} N_h \left(\frac{N}{n} - 1\right) S^2_{y,h}$$

$$= \frac{N-n}{n} \sum_{h=1}^{H} N_h S^2_{y,h}$$

Remarque : lorsque les tailles de strates N_h sont suffisamment grande, alors $S^2_{y,h} \approx \sigma^2_{y,h}$ et donc : $\text{Var}(\hat{t}_{y,\text{strat prop}}) = \frac{N-n}{n} \sum_{h=1}^{H} N_h \sigma^2_{y,h}$

$$= N(N-n) \frac{\sigma^2_{y\,\text{intra}}}{n}$$

Alors que la variance de l'estimateur du total pour un plan simple sans remise vaut : $\text{Var}(\hat{t}_{y,\pi}) \approx N(N-n) \frac{\sigma^2_y}{n}$. Les deux expressions sont quasiment identiques, mais plus que $\sigma^2_y = \sigma^2_{y\,\text{intra}} + \sigma^2_{y\,\text{inter}}$, la 1ère expression est plus petite c'a-d que l'on obtient de meilleurs résultats avec un plan stratifié avec l'allocation proportionnelle qu'avec un plan simple sans remise.

Ceci sera d'autant plus vrai que la variance inter strat sera grande, ceci est le

Cas lorsque le caractère d'intérêt y dépend fortement du caractère servant à la stratification, ici les tailles N_h.

On obtient alors biais. Cette variance par : $Var_1(\bar{y}_{st}, \pi) \leq N(N-n)\frac{\sigma^2}{n}$

$$Var(\bar{y}_{strat\ prop}) = \frac{N-n}{n} \sum_{h=1}^{H} N_h \hat{S}^2_{y,h}$$

$$\hat{S}^2_{y,h} = \frac{1}{n_h - 1} \sum_{k \in s_h} (y_k - \bar{y}_h)^2 \quad h = 1, 2, \ldots, H$$

Exple :

On donne dans le tableau suivant pour chaque étudiante de l'Université de Gaza :
- Son âge
- Sa catégorie : 1 = étudiant, 2 = enseignant, 3 = personnel administratif
- Son couleur de cheveux : a = brun, b = blond, c = châtain

Pour simplifier, on considère une population de 20 étudiantes.

Age	Catégories	cheveux
24	1	C
52	2	A
42	3	B
19	1	C
38	3	A
26	1	B
45	2	C
23	1	A
31	2	A
24	1	B

Age	Catégorie	cheveux
22	1	C
48	2	A
24	1	A
38	3	A
26	1	B
36	3	B
46	2	B
23	1	C
31	2	A
18	1	C

$n = 10$

1- On veut estimer la moyenne d'âge μ à l'aide d'un plan simple. Quelle est la variance de l'estimateur ?

2- On désire stratifier la population suivant la catégorie. Quelle est la variance de l'estimateur $\hat{\mu}$ pour un tel plan ?

3- On choisit maintenant de stratifier suivant la couleur des cheveux. Quelle est la variance de l'estimateur pour un tel plan.

Solution

20-11-2024

1 - $Var(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$

$= \left(1 - \frac{10}{20}\right) = \frac{115,305}{10}$

$= 5,77$

2 - $S^2_{y,h} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \bar{y}_{U_h})$

$N_1 = 22,3$; $S^2_1 = 6,99$
$N_2 = 44,83$; $S^2_2 = 26,17$
$N_3 = 35,5$; $S^2_3 = 6,33$

* On en déduit la variance de l'estimateur

$Var(\hat{\bar{y}}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) \frac{1}{N} \sum_{h=1}^{H} N_h S^2_h$

$= \frac{1}{10}\left(1 - \frac{10}{20}\right) \frac{1}{20} \left[10 \times 6,99 + 6 \times 26,17 + 4 \times 6,33\right]$

$= 0,63.$

3 - Variance de l'estimateur est de 4,86.

V - PLAN STRATIFIÉ OPTIMAL DU LT TOTALE

Si votre objectif est d'estimer un total ou une moyenne, alors il existe une taille optimale pour les strates. On cherche donc les tailles d'échantillon $n_1, \dots, n_H$ minimisant la variance des estimateurs du total t_y pour une taille d'échantillon fixée n, c-à-d à minimiser la variance.

$Var(\hat{t}_{y,strat}) = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S^2_{y,h}}{n_h}$

Par rapport à n_h et sous contrainte de $\sum_{h=1}^{H} n_h = n$

Le Lagrangien de ce problème de minimisation est :

$\mathcal{L}(n_1, \dots, n_H, \lambda) = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S^2_{y,h}}{n_h} + \lambda \left(\sum_{h=1}^{H} n_h - n \right)$

On a donc :

$\frac{\partial \mathcal{L}}{\partial n_h} = 0 \Rightarrow -\frac{N_h^2}{n_h^2} S^2_{y,h} + \lambda = 0$

$\Rightarrow n_h = \frac{N_h S_{y,h}}{\sqrt{\lambda}}$

Mais plus que $\sum n_h = n$, on a donc

$\frac{1}{\sqrt{\lambda}} N_h S_{y,h} = n \quad \lambda^{-\frac{1}{2}} \sum_{h=1}^{H} N_h S_{y,h} = n$

et il vient : $n_h = n \frac{N_h S_{y,h}}{\sum_{j=1}^{H} N_j S_{y,j}}$

$h = 1, \dots, H$

Remarque : la taille optimale pour une strate U_h est donc proportionnelle au produit de la taille de cette strate et de l'écart type du caractère y de cette strate.

Bien entendu en pratique, on ne connaît pas $S_{y,h}$ et donc la formule précédente n'est pas d'un grand intérêt. Elle est cependant assez instructive et intuitive ! Instructive puisque elle indique qu'il faut sur-représenté les strates qui ont une forte variabilité.

Remarque : En pratique les tailles $n_h \notin \mathbb{N}$ et on arrondira les résultats. De plus il peut arriver également que $n_h > N_h$ pour au moins un $h \in \{1, \dots, H\}$. Dans de tel cas, on prendra alors $n_h = N_h$ et on déterminera les tailles optimales sur les strates restantes — en procédant par itération si nécessaire.

Supposons que les tailles optimales soient des entiers et telles que $n_h < N_h \ \forall h$. Alors la variance du T-estimateur est donc :

$$\mathrm{Var}\left(\hat{t}_{y,\mathrm{strac}}\right) = \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S^2_{y,h}}{n_h}$$

$$= \sum_{h=1}^{H} N_h^2 \frac{\sum_{h=1}^{H} N_h S_{y,h}}{n N_h} S^2_{y,h} - \sum_{h=1}^{H} N_h S^2_{y,h}$$

$$= \left(\sum_{h=1}^{H} N_h \frac{S_{y,h}}{n}\right) \sum_{h=1}^{H} N_h S_{y,h} - \sum_{h=1}^{H} N_h S^2_{y,h}$$

$$= \frac{1}{n} \left(\sum_{h=1}^{H} N_h S_{y,h}\right)^2 - \sum_{h=1}^{H} N_h S^2_{y,h}.$$