REVUE DE STATISTIQUE APPLIQUÉE

W. E. DEMING

Quelques méthodes de sondage

Revue de statistique appliquée, tome 12, nº 4 (1964), p. 11-55
<http://www.numdam.org/item?id=RSA_1964__12_4_11_0>

© Société française de statistique, 1964, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

QUELQUES MÉTHODES DE SONDAGE

W. E. DEMING

NEW-YORK University

I - GENERALITES

I. 1 - SONDAGES PROBABILISTES

Dans ce qui suit nous envisageons uniquement les sondages probabilistes, c'est-à-dire ceux qui font appel à la théorie statistique pour fixer le plan de l'enquête afin de permettre la généralisation, à l'ensemble de la population à laquelle on s'intéresse, des résultats obtenus par l'observation de l'échantillon.

Dans cet article on étudiera les méthodes d'échantillonnage du point de vue de la recherche, particulièrement dans les sciences sociales, en indiquant seulement les points théoriques essentiels.

Dès que le questionnaire ou la méthode d'essai sont fixés, l'emploi judicieux de la théorie statistique, pour choisir le plan de sondage, permet d'obtenir l'information maximale pour une dépense donnée. Cette même théorie permet ensuite de calculer les marges d'incertitude dues à des causes accidentelles de natures diverses telles que la variabilité naturelle entre les unités qui constituent la population (fluctuations d'échantillonnage), les variations accidentelles dans les techniques des enquêteurs, aussi bien que celles dues aux procédés de dépouillement des résultats.

Ces considérations théoriques relatives à l'étude de l'incidence des fluctuations d'échantillonnage ne s'appliquent pas aux procédés d'échantillonnage par choix raisonné (par ex. méthode des quotas) dans lesquels l'échantillon est choisi suivant des règles résultant d'un jugement basé sur la connaissance préalable de la population ou est simplement constitué par une sous-population dont les unités sont déjà réunies (par ex. les élèves d'une classe).

---

(1) Conférence présentée à l'institut de Statistique de l'Université de Paris (Mai 1964). Traduction de E. MORICE.
Ce texte est tiré pour une part importante de DEMING "Sample designs in Business Research" (John Wiley - 1960).

Certaines de ces techniques, par ailleurs fort utiles et d'autre part moins coûteuses à metre en oeuvre ne permettent pas d'évaluer par des méthodes objectives l'ordre de grandeur possible des erreurs des résultats auxquels elles conduisent.

Un sondage n'est pas seulement la sélection et l'observation (interview, essai, ...) d'une partie de la population, mais c'est surtout l'emploi des théories probabilistes en vue d'acquérir un degré déterminé de confiance dans les résultats et ceci au moindre coût.

I.2 - SONDAGES ET CALCULATRICES MODERNES

Les calculatrices modernes, qui peuvent emmagasiner de grandes quantités d'informations concernant chaque unité, pouvant être récupérées et utilisées ultérieurement dans n'importe quelle combinaison en poussant un bouton, ne suppriment pas l'intérêt de sondages bien organisés.

L'expérience montre fréquemment que la masse d'informations ainsi accumulée ne contient pas celle dont on a besoin dans une étude particulière. De plus, même si cette information existe sur la carte perforée ou est emmaganisée sur le ruban magnétique, elle peut ne pas avoir la précision souhaitée pour tel usage particulier.

Dans d'autres cas, cette information globale sera cependant utile pour choisir un échantillon : stratification, estimation par la méthode du quotient (voir ci-après).

D'autre part, il ne faut pas oublier que les heures de machines sont très onéreuses et que même si l'information possédée par la machine correspond au problème que l'on veut étudier, il est souvent préférable de faire cette étude à partir d'un échantillon en réservant les heures-machine à des travaux plus productifs - compte tenu de leur coût - que cette tabulation de masse.

I.3 - QUELQUES EMPLOIS DES METHODES DE SONDAGE

- Enquêtes sociologiques et démographiques
- Etudes de budgets de familles et enquêtes sur les consommations
- Enquêtes statistiques gourvernementales (indice des prix, salaires et chômage, maladies, ...)
- Sondages effectués dans les données recueillies lors d'un recensement (réduction du coût, rapidité d'exécution et de publication des résultats)
- Vérification d'un recensement (enquêtes de contrôle par vérification d'un petit échantillon)
- Etudes agronomiques (comparaisons de variétés ou de traitements, essais d'insecticides)
- Recherche médicale (traitements, études de corrélation entre maladies et causes possibles)

- Contrôle du travail, des hommes ou des machines (méthode des observations continues, étude des temps élémentaires)
- Etudes de trafic (rail, routes, air).
- Etudes commerciales et industrielles (production, distribution, inventaire, gestion des stocks)
- Spécification et essais de matériaux
- Recherche de conditions opérationnelles optimales
- Solution d'équations méthématiques compliquées (méthode de Monte-Carlo).

Deux problèmes particulièrement importants relèvent du domaine industriel.

a) technique des cartes de contrôle pour la recherche d'éventuelles causes systématiques de variation dans les caractéristiques mesurables de qualité, les pourcentages de défectueux, les taux de production.

b) techniques de contrôle de réception par inspection d'échantillons prélevés dans les lots présentés.

I.4 - AVANTAGES ET DESAVANTAGES DES SONDAGES

Un sondage est non seulement plus rapide et moins coûteux qu'une enquête exhaustive, mais il peut être organisé avec des buts mieux précisés dans leurs détails et plus soigneusement réalisés.

Cependant, il ne fournit pas d'information détaillée concernant les unités, de plus l'erreur d'échantillonnage sur les résultats concernant de petits lots ou de petites sous-populations peut être importante.

I.5 - LES PRINCIPALES ETAPES D'UNE ENQUETE PAR SONDAGE

(Dans ce qui suit on désignera par D les responsabilités du spécialiste du domaine d'études, par S celles du statisticien).

1/ Formulation du problème en termes statistiques (D, S) : Le problème est d'abord un problème technique particulier (industriel, commercial, sociologique.....), sa transposition basée sur la théorie statistique vient ensuite : c'est alors que prennent place des considérations sur le type et la grandeur des unités d'échantillonnage, sur la précision probable de tel ou tel type de sondage et sur le coût correspondant.

2/ Définition de la population globale qui est l'objet de l'enquête (D).

3/ Définition de la population réelle dans laquelle sera prélevé l'échantillon (D), définition du type et de la taille des unités d'échantillonnage qui constituent cette base de sondage (S).

4/ Techniques de sélection de l'échantillon (S) : choix du plan de sondage donnant le maximum d'informations pour un coût donné, règles fixées pour le choix des unités d'échantillonnage (emploi des tables de

nombres au hasard, stratification....), règles relatives à l'exécution sur le terrain, rappels.....

5/ Méthodes de calcul des estimations des caractéristiques désirées (totaux, moyennes, proportions) (S)

6/ Méthodes de calcul des erreurs d'échantillonnage (S)

7/ Organisation des contrôles pour évaluation des erreurs autres que les erreurs d'échantillonnage (S)

8/ Présentation des résultats (tableaux, graphiques....) (D, S)

9/ Estimation d'ensemble de la confiance que l'on peut accorder aux résultats (S).

10/ Décisions à prendre compte tenu des résultats (D)

Le choix d'un plan de sondage implique préalablement la comparaison de divers plans possibles, dans le but de choisir le plan optimal compte tenu de la précision et du coût. La théorie statistique a permis d'élaborer une large variété de plans et fournit les moyens rationnels de comparer leurs efficacités.

I.6 - ENQUETES PAR VOIE POSTALE

Elles demandent une préparation particulièrement soignée en raison des risques de non réponse : une comptabilité des réponses reçues permettra les rappels nécessaires et éventuellement les enquêtes directes près des non réponses (ou d'une fraction de celles-ci).

Des économies importantes peuvent souvent être réalisées en commençant une enquête par voie postale et en la terminant par des interviews directs des non répondants. Les enquêtes postales sont particulièrement faciles à adapter aux études portant sur des groupes d'unités dont on possède la liste.

II - DEFINITIONS

A - POPULATION, BASE DE SONDAGE ET ECHANTILLON

II.1 - LA POPULATION OU UNIVERS

Sous ce terme il faut entendre l'ensemble de toutes les unités (individus, entreprises, pièces manufacturées, matériaux.....) que l'on veut étudier, qu'elles soient accessibles ou non.

Sa définition précise exige que soient clairement définis le problème que l'on veut résoudre et l'usage que l'on veut faire des résultats.

Par exemple, la population peut être constituée par l'ensemble des femmes d'une certaine tranche d'âges, habitant une certaine région ou

l'ensemble des firmes qui fabriquent un produit donné ou tous les enfants d'âge scolaire d'une région (population qui dans certains cas sera différente de celle des écoliers), ou toutes les vaches laitières d'une région (à la fois dans les secteurs ruraux ou urbains), ou toutes les pièces mécaniques correspondant à un contrat ou à une spécification, ou encore toutes les factures présentées (ou payées) relatives à une certaine période.

II.2 - LA BASE DE SONDAGE

La base de sondage constitue le moyen d'accès réel à la population ou à une fraction suffisante de cette population : elle est constituée par un certain ensemble d'unités d'échantillonnage appartenant à la population.

Une unité d'échantillonnage, dans une enquête sur les logements, peut être un logement ou un groupe d'un certain nombre de logements voisins (même étage ou même immeuble....).

Dans une étude portant sur un groupe professionnel, la base peut être la liste des membres d'une société professionnelle, l'unité d'échantillonnage étant une ligne de cette liste ou un groupe d'un nombre fixé de lignes consécutives.

Dans une étude des caractéristiques de trafic marchandises par camions, la base peut être constituée par les fiches de transport établies par les sociétés exerçant cette activité. Dans une comparaison de traitements médicaux la base pourra être constituée par les fiches des hôpitaux.

Chaque individu, chaque unité élémentaire figurant dans la base doit appartenir à une unité d'échantillonnage bien définie ou avoir une probabilité déterminée (ou pouvant l'être) d'appartenir à une certaine unité d'échantillonnage.

Sans base convenable, il ne sera pas possible d'organiser un sondage probabiliste.

On peut envisager des règles donnant un numéro à chaque unité d'échantillonnage de la base : un nombre au hasard fourni par des tables définira une unité d'échantillonnage et conduira à l'examen ou à l'interview des individus ou unités élémentaires de cette unité d'échantillonnage qui appartiennent à la population à moins qu'il n'ait été prévu par le plan de sondage de ne faire porter l'examen ou l'interview que sur un sous-échantillon au hasard de l'unité d'échantillonnage (sous-échantillon choisi avec une probabilité conditionnelle préfixée : sondages à plusieurs degrès).

La base de sondage, répertoire ou fichier dans lequel sont consignées les unités de sondage de l'ensemble à étudier, doit, dans la mesure du possible, être à jour, et ne comporter ni omissions, ni doubles emplois, afin de permettre le tirage des unités de sondage conformément au schéma probabiliste défini par le plan de sondage.

Dans les grandes enquêtes on dispose assez exceptionnellement de bases de sondage répondant rigoureusement à ces conditions : diverses méthodes peuvent être envisagées pour adapter au mieux les bases que l'on possède à l'enquête que l'on veut faire(1).

---

(1) cf Thionet : Application des méthodes de sondage aux enquêtes statistiques (Imprimerie Nationale, I. N. S. E. E. - 1953) N. d. I. R.

Ainsi, une liste de familles dans une région peut ne pas contenir tous les consommateurs d'un certain produit, par exemple les consommateurs habitant des pensions, des hôtels, des hôpitaux, des casernes situées dans cette région. L'extension de la base à ces secteurs particuliers et l'enquête relativement à ces consommateurs pose des problèmes différents de ceux de l'enquête pour des familles courantes. Si la proportion des personnes habitant dans ces conditions particulières est faible, le responsable de l'étude peut décider de les négliger malgré le biais qui en résulte.

Mais une base peut être sans utilisation valable si une fraction trop considérable de catégories importantes de l'univers n'y figure pas.

Il est quelquefois possible de combler au moins en partie le vide qui existe entre la population et une base disponible.

Par exemple si la base originale est une liste d'établissements importants, on pourra pour les régions auxquelles on s'interesse rechercher des listes d'établissements de taille intermédiaire (annuaires divers).

Il est possible dans certaines enquêtes de fixer des règles à l'aide desquelles, chaque unité d'échantillonnage une fois tirée d'une base incomplète pourra conduire à des listes ou cartes ne figurant pas dans la base initiale et qui pourront alors être échantillonnées avec les probabilités voulues.

II.3 - ECHANTILLONNAGE POUR L'ETUDE DE CAS RARES

Soit, par exemple, à étudier des problèmes relatifs à une catégorie exceptionnelle d'individus, alors qu'il n'existe pas de liste correspondante. Si, comme il arrive quelquefois, ces cas exceptionnels sont plus particulièrement groupés dans certains secteurs connus le problème est grandement simplifié, on pourra utiliser un taux de sondage élevé dans les secteurs où sont particulièrement concentrés les cas exceptionnels et un taux faible dans les autres, en utilisant la répartition de Neyman (voir ci-après).

Mais cette procédure n'est efficace que si les concentrations des cas rares sont très nettement différentes dans les deux types de secteurs.

On peut quelquefois envisager un test rapide et peu couteux qui permettra d'identifier surement la rareté : un essai préliminaire portera sur un large échantillon que l'on divisera en deux groupes :

1/ celui des unités qui contiennent presque surement des cas exceptionnels,

2/ celui des unités qui, presque surement n'en contiennent pas.

Un tel plan est utile dans les études de cas rares dans les hopitaux dont les fiches permettront un partage rapide d'un gros échantillon préliminaire en deux groupes.

On tirera alors des échantillons de ces deux groupes et on les soumettra à l'enquête (Neyman 1938).

L'échantillon final contiendra une forte proportion (peut-être 100 %) de la première strate et une faible proportion de la seconde.

II. 4 - DEFINITION D'UNE EGALE ET COMPLETE COUVERTURE DE LA BASE

Par définition la couverture égale et complète de la base serait ce qu'on obtiendrait si toutes les unités d'échantillonnage de la base (secteurs géographiques, établissements, comptes, articles manufacturés) étaient soumises à l'enquête, celle-ci étant réalisée par les mêmes inspecteurs, utilisant avec le même soin, les mêmes définitions et les mêmes techniques au cours de la même période.

L'échantillon est constitué par une fraction, choisie au hasard, de la couverture complète.

La couverture égale et complète peut être conceptuelle ou réelle : le recensement peut être considéré comme représentant pratiquement une telle couverture.

Cependant pour des raisons de rapidité et d'économie, de nombreux tableaux publiés à la suite d'un recensement proviennent d'échantillons tirés du recensement qui constitue la couverture complète de ces échantillons.

Il est important de noter que la théorie statistique fournit des méthodes d'induction (estimations, erreurs d'échantillonnage, tests d'hypothèse) pour la base de sondage, mais non pour la population si la base n'en donne pas une image suffisamment fidèle : la généralisation à la population ne peut être faite que compte tenu de la connaissance du contexte dans lequel se place le problème statistique étudié.

B - DEFINITIONS FONDAMENTALES

II. 5 -

Soit une base de sondage dont les unités d'échantillonnage portent les numéros 1, 2, ... N.

Considérons deux variables (X) et (Y) dont les valeurs pour chacune des unités de la base sont ;

$$a_1 \ , \quad a_2 \ldots\ldots\ldots\ldots a_N \qquad \text{pour (X)}$$

$$b_1 \ , \quad b_2 \ldots\ldots\ldots\ldots b_N \qquad \text{pour (Y)}$$

Posons :

$$A = a_1 + a_2 + \ldots\ldots\ldots + a_N = N\bar{a}$$

$$B = b_1 + b_2 + \ldots\ldots\ldots + b_N = N\bar{b}$$

$$\varphi = \frac{A}{B} = \frac{\bar{a}}{\bar{b}} \qquad (1)$$

caractéristiques de la population que l'on voudrait estimer par un sondage.

Tirons de la base à l'aide d'une table de nombres au hasard, un échantillon de n unités.

Soient x_i et y_i les valeurs observées sur l'unité de sondage N° i, définie par exemple par le $i^{ème}$ nombre au hasard.

Considérons pour l'instant les estimations respectives de $\bar{a}$, $\bar{b}$, A, B, φ, définies à partir de cet échantillon par :

$$\bar{x} = \frac{1}{n}(x_1 + \ldots\ldots + x_n)$$

$$\bar{y} = \frac{1}{n}(y_1 + \ldots\ldots + y_n)$$

$$X = n\bar{x} \qquad (2)$$

$$Y = n\bar{y}$$

$$f = \frac{\bar{x}}{\bar{y}}$$

Les erreurs d'échantillonnage de ces estimateurs sont :

$$\Delta\bar{x} = \bar{x} - \bar{a} \qquad \Delta\bar{y} = \bar{y} - \bar{b}$$

$$\Delta X = X - A \qquad \Delta Y = Y - B \qquad (3)$$

$$\Delta f = f - \varphi$$

Les résultats du sondage permettront non de calculer ces erreurs, mais d'estimer la marge de variation associée à une probabilité donnée d'être dépassée.

II.6 - ESPERANCE MATHEMATIQUE, ERREUR-TYPE ET BIAIS D'UNE ESTIMATION

Si dans la population on prélevait de manière indépendante un grand nombre d'échantillons de n observations, chacun donnant une moyenne $\bar{x}$, on pourrait construire une distribution qui serait une approximation de la distribution théorique de $\bar{x}$, distribution caractérisée par une espérance mathématique $E\bar{x}$ et une variance $\sigma^2_{\bar{x}} = E(\bar{x} - E\bar{x})^2$.

La moyenne et l'écart-type de la distribution réelle envisagée ci-dessus seraient respectivement des approximations de $E\bar{x}$ et $\sigma_{\bar{x}}$.

D'une manière générale, si $\hat{u}$ est l'estimation par le sondage d'une caractéristique dont la vraie valeur dans la population est u et si :

$$E\hat{u} = u \qquad (4)$$

la technique d'échantillonnage est dite sans biais, mais si :

$$E\hat{u} = u + c \qquad (5)$$

cette technique est caractérisée par un biais égal à c. Dans tous les cas la variance de la distribution de $\hat{u}$ est

$$\sigma^2_{\hat{u}} = E(\hat{u} - E\hat{u})^2 , \qquad (6)$$

ainsi pour une procédure donnée d'échantillonnage un estimateur a une espérance mathématique, une erreur-type, et éventuellement un biais qui, s'il existe, décroît lorsque l'effectif (ou taille) de l'échantillon augmente.

Ce biais théorique, résulte de la conception même du plan de sondage mais non des insuffisances de sa réalisation.

II.7 - MARGE D'INCERTITUDE DUE AUX VARIATIONS ACCIDENTELLES

La marge d'incertitude de l'estimateur $\bar{x}$, qui peut être attribuée au plan de sondage et aux autres variations accidentelles, est calculée sous la forme $t\,\hat{\sigma}_{\bar{x}}$, $\hat{\sigma}_{\bar{x}}$ étant une estimation de $\sigma_{\bar{x}}$. Le facteur t correspond au niveau de probabilité choisi pour cette marge, compte tenu aussi des risques qu'elle implique. Eventuellement le facteur t est aussi associé au nombre de degrés de liberté de l'estimation $\sigma_{\bar{x}}$.

Pour ces échantillons importants, la distribution de l'estimateur $\bar{x}$ et même celle de beaucoup d'autres caractéristiques est voisine de la distribution normale, excepté dans le cas où la distribution de x dans la base présente une forme très exceptionnelle.

L'écart-type $\sigma_{\bar{x}}$ contient alors toute l'information relative à la marge d'incertitude de $\bar{x}$ pouvant être attribuée aux variations accidentelles (Fisher 1922). Mais cet écart-type est une mesure insiffusante des variations accidentelles de $\bar{x}$ lorsque la distribution de x n'est pas normale : dans ce cas la présentations des résultats d'une enquête exige d'être soigneusement examinée (Fisher 1956, p. 152, Shewart 1939 p. 106).

II.8 - LIMITATION DE L'INDUCTION STATISTIQUE

L'erreur-type, ainsi que d'autres calculs qui y sont liés, ne permet ni de découvrir, ni de mesurer :

a) les composantes d'erreurs systématiques qui pourraient exister dans la base de sondage,

b) les incertitudes du type I, décrites ci-après : aucun calcul statistique ne permettra de combler les lacunes de la base.

Une petite erreur type d'échantillonnage signigie :

a) que les variations entre des échantillons répétés seraient petites,

b) que les imperfections accidentelles de l'enquête sont petites.

c) que le résultat du sondage est en bon accord avec ce que donnerait une enquête portant sur l'ensemble de la base (couverture complète).

Elle ne prouve pas que les erreurs systématiques persistantes sont petites : ces dernières pourront être mises en évidence par un contrôle statistique (nouvelle enquête portant sur un sous-échantillon de l'échantillon principal) ou par des comparaisons avec d'autres résultats connus.

C - CHOIX AU HASARD

II.9 - MELANGE ET HASARD

Il n'est jamais sans danger d'admettre a priori que les unités d'échantillonnage sont convenablement mélangées dans la base.

Celle-ci se présente souvent en groupes différents en raison de leur origine (géographique, ordre de production.....).

Le statisticien doit prendre la précaution de subdiviser la base en régions, et de tirer un échantillon de chaque région, pouvant ainsi tirer profit d'une stratification naturelle qui pourrait exister dans la base.

II.10 - VARIABLE ALEATOIRE

Une variable aléatoire est le résultat d'une opération dans laquelle le hasard seul conditionne ce résultat : les tables de nombres au hasard désignant les numéros affectés aux unités de la base permettent de constituer un échantillon au hasard, et constituent l'outil fondamental de sélection des unités de l'échantillon.

Des méthodes physiques de mélange ou de battage de cartes ne doivent pas être considérées comme satisfaisantes : il en est de même, évidemment, de celles qui seraient fondées uniquement sur le jugement de l'enquêteur.

III - CAUSES D'INCERTITUDE DANS LES OBSERVATIONS

Toutes les données recueillies, aussi bien à partir d'un échantillon que de la base complète, sont affectées de types variés d'incertitude pouvant provenir du questionnaire, des enquêteurs, des mesures physiques ou de données extraites de documents existants.

Les différences essentielles entre un sondage et un recensement sont les suivantes :

a) le sondage peut être réalisé avec plus de soin et fournir des informations auxquelles on peut accorder une plus grande confiance,

b) les résultats du sondage peuvent être affectés par des erreurs aléatoires d'échantillonnage mais celles-ci peuvent être estimées à partir des observations.

III.1 - CLASSIFICATION DES CAUSES D'ERREURS

Connaître leur existence, les classer selon leur origine, est nécessaire pour rechercher les améliorations à apporter aux enquêtes ultérieures.

La mise au point et la réalisation d'un plan d'enquête exigent la recherche d'un compromis économique entre les diverses causes d'erreurs : ainsi, par exemple, il serait sans intérêt pratique de vouloir, au prix d'un effort coûteux, réduire l'erreur d'échantillonnage à un niveau très inférieur à celui des erreurs d'autres sources.

Nous envisagerons trois types de causes d'erreurs.

III.1.1 - Causes d'erreurs dues à l'organisation du plan d'enquête et aux méthodes de réalisation

- L'information utile à recueillir n'a pas été suffisamment précisée. Le questionnaire est mal présenté (questions mal rédigées, ordre des questions),
- Les unités d'échantillonnage ne sont pas définies de manière assez précise.
- La base ne contient pas certaines classes importantes de la population.
- La date de l'enquête est mal choisie.
- Les règles de codification ne sont pas assez claires pour être correctement appliquées.
- Le programme de tabulation ne correspond pas aux intervalles de classes ou aux classifications dont on aura ultérieurement besoin.
- Coefficients de pondération mal choisis, ajustements incorrects.
- Interprétation abusive des résultats, rapport final n'attirant pas suffisamment l'attention des utilisateurs sur les marges d'erreur des résultats, et sur l'incidence possible des erreurs dues aux difficultés rencontrés dans l'éxécution de l'enquête.

III.1.2 - Erreurs dues aux imperfections et maladresses dans la réalisation de l'enquête sur le terrain

- Couverture incomplète de l'échantillon.
- Mauvaise interprétation de l'unité d'échantillonnage (oubli de certains éléments d'unités complexes ou au contraire, extension à des éléments étrangers à l'enquête, remplacement d'unités prévues par d'autres).
- Questionnaire non scrupuleusement respecté.
- Erreurs matérielles diverses (comptages, calculs, transcription.....)
- Non-réponses et refus.

III.1.3 - Erreurs aléatoires d'échantillonnage

Aux causes d'erreurs précédentes qui existent dans les recensements aussi bien que dans les sondages, il faut ajouter les variations aléatoires dues :

- aux différences entre les unités de la base dont l'échantillon a été tiré,

- aux variations accidentelles au cours du temps de tous ceux qui participent à l'enquête et aux différences qui existent entre eux.

Les erreurs de la première catégorie sont indépendantes du type de sondage et de la taille de l'échantillon : une nouvelle enquête ne permettra pas de les découvrir. Au contraire une seconde réalisation de l'enquête permettra de constater l'existence des causes d'erreurs de la seconde catégorie.

La marge d'incertitude due aux variations aléatoires de la troisième catégorie peut être estimée dans un sondage probabiliste.

L'erreur type du résultat tient compte de toutes les espèces de variations aléatoires. Il est en général possible, moyennant un plan d'enquête convenable, de mesurer les différences entre enquêteurs, contrôleurs, codifieuses.....

Un échantillon étant plus petit qu'un recensement complet pourra en général être réalisé avec plus de soin et, par conséquent être moins affecté par les causes d'erreur de la seconde catégorie.

En revanche, l'effet des erreurs aléatoires aura en général tendance à augmenter lorsque la taille de l'échantillon diminue.

III.2 - COMPARAISON D'ENQUETES

La comparaison entre enquêtes, lorsqu'elle est possible, est utile pour comprendre la nature et l'origine des résultats obtenus : des différences non négligeables peuvent provenir de petites différences dans les questionnaires, dans la formation et le contrôle du travail des enquêteurs et des codifieuses, leurs salaires, les époques.....

III.3 - CONDITIONS D'UN BON PLAN DE SONDAGE

L'augmentation de la taille de l'échantillon diminue la marge d'incertitude due aux causes d'erreurs accidentelles, mais elle n'a aucun effet sur les causes d'erreurs de la première catégorie et elle peut éventuellement accroître les erreurs dues aux causes de la seconde catégorie.

Le plan de sondage devra tenir compte de ces considérations qui impliquent, en ce qui concerne les erreurs aléatoires, des exigences raisonnables tenant compte des autres causes d'erreur.

Le plan de sondage devra prévoir les moyens statistiques de contrôle du travail des enquêteurs afin de mettre en évidence et de mesurer les effets des principales imperfections ou maladresses de leur travail (cf. exemples dans Deming, Ch. 1.9.12).

III.4 - REDUCTION DU NOMBRE DE NON REPONSES

De nombreuses caractéristiques des individus qui ne sont pas à leur domicile à la première visite ou qu'il est difficile de rencontrer, sont très différentes de la moyenne : le biais des réponses manquantes constitue l'un des plus sérieux problèmes posés par les enquêtes.

D'autres difficultés sont dues (enquêtes par voie postale ou questionnaire rempli par l'enquête), lors du dépouillement, aux réponses manquantes, illisibles, obscures ou manifestement fausses.

Accroître la taille de l'échantillon n'est pas une solution : ceci conduira à accroître ce que l'on possède déjà mais non à obtenir ce qui manque. Seuls des rappels ou de nouvelles visites permettront de réduire le biais des non réponses : on ne peut généralement pas le supprimer complètement.

Pour essayer d'estimer la valeur moyenne de la caractéristique cherchée on peut encore opérer de la façon suivante, proposée par Hansen, pour des enquêtes par interview :

Parmi les non répondants on prélève un sous-échantillon et on leur envoie d'excellents enqueteurs (l'expérience montre que certains enquêteurs n'essuient presque jamais de refus).

On calcule les valeurs moyennes, $\bar{x}_1$, pour l'échantillon des n_1 répondants à la première enquête et $\bar{x}_2$ pour le sous-échantillon des n'_2 répondants à cette seconde enquête partielle et prélevés au hasard parmi les n_2 non-répondants à l'enquête initiale.

On prendra comme estimation finale :

$$\bar{x} = \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}_2$$

avec

$$n_1 + n_2 = n$$

Cette estimation serait sans biais, si n'_2 était l'effectif total n_2 du sous-échantillon des non répondants à la première enquête.

D'autres méthodes ont été proposées, mais toutes ne sont que des palliatifs plus ou moins efficaces : le maximum d'efforts doit être fait pour obtenir effectivement une réponse de la quasi totalité de l'échantillon initial, quitte à réduire son effectif pour rester dans la limite des crédits disponibles.

IV - METHODES ELEMENTAIRES DE SONDAGE

IV.1 - PROBLEMES DE DENOMBREMENT OU D'ESTIMATION ET PROBLEMES DE COMPARAISON

Le but d'une étude énumérative est de compter le nombre d'individus qui, dans une certaine population, possèdent tels ou tels caractères sans se préoccuper de la manière dont ils les ont acquis.

Le but d'une étude analytique est de découvrir les différences et les causes de différence entre classes et de rechercher les causes de ces différences, en un mot de découvrir pourquoi les individus possèdent certaines caractéristiques.

Par exemples, envisageons une étude de la fertilité d'un certain type de malades (schizophréniques).

Le but d'une étude de dénombrement peut par exemple être de compter le nombre de malades par sexe et par âge admis et réadmis dans dans une certaine population hospitalière pendant une certaine période, ou encore d'estimer le nombre d'enfants nés avant la première attaque de la maladie, ou avant la première admission ou avant la seconde.

Le but d'une étude analytique peut être de découvrir les différences causées par divers types de traitement, ou les différences entre deux périodes de temps. On peut encore se demander si le nombre de réadmissions par patient a augmenté entre les deux périodes, ou si la proportion de malades mariés a changé, ou si l'âge moyen d'admission a varié.

Les équations permettant de définir le plan optimum de sondage seront en général différentes selon le type d'étude envisagé : il en résulte que dans une enquête où interviennent ces deux types de problèmes, les exigences des deux plans seront plus ou moins en désaccord et il sera nécessaire de trouver une solution de compromis.

Nous allons examiner plus particulièrement les problèmes du premier type.

IV.2 - EXEMPLE D'UN CHOIX ELEMENTAIRE AU HASARD

Considérons une base constituée par N unités d'échantillonnage, numérotées 1, 2, ...N.

Soient a_i et b_i les valeurs des deux caractères (X) et (Y) dans la $i^{è}$ unité ; telles qu'elles pourraient être fournies par un recensement complet.

Soient $\bar{a}$ et $\bar{b}$ les moyennes par unité d'échantillonnage, A et B les totaux, dans la base de sondage.

Désignons par :

$$\sigma_a^2 = \frac{1}{N} \sum_i^N (a_i - \bar{a})^2 \qquad (1)$$

la variance du caractère (X) entre les unités d'échantillonnage de la base et par $C_a = \sigma_a/\bar{a}$ le coefficient de variation correspondant.

A l'aide d'une table de nombres au hasard entre 1 et N, tirons un échantillon de n unités.

Le tableau ci-après donne les résultats de ce sondage relativement aux deux caractères (X) et (Y).

Caractère	Echantillon	Total	Moyenne par unité dans l'échantillon
(X)	$x_1, x_2, \ldots\ldots x_n$	x	$\bar{x} = x/n$
(Y)	$y_1, y_2, \ldots\ldots y_n$	y	$\bar{y} = y/n$

IV.3 - ESTIMATIONS DE LA MOYENNE ET DU TOTAL DANS LA BASE

Posons :

$$X = N\bar{x} \qquad Y = N\bar{y} \tag{2}$$

Le procédé de tirage donnant à chaque unité de la base la même probabilité de sélection, on aura :

$$E\bar{x} = \bar{a} \qquad EX = A \qquad E\bar{y} = \bar{b} \qquad EY = B \tag{3}$$

$\bar{x}$, $\bar{y}$, X, Y sont respectivement des estimations sans biais de $\bar{a}$, $\bar{b}$, A, B.

Considérons maintenant le rapport

$$\varphi = \frac{A}{B} = \frac{\bar{a}}{\bar{b}} \tag{4}$$

et à partir de l'échantillon, calculons l'estimation

$$f = \frac{X}{Y} = \frac{\bar{x}}{\bar{y}} = \frac{x}{y} \tag{5}$$

Si le total B dans la base est connu à partir d'une autre source (recensement, autre enquête plus importante), on peut alors estimer A par la formule

$$X' = Bf \tag{6}$$

Cette estimation de A sera appelée : estimation par le quotient.

On avait déjà l'estimation $X = N\bar{x}$, d'autres encore peuvent être envisagées (par exemple estimation par régression, voir ci-après).

La théorie statistique fournira le moyen de choisir l'estimation optimale.

IV.4 - BIAIS DANS L'ESTIMATION D'UN RAPPORT

Bien que X et Y soient des estimations sans biais de A et B, le rapport $f = X/Y$ n'est généralement pas une estimation sans biais de $\varphi = A/B$.

Le biais est approximativement égal à

$$B(f) = Ef - \varphi = \frac{\varphi}{n}(C_b^2 - C_{ab}) \tag{7}$$

avec

$$C_{ab} = \frac{1}{N\bar{a}\bar{b}} \sum_{1}^{N} (a_i - \bar{a})(b_i - \bar{b}) = \rho C_a C_b \tag{8}$$

(covariance relative entre les a_i et les b_i dans la population)

$$C_b^2 = \left(\frac{\sigma_b}{\bar{b}}\right)^2 = \frac{1}{N\bar{b}^2} \sum_1^N (b_i - \bar{b})^2 \qquad (9)$$

(variance relative des b_i dans la population).

Le biais diminue lorsque n augmente, dans de nombreux cas il est négligeable, cependant il faut y faire attention lorsque n est très petit.

Par exemple, si on utilise le département comme unité d'échantillonnage dans un sondage national ou sur une grande région, il faut éviter de calculer les rapports département par département et de calculer ensuite la moyenne pour l'ensemble car dans ce cas on aura le biais maximum (n = 1).

Le rapport $f = \frac{x}{y}$ des totaux pour l'ensemble des départements est préférable, le biais étant alors réduit dans le rapport 1/n.

IV.5 - VARIANCES DES ESTIMATIONS

Dans ce qui suit, on envisagera deux cas :

1/ Tirage sans remise des unités tirées successivement une à une,

2/ Tirage avec remise de chaque unité tirée avant nouveau tirage.

Pour l'estimation de la moyenne on aura :

1/ $$\sigma_{\bar{x}}^2 = \frac{N - n}{N - 1} \frac{\sigma^2}{n} \sim \left(\frac{1}{N} - \frac{1}{N}\right) \sigma^2 \qquad (10)$$

2/ $$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

Pour le rapport f, les approximations :

1/ $$\sigma_f^2 \sim \left(\frac{1}{n} - \frac{1}{N}\right) f^2 (C_a^2 + C_b^2 - 2\, C_{ab})$$

2/ $$\sigma_f^2 \sim \frac{f^2}{n} (C_a^2 + C_b^2 - 2\, C_{ab}) \qquad (11)$$

sont très valables, particulièrement si n n'est pas trop petit.

IV.6 - ESTIMATIONS DES VARIANCES

Elles sont calculées à partir des résultats du sondage pourvu que celui-ci soit bien réalisé conformément aux principes envisagés ci-dessus.

La variance de x dans la base sera estimée à partir de l'échantillon par

1/ $$(\hat{\sigma}_x)^2 = \frac{N - 1}{N} \frac{\sum_1^n (x_i - \bar{x})^2}{n - 1} \sim \frac{\sum (x_i - \bar{x})^2}{n - 1} \qquad (12)$$

$$2/ \qquad (\hat{\sigma}_x)^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n - 1}$$

Pour l'estimation de la variance de la moyenne, on aura :

$$1/ \qquad (\hat{\sigma}_{\bar{x}})^2 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n - 1} \sum_1^n (x_i - \bar{x})^2$$

$$2/ \qquad (\hat{\sigma}_{\bar{x}})^2 = \frac{1}{n(n - 1)} \sum_1^n (x_i - \bar{x})^2 \qquad (13)$$

et pour les covariances :

$$1/ \qquad \hat{\sigma}_{\bar{x}\bar{y}} = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n - 1} \sum_1^n (x_i - \bar{x})(y_i - \bar{y})$$

$$2/ \qquad \hat{\sigma}_{\bar{x}\bar{y}} = \frac{1}{n(n - 1)} \sum_1^n (x_i - \bar{x})(y_i - \bar{y}) \qquad (14)$$

Ces formules sont sans biais avec n - 1 degrés de liberté, quelle que soit la distribution des a_i ou des b_i dans la base.

Mais $\hat{\sigma}_{\bar{x}}$ est pour de petits échantillons, une estimation légèrement biaisée de l'erreur-type $\sigma_{\bar{x}}$.

Si n est grand, disons 25 ou 30 ou davantage, la distribution de

$$t = \frac{\bar{x} - E\bar{x}}{\hat{\sigma}_{\bar{x}}}$$

sera en pratique, dans la plupart des cas, très voisine d'une distribution normale. La distribution de Student, qui dépend du nombre n - 1 degrés de liberté, tient compte de la non-normalité de t pour des échantillons issus d'une population normale.

Une estimation rapide de l'erreur-type de $\bar{x}$ peut être calculée à l'aide de la formule

$$\hat{\sigma}_{\bar{x}} \sim \frac{\bar{w}}{g}, \quad (g \leqslant 10),$$

où $\bar{w}$ est la moyenne des étendues (écarts entre valeurs extrêmes) des groupes de g valeurs successives de l'échantillon. La valeur optimale de g pour ce calcul est g = 8 pour une base normalement distribuée, mais la formule est remarquablement précise jusqu'à g = 10, à la fois pour une distribution normale ou une distribution rectangulaire (Grubbs and Weaver, 1947, Nathan Mantel, 1951).

Une estimation utile de la variance de $f = \bar{x}/\bar{y}$ est donnée par les formules :

$$1/ \qquad (\hat{\sigma}_f)^2 \sim \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{(n - 1)\bar{x}^2} \sum_1^n (x_i - fy_i)^2$$

2/ $$(\hat{\sigma}_f)^2 \sim \frac{1}{n(n-1)\bar{x}^2} \sum_1^n (x_i - fy_i)^2$$

obtenues en substituant dans les équations (11) les estimations données par les équations (13) et (14).

Nous pouvons maintenant comparer les variances des estimations X et X' (équations (2) et (6)). On voit que X' sera la plus précise si le coefficient de corrélation ρ entre les x_i et les y_i est assez élevé.

De manière plus précise, on aura :

$$C_{x'} < C_x$$

si

$$\rho > \frac{1}{2} \; C_b / C_a$$

IV.7 - IMPORTANCE PEU SENSIBLE DE LA TAILLE DE LA BASE

Le rôle joué par N dans les formules ci-dessus montre que l'effectif de la base a peu d'importance sur la taille de l'échantillon nécessaire pour obtenir une précision donnée, à moins que l'échantillon représente 20 % ou plus de la base (dans le cas des tirages sans remise).

L'échantillon nécessaire pour estimer la proportion de grains de sable noirs ou blancs dans un boisseau de sable serait le même que celui qui serait nécessaire pour estimer cette proportion dans le chargement d'un camion.

La seule condition - mais elle est importante - pour qu'il en soit ainsi est que le boisseau ou le camion de sable soient parfaitement mélangés.

IV.8 - ESTIMATION D'UNE PROPORTION

Dans de nombreuses enquêtes, l'examen d'une unité d'échantillonnage donne lieu à l'une ou l'autre de deux réponses : oui ou non, pile ou face, dans une inspection par calibres, une pièce passe ou ne passe pas, une personne appartient ou non à un certain groupe d'âges.

Ce cas particulier se rattache aisément au cas général : pour chaque unité dans la base on pourra écrire $a_i = 1$ pour oui, ou $a_i = 0$ pour non.

Dans ce cas A correspond à l'effectif total des "oui" dans la base et a sera la proportion des "oui" dans cette base, proportion généralement désignée par p.

On a alors :

$$\sigma^2 = pq, \qquad (15)$$

avec $p + q = 1$.

La variable aléatoire x_i prendra l'une des valeurs 0 ou 1,

$$r = \sum_{1}^{n} x_i \tag{16}$$

sera le nombre des "oui" dans l'échantillon et

$$\hat{p} = \frac{r}{n} \tag{17}$$

sera la proportion des "oui" dans cet échantillon.

En remplaçant $\bar{x}$ par $\hat{p}$ dans les équations précédentes on voit que $\hat{p}$ est une estimation de p et que, suivant le mode le tirage (sans remise ou avec remise), on a :

$$\begin{aligned} &1/ \qquad \sigma^2_{\hat{p}} = \frac{N-n}{N-1}\,\frac{pq}{n} \sim \left(\frac{1}{n} - \frac{1}{N}\right) pq \\ &2/ \qquad \sigma^2_{\hat{p}} = \frac{pq}{n} \end{aligned} \tag{18}$$

La validité de ces formules exige que chaque unité d'échantillonnage ne puisse donner que l'une des valeurs 0 ou 1 (classement en deux catégories et deux seulement).

La variance de $\hat{p}$, sous les conditions ci-dessus, sera estimée par :

$$\begin{aligned} &1/ \qquad \sigma^2_{\hat{p}} = \left(1 - \frac{n}{N}\right) \frac{\hat{p}\hat{q}}{n-1} \sim \left(\frac{1}{n} - \frac{1}{N}\right) \hat{p}\hat{q} \\ &2/ \qquad \sigma^2_{\hat{p}} = \frac{\hat{p}\hat{q}}{n-1} \sim \frac{\hat{p}\hat{q}}{n} \end{aligned} \tag{19}$$

avec $\hat{p} + \hat{q} = 1$ (n étant généralement grand dans ces problèmes)

IV.9 - VARIANCE DE L'ESTIMATION D'UNE VARIANCE

Les variances ci-dessus sont estimées à partir des résultats de l'échantillon, ce sont donc aussi des variables aléatoires affectées d'une certaine variance.

Cette variance dépend de l'effectif de l'échantillon, mais aussi d'un certain paramètre γ_2, moment centré du quatrième ordre de la distribution des unités dans la base :

$$\gamma_2 = \frac{1}{N} \sum_{1}^{N} \left(\frac{a_i - \bar{a}}{\sigma}\right)^4 = E\left(\frac{a_i - \bar{a}}{\sigma}\right)^4 \tag{20}$$

La variance de $(\hat{\sigma}_x)^2$ dépend de n et γ_2, elle est donnée, dans le cas d'un tirage avec remise, par les formules

$$\left.\begin{aligned} V(\hat{\sigma}_x^2) &= \frac{2\sigma^4}{n-1}\left[1 + \frac{n-1}{2n}(\gamma_2 - 3)\right] \\ &\sim \sigma^4 \frac{\gamma_2 - 1}{n} \qquad \text{(pour n grand)} \\ &= \frac{2\sigma^4}{n-1} \qquad \text{(quel que soit n si } \gamma_2 = 3) \end{aligned}\right\} \tag{21}$$

(Cette dernière formule correspond au cas d'une distribution normale pour laquelle on a précisément $\gamma = 3$).

L'estimation de $\hat{\sigma}_x^2$ peut donc être entachée d'une erreur importante si γ_2 est grand.

Ainsi, par exemple, dans le cas d'une distribution binomiale de la variable x_i, dont l'espérance mathématique est p et la variance pq, et qui ne peut prendre que les valeurs 0 ou 1 avec les probabilités q et p on aura :

$$\gamma_2 = \frac{q(0-p)^4 + p(1-p)^4}{p^2q^2} = \frac{1}{pq} - 3 \tag{22}$$

Pour $p = q = 1/2$, la valeur minimale de γ_2 est $\gamma_2 = 1$, mais pour $p = 0,05$ on aurait $\gamma_2 \sim 18$.

IV.10 - SONDAGE SYSTEMATIQUE

Les variances calculées ci-dessus constituent la base essentielle de la théorie des sondages, bien que pratiquement on utilisera rarement le mode élémentaire de tirage envisagé ici.

On peut par exemple diviser la base en un certain nombre de zones et faire les tirages dans chaque zone pour tirer parti de la stratification naturelle qui peut exister dans la base.

Un moyen simple et souvent employé consiste à prendre dans la base dont les unités ont été numérotées de 1 à N, les n unités numérotées :

$$h \qquad h + k \;\ldots\ldots\ldots\; h + (n-1)\,k,$$

h étant pris au hasard parmi les k premiers nombres entiers, h et k satisfaisant à la condition

$$h + (n-1)\,k \leqslant N < h + nk$$

Cette procédure est appelée sondage systématique.

Un désavantage d'un tel tirage systématique est qu'il n'existe pas de moyen de calculer l'erreur type de l'estimation obtenue. On peut cependant calculer une valeur surestimée de $V(\bar{x})$ en prenant la moyenne des variances sur des paires consécutives d'unités d'échantillonnage.

La base est partagée en n/2 zones égales : chacune contenant 2N/n unités consécutives ; dans chaque zone on tire au hasard 2 unités.

Si 1, 2, ... 2 i - 1, 2i, ..., n, sont les rangs des unités de l'échantillon, classées dans l'ordre qu'elles occupaient dans la base on aura :

$$\bar{x} = \frac{2}{n} \sum_{i=1}^{n/2} \bar{x}_i \qquad \text{avec} \qquad \bar{x}_i = \frac{1}{2} (x_{2i-1} + x_{2i})$$

et

$$V(\bar{x}) = \frac{4}{n^2} \sum_{1}^{n/2} V(\bar{x}_i)$$

$V(\bar{x}_i)$ étant estimé conformément à la formule (13) par

$$V(\bar{x}_i) = \left(\frac{1}{2} - \frac{1}{2N/n}\right)\frac{1}{2} \left[(x_{2i-1} - \bar{x}_i)^2 + (x_{2i} - \bar{x}_i)^2\right]$$

d'où finalement :

$$V(\bar{x}) = \frac{1}{2n} \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{n/2} [(x_{2i} - x_{2i-1})^2] \tag{23}$$

Un autre désavantage du sondage systématique tel qu'il a été précédemment défini, est que, si des périodicités inconnues existent dans la base, la variance de l'estimation pourra être considérablement augmentée si la période est un sous-multiple de la raison de la progression arithmétique.

Par contre, si dans la base, les unités qui occupent des rangs voisins ont tendance à se ressembler, le sondage systématique sera plus efficace que le sondage élémentaire.

L'efficacité du sondage systématique est la même que celle du sondage élémentaire si le rangement des unités dans la base peut être considéré comme aléatoire.

Au lieu d'un unique point de départ entre 1 et N/n, on peut évidemment prendre au hasard deux points de départ entre 1 et 2N/n, et prendre ensuite chaque $\left(\frac{2N}{n}\right)^{\text{ème}}$ unité, l'extension à un nombre quelconque de points de départ étant évidente. L'utilisation de 2 ou plusieurs origines permettra alors une estimation de l'erreur-type de la moyenne.

Pour des études préliminaires des procédés plus rapides que le comptage pourront être utilisés : lignes équidistantes d'une liste, fiches équidistantes dans une pile, ..., mais de tels procédés qui risquent, dans leur application par le personnel qui en est chargé, de ne pas respecter le caractère probabiliste fondamental, sont beaucoup moins sûrs que le comptage.

IV.11 - MESURE QUANTITATIVE DE L'EFFICACITE D'UN PLAN DE SONDAGE

Par définition, les efficacités I_1 et I_2 de deux plans d'échantillonnage, pour une même taille n de l'échantillon sont proportionnelles aux inverses des variances dans les échantillons

$$\frac{I_1}{I_2} = \frac{\sigma_2^2}{\sigma_1^2} \qquad (\text{si } n_1 = n_2) \tag{24}$$

Si n_1 et n_2 sont quelconques, mais $\sigma_1 = \sigma_2$, on pourra écrire :

$$\frac{I_1}{I_2} = \frac{n_1}{n_2} \tag{25}$$

La comparaison des coûts est aussi un élément important : pour une même variance des estimations, Hansen a proposé de définir le rapport des efficacités par le rapport des coûts.

Le but d'un plan de sondage est d'obtenir une efficacité maximale pour un coût donné ou, ce qui revient au même, de construire un plan de sondage qui aura le coût minimal pour une erreur type acceptable, fixée à l'avance.

Dans la pratique des enquêtes, une difficulté se présente souvent, du fait que le plan de sondage qui est plus efficace qu'un autre pour une caractéristique donnée, peut être moins efficace pour une autre caractéristique.

V - LA STRATIFICATION DANS LES SONDAGES

A - PROBLEMES DE DENOMBREMENT ET D'ESTIMATION

V.1 - BUT DE LA STRATIFICATION

D'une manière générale, la stratification a pour but essentiel d'utiliser certaines informations contenues dans la base, autres que celles qui seront fournies par l'échantillon final ; il existe d'autres moyens d'utiliser de telles informations, par exemple, la méthode du quotient, la méthode de régression. Elles seront examinées ci-après.

La stratification peut être envisagée de différentes manières :

Une première façon de faire est de grouper les unités d'échantillonnage de la base en un certain nombre de classes (ou strates) dont les unités présentent quelque ressemblance de point de vue étudié ; on tire ensuite un échantillon de chacune des strates ainsi formées.

Une autre façon d'envisager la stratification est de tirer de la base de sondage un échantillon préliminaire sans stratification préalable, de classer ensuite en strates les unités fournies par cet échantillon préliminaire et enfin de tirer un échantillon de chacune de ces strates pour constituer l'échantillon final. Cette classification a posteriori sera moins coûteuse que la classification de l'ensemble à moins que celle-ci ne soit réalisée à priori pour des raisons physiques (par exemple : répartition géographique).

Dans tous les cas, les estimations finales pour la base de sondage seront des combinaisons des estimations faites dans chaque strate.

V.2 - NOTATIONS ET DEFINITIONS

Nous examinerons successivement quelques plans de sondage utilisant la stratification afin de les comparer au plan sans stratification (A) dont les formules ont été rappelées précédemment (cf. Ch. IV).

Rappelons que dans ce plan (A), l'échantillon fournit les estimations :

$$\bar{x} = \frac{1}{n}(x_1 + \ldots\ldots + {}_n) = \frac{x}{n} \qquad X = N\bar{x}$$

avec

$$V(\bar{x}) = \frac{N-n}{N-1}\,\frac{\sigma^2}{n} \sim \frac{1}{n} - \frac{1}{N}\,\sigma^2 \qquad \text{(tirage sans remise)}$$

ou

$$V(\bar{x}) = \frac{\sigma^2}{n} \qquad \text{(tirage avec remise),}$$

l'estimation de σ^2 à partir de l'échantillon étant, suivant le cas :

$$(\hat{\sigma}_x)^2 = \frac{N-1}{N}\,\frac{\sum_1^n (x_i - \bar{x})^2}{n-1} \sim \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

$$(\hat{\sigma}_x)^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

Notons ainsi que dans ce plan (A), comme dans les plans (B), (D), (F), (H), ci-après les unités d'échantillonnage de la base ont une égale probabilité d'être tirées, soit n/N et que $E\bar{x} = \bar{a}$, $EX = A$.

Les notations utilisées sont indiquées dans les tableaux (1) et (2) présentés pour le cas de 3 strates, l'extension à un nombre quelconque M de strates étant immédiate.

V.3 - LES PROPORTIONS P_i SONT CONNUES OU PEUVENT ETRE DETERMINEES

V.3.1 - Stratification a priori de la base

V.3.1.1 - *Plan (B) : Echantillon représentatif*

Dans chaque strate, l'échantillon est prélevé avec un taux de sondage uniforme :

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots = \frac{n}{N} \tag{8}$$

L'effectif total n de l'échantillon étant fixé, compte tenu de considérations de précision et de coût, on a :

$$n_i = n\frac{N_i}{N} = n\,P_i \tag{9}$$

TABLEAU 1

NOTATIONS ET DEFINITIONS POUR LA BASE (Cas de trois strates)

Strate numéro	Nombre d'unités d'échantillonnage		Proportion des unités dans la base	Population		Variance à l'intérieur des strates
	Base	Echantillon		Moyenne par unité dans la strate	Total dans la strate	
1	N_1	n_1	$P_1 = N_1/N$	$\bar{a}_1$	$A_1 = N_1\bar{a}_1$	σ_1^2
2	N_2	n_2	$P_2 = N_2/N$	$\bar{a}_2$	$A_2 = N_2\bar{a}$	σ_2^2
3	N_3	n_3	$P_3 = N_3/N$	$\bar{a}_3$	$A_3 = N_3\bar{a}$	σ_3^2
Total pour la base	N	n	1		A	
Moyenne simple par strate	$\bar{N} = N/M$	$\bar{n} = n/M$	$1/M$		$\bar{A} = A/M$	
Moyenne pondérée par unité				$\bar{a} = A/N$		σ_w^2

$$\bar{a} = P_1\bar{a}_1 + P_2\bar{a}_2 + P_3\bar{a}_3 = A/N \quad (1)$$

$$\sigma_w^2 = P_1\sigma_1^2 + P_2\sigma_2^2 + P_3\sigma_3^2 \quad (2)$$

$$\sigma_R^2 = Q_1\sigma_1^2 + Q_2\sigma_2^2 + Q_3\sigma_3^2 \quad (3)$$

$$\bar{\sigma}_w = P_1\sigma_1 + P_2\sigma_2 + P_3\sigma_3 \quad (4)$$

$$\bar{\sigma}_R = Q_1\sigma_1 + Q_2\sigma_2 + Q_3\sigma_3 \quad (5)$$

$$\sigma_b^2 = P_1(\bar{a}_1 - \bar{a})^2 + P_2(\bar{a}_2 - \bar{a})^2 + P_3(\bar{a}_3 - \bar{a})^2 = P_1\bar{a}_1^2 + P_2\bar{a}_2^2 + P_3\bar{a}_3^2 - \bar{a}^2 \quad (6)$$

$$\sigma^2 = \sigma_b^2 + \sigma_w^2 \quad (7)$$

TABLEAU 2

NOTATIONS ET DEFINITIONS POUR L'ECHANTILLON

Strate n°	Total par strate dans l'échantillon	Moyenne par unité de sondage	Total estimé dans la population	Variance de cette estimation
1	x_1	$\bar{x}_1 = x_1/n_1$	$X_1 = N_1x_1/n_1$	$V(X_1)$
2	x_2	$\bar{x}_2 = x_2/n_2$	$X_2 = N_2x_2/n_2$	$V(X_2)$
3	x_3	$\bar{x}_3 = x_3/n_3$	$X_3 = N_3x_3/n_3$	$V(X_3)$
Total	x		X	V(X)(1)

(1) Les variances ne sont additives que si les N_i (ou P_i) sont connus et utilisés dans l'estimation de X/.

Dans le plan (B) ci-après, sondage proportionnel, X se réduit à Nx/n.

Ayant tiré dans chaque strate, à l'aide d'une table de nombres au hasard, un échantillon de taille n_i qui donne un total x_i, on aura les estimations

$$\ldots\ldots\ldots \; X_i = N_i \frac{x_i}{n_i} = \frac{N}{n} x_i \; \ldots\ldots \tag{10}$$

d'où, pour la population

$$X = X_1 + X_2 + \ldots\ldots = \frac{N}{n}(x_1 + x_2 \ldots\ldots) = \frac{N}{n} x = N\bar{x} \tag{11}$$

Si les N_i sont connus, et si les tailles n_i des échantillons sont indépendantes, les variances dans les strates sont additives et on aura :

$$\begin{aligned} V(X) &= V(X_1) + V(X_2) + \ldots\ldots\ldots \\ &= N_1^2 \left(1 - \frac{n_1}{N_1}\right)\frac{\sigma_1^2}{n_1} + N_2^2\left(1 - \frac{n_2}{N_2}\right)\frac{\sigma_2^2}{n_2} + \ldots\ldots \\ &= \left(1 - \frac{n}{N}\right)\frac{N^2}{n}\left(P_1\sigma_1^2 + \ldots\ldots\ldots\right) \end{aligned}$$

d'où

$$V(\bar{x}) = \left(1 - \frac{n}{N}\right)\frac{\sigma_w^2}{n} \tag{12}$$

qui dans le cas d'un tirage avec remise, se réduit à

$$V(\bar{x}) = \frac{\sigma_w^2}{n} \tag{13}$$

V. 3. 1. 2 - *Plan (C) : Echantillon avec répartition optimale de Neyman*

Lorsque les variances σ_1^2, σ_2^2, ..., dans les strates sont nettement différentes, il y a intérêt, pour un échantillon d'effectif total n donné, à envisager une répartition n_1, n_2, ... qui conduise à une estimation finale de variance minimale.

Pour cela on déterminera n_1 ; n_2, ... par la condition :

$$\frac{n_1}{N_1\sigma_1} = \ldots\ldots\ldots = \frac{n_i}{N_i\sigma_i} = \ldots\ldots\ldots = \frac{n}{\sum N_i\sigma_i}$$

soit

$$n_i = n \frac{N_i\sigma_i}{\sum N_i\sigma_i} = n\, P_i\sigma_i/\bar{\sigma}_w \tag{14}$$

avec evidemment

$$n_1 + n_2 + \ldots\ldots = n$$

Dans chacune des strates, on déterminera comme ci-dessus les estimations X_1, X_2 .... d'où

$$X = X_1 + X_2 \ldots\ldots\ldots , \quad \bar{x} = \frac{X}{N}$$

Dans ce cas, on aura :

$$V(\bar{x}) = \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N} \qquad (15)$$

qui pour des tirages avec remise se réduit à :

$$V(\bar{x}) = \frac{(\bar{\sigma}_w)^2}{n}$$

En général, le terme correctif du second membre de l'équation (15) peut être négligé.

Une application rigoureuse des formules de Neyman est très généralement impossible, mais une approximation raisonnable des ordres de grandeur relatifs des σ_i^2 donnera d'excellents résultats.

Il arrive quelquefois que N et les N ne sont connus qu'approximativement avant le sondage, on devra alors se contenter d'approximations pour les n_i, les nombres N, N_i et P_i pouvant être définitivement fixés après le sondage.

V. 3.2 - Stratification a prosteriori (après tirage d'un échantillon priliminaire)

Si la base contient des milliers d'unités de sondage, le coût de classification de chaque unité dans une strate appropriée peut dépasser les économies à attendre des plans B ou C. La dépense et le temps nécessaire seront spécialement importants si l'information permettant le classement doit être recherchée par un essai ou un interview portant sur chaque unité.

Il est heureusement possible, dans de nombreux cas, d'utiliser des plans de sondage stratifié qui exigent seulement le classement des unités d'un échantillon préliminaire. Savoir si c'est le procédé le plus efficace dans tel ou tel cas particulier dépend des coûts comparatifs et des variances présumées.

Dans la stratification a posteriori, on tirera d'abord de la base un échantillon préliminaire sur lequel on effectuera la stratification, ensuite de chaque strate de cette sous-population on tirera un échantillon pour constituer avec leur ensemble l'échantillon final.

Pour l'instant, nous continuerons à admettre, dans les plans (D), (E), (F), (G) ci-après que les proportions P_i des différentes classes dans la population initiale (base) sont connues.

V. 3.2.2 - *Plan (D)*

L'effectif total n de l'échantillon final étant fixé, on tirera de la base, comme dans le plan (A), un échantillon de taille n.

Ces unités seront réparties en strates : les nombres n_1, n_2, ... d'unités de sondage dans les diverses strates seront alors des variables aléatoires.

On examinera chacune de ces unités et on formera les estimations X_1, X_2, ... X et $\bar{x}$ comme dans le plan (B).

La variance de $\bar{x}$ est alors :

$$V(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left(\sigma_w^2 + \frac{1}{n}\sigma_R^2\right)$$

$$\sim \frac{1}{n}\left(\sigma_2^2 + \frac{1}{n}\,\sigma_R^2\right) \tag{16}$$

Si le terme $\sigma_R^2 = Q_1\sigma_1^2 + Q_2\sigma_2^2 + \ldots\ldots$ est petit, la variance du plan (D) sera seulement légèrement supérieure à celle du plan (B). Cependant le terme σ_R^2/n peut être très important lorsque la strate des grandes unités est ouverte vers le haut, auquel cas le plan (D) serait un mauvais choix.

V.3.2.2 - *Plan (E)*

L'effectif total n de l'échantillon final étant fixé ainsi que l'effectif n' de l'échantillon préliminaire on tirera ce dernier comme dans le plan (A).

Si n_1', n_2' .... sont les effectifs (aléatoires) résultant de l'affectation des n' unités de l'échantillon préliminaire dans les strates, on déterminera les effectifs n_1, n_2 ..... des échantillons composant l'échantillon final de manière que :

$$n_1 + \ldots\ldots + n_i + \ldots\ldots = n$$

avec :

$$\frac{n_1}{n_1'\sigma_1} = \ldots\ldots\ldots\ldots = \frac{n_i}{n_i'\sigma_i} = \ldots\ldots = \frac{n}{\sum_i n_i'\sigma_i} \tag{17}$$

d'où

$$n_i = n\,\frac{n_i'\sigma_i}{\sum_i n_i'\sigma_i}$$

On prélevera au hasard n_i unités parmi les n_i' de manière à constituer l'échantillon final.

On aura intérêt à choisir n' de manière à ce que la classe correspondant à la plus grande variance soit utilisée sans réduction d'effectif.

On examinera chaque unité de l'échantillon final et on formera encore les estimations X_1, X_2 .....

$$X = X_1 + X_2 + \ldots\ldots$$

$$\bar{x} = X/N$$

On aura cette fois :

$$\begin{aligned} V(\bar{x}) &= \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N} + \frac{1}{n}\left(\frac{1}{n'} - \frac{1}{N}\right)\bar{\sigma}_w\,\bar{\sigma}_R \\ &\sim \frac{1}{n}\left[(\bar{\sigma}_w)^2 + \frac{1}{n'}\,\bar{\sigma}_w\,\bar{\sigma}_R\right], \end{aligned} \tag{18}$$

si N est grand par rapport à n'.

Nous allons examiner maintenant deux plans (F) et (G), dans lesquels les effectifs n_i' sont réalisés progressivement avec une importante économie.

V.3.2.2 - *Plan (F)*

Les effectifs n_i des échantillons partiels étant fixés comme dans le plan (B), on tirera une à une, à l'aide des nombres au hasard, des unités de la base de sondage et on les classera dans leurs strates respectives, on continuera jusqu'à ce que les quotas n_i soient satisfaits.

On calculera X et $V(\bar{x})$ comme dans le plan (B).

V. 3.2.4 - *Plan (G)*

La technique est la même que dans le plan (F) sauf que les tailles n_i sont déterminées comme dans le plan (C). La variance dans le plan (G) est identique à la variance du plan (C).

V. 4 - LES PROPORTIONS P_i NE SONT PAS CONNUES

Lorsque les proportions $P_i = N_i/N$ dans la base ne sont pas connues, on peut les estimer à l'aide d'un échantillon préliminaire de taille $N' > n$, n étant la taille totale de l'échantillon final.

L'échantillon préliminaire N' servira de nouvelle base, ceci étant les plans (H) et (I) ci-après ressemblent respectivement aux plans (B) et (C).

V.4.1 - Plan (H)

La taille finale n étant fixée, on déterminera d'abord la taille optimale N' de l'échantillon préliminaire par :

$$\frac{n}{N'} = \frac{\sigma_w}{\sigma_b} \sqrt{\frac{c_1}{c_2}} \tag{19}$$

dans laquelle c_1 est le coût moyen de classification d'une unité d'échantillonnage dans l'échantillon préliminaire et c_2 le coût moyen d'enquête par unité de l'échantillon final.

On tirera d'abord, comme dans le plan (A), un échantillon préliminaire de taille N' que l'on répartira en strates de tailles N'_1, N'_2,

Les rapports $\hat{P}_i = N'_i/N'$ donneront une estimation des proportions P_i d'unités d'échantillonnage appartenant aux strates i dans la base.

On réduira ensuite proportionnellement les strates de l'échantillon préliminaire de manière à ramener son effectif à la taille finale n.

Si x_i est le total observé pour les n_i unités de la $i^{ème}$ strate de l'échantillon final, on calculera :

$$\bar{x}_i = \frac{x_i}{n_i}$$

et ensuite

$$\bar{x} = \frac{X}{N} = \frac{1}{N} \sum N\hat{P}_i \bar{x}_i = \sum \frac{N'_i}{N'} \frac{x_i}{n_i}$$

$$= \frac{N'}{n} \frac{1}{N'} \sum x_i = \frac{1}{n} \sum x_i = \frac{x}{n}$$

Dans ce cas on aura :

$$V(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) \sigma_w^2 + \frac{1}{n}\left(\frac{1}{N'} - \frac{1}{N}\right) \sigma_R^2 + \left(\frac{1}{N'} - \frac{1}{N}\right) \sigma_b^2$$

$$\sim \frac{\sigma_w^2}{n} + \frac{\sigma_b^2}{N'} \qquad (20)$$

Si N est grand, comparé à N' et à n.

La variance du plan (H) est supérieure à celle du plan (B) d'environ $\frac{\sigma_b^2}{N'}$. Si les moyennes des strates sont nettement différentes, l'équation (19) conduira à un échantillon préliminaire N' important ce qui diminuera le terme $\frac{\sigma_b^2}{N'}$ de l'équation (20).

V.4.2 - Plan (I)

L'effectif n de l'échantillon final étant fixé, on calculera d'abord la taille optimale N' de l'échantillon préliminaire à l'aide de l'équation :

$$\frac{n}{N'} = \frac{\bar{\sigma}_w}{\sigma_b} \sqrt{\frac{c_1}{c_2}} \qquad (21)$$

On tirera d'abord de la base, comme dans le plan (A), un échantillon préliminaire de taille N' dont on répartira les unités en strates comme dans le plan (H).

On réduira ces strates de manière à satisfaire à la relation :

$$\frac{n_1/N_1'}{\sigma_1} = \frac{n_2/N_2'}{\sigma_2} = \ldots\ldots \qquad (22)$$

de manière à atteindre l'effectif final fixé n.

A partir de cet échantillon final total on estimera les proportions P_i de l'échantillon préliminaire comme dans le plan (H) et on calculera $\bar{x}_i$ comme dans ce dernier plan.

On obtiendra finalement l'estimation :

$$\bar{x} = \sum \hat{P}_i \bar{x}_i = \frac{1}{N'} \sum \frac{N_i'}{n_i} x_i \qquad (23)$$

avec

$$V(\bar{x}) = \frac{|\sigma_w|^2}{n} - \frac{\sigma_w^2}{N} + \frac{1}{n}\left(\frac{1}{N'} - \frac{1}{N}\right) \bar{\sigma}_w \bar{\sigma}_R + \left(\frac{1}{N'} - \frac{1}{N}\right) \sigma_b^2$$

$$\sim \frac{(\bar{\sigma}_w)^2}{n} + \frac{\sigma_b^2}{N'}, \qquad (24)$$

si N est grand comparé à N' et à n.

La variance du plan (I) est supérieur à celle du plan (C) d'environ $\frac{\sigma_b^2}{N'}$.

On notera que dans les plans (H) et (I) l'échantillon préliminaire donne des estimations des P_i, tandis que l'enquête sur l'échantillon final donne des estimations $\bar{x}_i$ des $\bar{a}_i$.

On peut employer les plans (F) ou (G) en liaison avec les plans (H) ou (I) de manière à recueillir le bénéfice de strates nombreuses sans classer effectivement l'échantillon préliminaire complètement.

On fixe les tailles finales n_i pour un nombre quelconque de strates (définies par exemple par l'âge, le sexe, le revenu, le nombre de personnes par famille) en accord avec les proportions du recensement, en tenant compte si on le désire de la répartition optimale de Neyman. Ensuite on tire les unités de l'échantillon préliminaire de manière à atteindre les quotas n_i.

Cette méthode a été utilisée par Koller (Statistiches Bundensant, Wiesbaden).

V.5 - BENEFICE DE L'ECHANTILLONNAGE STRATIFIE

Si on désigne par A, B, C les variances de l'estimation de $\bar{a}$ par les plans (A), (B), (C) on a, à partir de la formule générale

$$V(\bar{x}) = \sum P_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\sigma_i^2}{n_i}$$

qui se réduit à

$$V(\bar{x}) \sim \sum P_i^2 \frac{\sigma_i^2}{n_i}$$

si N est grand.z

(A) $A = V(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$ pas de stratification

(B) $B = V(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{\sigma_w^2}{n}$ répartition proportionnelle

(C) $C = V(\bar{x}) = \frac{(\bar{\sigma}_w)^2}{n} - \frac{\sigma_w^2}{N}$ répartition de Neyman

Si n est le même pour les trois plans, on a :

$$\frac{A - B}{A} = 1 - \left(\frac{\sigma_w}{\sigma}\right)^2 = \left(\frac{\quad}{\quad}\right)^2$$

$$B = A \left(\frac{\sigma_w}{\sigma}\right)^2 \tag{25}$$

$$\frac{B - C}{B} = 1 - \left(\frac{\bar{\sigma}_w}{\sigma_w}\right)^2 = \frac{C_{\sigma_i}^2}{1 + C_{\sigma_i}^2} \tag{26}$$

$$C = B \left(\frac{\bar{\sigma}_w}{\sigma_w}\right)^2 = B \frac{1}{1 + C^2_{\sigma_i}} \qquad (27)$$

$$= A \left(\frac{\bar{\sigma}_w}{\sigma}\right)^2$$

en désignant par

$$C^2_{\sigma_i} = \left(\frac{1}{\bar{\sigma}_w}\right)^2 \sum P_i \ (\sigma_i - \bar{\sigma}_w)^2 \qquad (28)$$

la variance relative des σ_i .

Remarque

On peut montrer que pour deux strates on a :

$$\frac{A - B}{A} = P_1 P_2 \left(\frac{\bar{a}_2 - \bar{a}_1}{\sigma}\right)^2$$

$$\frac{B - C}{B} = P_1 P_2 \left(\frac{\sigma_2 - \sigma_1}{\sigma_w}\right)^2$$

et plus généralement :

$$\frac{A - B}{A} = \sum_{j>i} P_i P_j \left(\frac{\bar{a}_i - \bar{a}_j}{\sigma}\right)^2$$

$$\frac{B - C}{B} = \sum_{j>i} P_i P_j \left(\frac{\sigma_j - \sigma_i}{\sigma_w}\right)^2$$

Le bénéfice du plan (C) sur le plan (B) provient de la variance relative des σ_i, on voit ainsi pourquoi l'échantillonnage de Neyman augmente souvent le bénéfice dans l'échantillonnage de distributions très dissymétriques (comme dans le cas de revenus, nombre d'employés, production industrielle) et pourquoi, dans certains cas, il ne procure pas ces mêmes avantages (par exemple dans les enquêtes d'opinion où chaque personne compte pour 0 ou 1).

Les gains de précision des plans (F) et (G) sur le plan (A) sont les mêmes que pour (B) et (C) respectivement. Les gains des plans (D) et (E) sont moindres en raison du fait que les tailles n_i sont alors des variables aléatoires.

Des équations ci-dessus il résulte que le bénéfice à escompter de la formation d'une nouvelle strate (i) ne sera pas important à moins que la proportion P_i soit importante ou encore que σ_i ou a_i soient largement différents de la valeur moyenne.

Exemple[1]. Supposons que l'on veuille faire un sondage dans une région pour estimer le nombre total des lecteurs d'un certain journal. La liste

---

(1) Deming. "Sample Designs in Business Research". p. 299

d'adresses des abonnés du journal et le nombre d'exemplaires vendus par les marchands permettent de partager la région en 2 strates de telle sorte que, par exemple :

$p_1 = 0,10$ proportion des lecteurs dans la strate 1

$p_2 = 0,03$ " " " " 2

$P_1 = 0,5$ proportion des unités d'échantillonnage dans la strate 1

$P_2 = 0,5$ " " " " " 2

Les unités d'échantillonnage dans les deux strates étant par exemple des régions élémentaires contenant environ 30 personnes en âge d'être des lecteurs éventuels : il y a environ 3 lecteurs par région élémentaire dans la strate n° 1 et un lecteur dans la strate n°2.

Si l'on admet que les nombres de lecteurs par région élémentaire suivent une distribution de Poisson, on aurait alors

$$\sigma_1^2 = 3, \qquad \sigma_2^2 = 1$$

La variation des tailles réelles des régions élémentaires augmente ces variances : pour plus de surêté on prendra :

$$\sigma_1^2 = 5, \qquad \sigma_2^2 = 2.$$

On a alors

$$\bar{\sigma}_w = P_1\sigma_1 + P_2\sigma_2 = \frac{1}{2}(\sqrt{5} + \sqrt{2}) = 1,83$$

$$(\bar{\sigma}_w)^2 = 3,35$$

$$\sigma_w^2 = P_1\sigma_1^2 + P_2\sigma_2^2 = \frac{1}{2}(5 + 2) = 3,5$$

$p = P_1p_1 + P_2p_2 = 0,065$, proportion des lecteurs dans l'ensemble

$$\sigma_b^2 = P_1(p_1 - p)^2 + P_2(p_2 - p)^2 = 0,0012$$

$$\sigma^2 = \sigma_w^2 + \sigma_b^2 = 3,501$$

La comparaison des variances pour les trois plans A, B, C donne :

$$B = A\left(\frac{\sigma_w}{\sigma}\right)^2 \sim A$$

Dans ce cas, le sondage élémentaire (A) donne pratiquement la même précision que la stratification avec taux de sondage uniforme.

Mais on a aussi :

$$C = B\left(\frac{\bar{\sigma}_w}{\sigma_w}\right)^2 = B\,\frac{3,35}{3,50} = 0,96\ B,$$

c'est-à-dire que dans le cas d'un sondage avec répartition optimale de Neyman, 96 enquêtés donneraient à peu près la même précision que 100 enquêtés par sondage élémentaire direct ou par sondage avec taux de sondage uniforme.

Avec les mêmes données, sauf cette fois :

$$\sigma_1^2 = 6, \qquad \sigma_2^2 = 1$$

on obtiendrait :

$$C = 0,85\ B$$

Les sondages en vue de déterminer un rapport suivent sensiblement la même théorie, mais avec d'importantes extensions, particulièrement dans le cas d'estimations calculées à l'aide de totaux marginaux dans les classes, par exemple pour les groupes d'âges basés sur un recensement (cf Hansen, Hurwitz et Madow, Vol II ch. 7).

B - MODIFICATIONS A ENVISAGER POUR TENIR COMPTE D'UNE VARIABILITE DES COUTS, ENTRE STRATES

Il peut, dans certains cas, être sage de diminuer la probabilité de sélection dans les strates où le coût d'enquête par unité devient excessif et de baser davantage le sondage sur des strates dans lesquelles le coût est raisonnable.

Les théories qui permettront de modifier les plans (B) et (C) (et les autres plans basés essentiellement sur la répartition proportionnelle ou celle de Neyman) reposent sur les conditions :

$$\frac{n_i}{n_j} = \frac{P_i}{P_j} \sqrt{\frac{c_j}{c}} \longrightarrow \text{Plan (B')} \tag{29}$$

ou

$$\frac{n_i}{n_j} = \frac{P_i\,\sigma_i}{P_j\,\sigma_j} \sqrt{\frac{c_j}{c_i}} \longrightarrow \text{Plan (C')} \tag{20}$$

dans lesquelles c_i et c_j sont les coûts respectifs unitaires d'enquête dans les strates i et j.

Comme les coûts entrent dans ces équations par leurs racines carrées, de telles modifications ne seront réellement efficaces que si c_i et c_j sont très différents.

Dans chaque cas particulier, la décision impliquera des calculs qui devront tenir compte de valeurs vraisemblables des rapports c_1/c_2 et σ_1/σ_2 .

L'exemple numérique(1) ci-après permet de comparer les variances des plans (B), (B'), (C), (C') dans l'hypothèse où l'on aurait :

$$c_1 = 5, \quad c_2 = 10, \quad P_1 = 0,60, \quad P_2 = 0,40, \quad \frac{\sigma_1}{\sigma_2} = 2,$$

la dépense totale pour l'enquête sur le terrain étant fixée à 2 000.

---

(1) Deming "Sampling Designs in Business Research" p. 304

Dans tous les cas on a :

$$V(\bar{x}) = \sum P_i^2 \frac{\sigma_i^2}{n_i} - \sum \frac{1}{N_i} P_i^2 \sigma_i^2$$

En négligeant le terme $\sum 1/N_i \, P_i^2 \sigma_i^2$ qui figure dans les quatre expressions de la variance, on aura les résultats suivants qui permettent de comparer les variances :

1/ <u>Plan (B)</u>

$$\frac{n_1}{n_2} = \frac{P_1}{P_2} = \frac{3}{2}$$

$$n_2 = \frac{2}{3} n_1$$

$$n_1 c_1 + n_2 c_2 = n_1 \left(5 + \frac{20}{3}\right) = 2\,000$$

$$n_1 = 174$$

$$n_2 = 114$$

$$n = n_1 + n_2 = 285$$

$$V(\bar{x}) = \frac{\sigma_w^2}{n} = \frac{P_1 \sigma_1^2 + P_2 \sigma_2^2}{n} = 0{,}00563 \, \sigma_2^2$$

2/ <u>Plan (B')</u>

$$\frac{n_1}{n_2} = \frac{6}{4}\sqrt{2} = 2{,}12$$

$$n_1 c_1 + n_2 c_2 = (10{,}6 + 10)\, n_2 = 2\,000$$

$$n_2 = 97$$

$$n_1 = 206$$

$$n = 303$$

$$V(\bar{x}) = P_1^2 \frac{\sigma_1^2}{n_1} + P_2^2 \frac{\sigma_2^2}{n_2} = 0{,}00515 \, \sigma_2^2$$

3/ <u>Plan (C)</u>

$$\frac{n_1}{n_2} = \frac{P_1 \sigma_1}{P_2 \sigma_2} = 2{,}12$$

d'où

$$n_1 = 206 \qquad n_2 = 97 \qquad n = 303$$

comme pour le plan (B'), de même que pour la variance que l'on peut calculer à partir de :

$$V(\bar{x}) = \frac{(\bar{\sigma}_w)^2}{n} = \frac{(P_1\sigma_1 + P_2\sigma_2)^2}{n} = \frac{(0,85 + 0,4)^2}{n} = 0,00515\ \sigma_2^2$$

4/ Plan (C')

$$\frac{n_1}{n_2} = \frac{P_1\sigma_1}{P_2\sigma_2}\sqrt{\frac{c_2}{c_1}} = 3$$

$$n_1c_1 + n_2c_2 = n_1\left(5 + \frac{10}{5}\right) = 2\,000$$

$$n_1 = 240 \qquad n_2 = 80 \qquad n = 320$$

$$V(\bar{x}) = P_1^2\frac{\sigma_1^2}{n_1} + P_2^2\frac{\sigma_2^2}{n_2} = 0,00500\ \sigma_2^2$$

Les quatres résultats sont pratiquement équivalents mais, dans les mêmes conditions sauf cette fois $c_1 = 5$, $c_2 = 45$, on obtiendrait les résultats :

Plan	$V(\bar{x})$
(B)	$0,0168\ \sigma_2^2$
(B')	$0,0107\ \sigma_2^2$
(C)	$0,0138\ \sigma_2^2$
(C')	$0,0105\ \sigma_2^2$

C - PROBLEMES POSES PAR L'APPLICATION DE CES METHODES

V.6 - HYPOTHESES NECESSAIRES

Le choix de l'effectif final, compte tenu de la précision caractérisée par $V(\bar{x})$, que l'on souhaite obtenir, exige que l'on puisse, au moins approximativement, faire les calculs que l'on rencontre dans les équations liant $V(\bar{x})$ à n de manière que les plans de sondage soient complètement établis avant que l'enquête commence.

Ceci exige que le statisticien puisse avoir à l'avance quelque information valable relativement à certains paramètres figurant dans ces équations soit σ, σ_w, $\bar{\sigma}_w$, σ_b.

De bonnes approximations des rapports σ_w/σ, $\bar{\sigma}_w/\sigma$, $\bar{\sigma}_w/\sigma_b$, c_1/c_2 (ou plus généralement c_i/c_j) permettront une bonne répartition, mais de mauvaises approximations pourraient conduire à des erreurs importantes.

Ces approximations pourront être fournies par des expériences précédentes ou par l'avis d'un expert connaissant bien le domaine technique

que l'on veut étudier. On peut, dans certains cas, connaître assez bien la forme vraisemblable des distributions, ou l'étendue vraisemblable des valeurs numériques qui existent dans les diverses strates et, à partir de là, estimer de façon suffisamment précise, pour le but envisagé, les variances que l'on pourra utiliser : les relations entre étendue et écart-type permettront d'estimer celles-ci (distribution normale ou distribution rectangulaire).

D'ailleurs, dans beaucoup de cas la forme de la distribution n'a pas une grosse influence sur cette relation. Le cas d'une strate dans laquelle la variable peut atteindre de très grandes valeurs peut être étudié à part, par exemple par une enquête à 100 %.

Enfin, il faut tenir compte du fait que vouloir obtenir $\bar{x}$ avec une variance fixée de manière très précise à l'avance est un problème qui ne se pôse presque jamais. Dans tous les cas la précision réelle obtenue dans un sondage probabiliste pourra toujours être estimée à partir des résultats du sondage lorsque celui-ci sera effectué.

V.7 - ECHANTILLONNAGE PROGRESSIF

Il est quelquefois possible, lorsque la décision relative à la taille de l'échantillon est difficile, de procéder de la manière suivante :

- choisir a priori une taille d'échantillon qui soit presque certainement assez grande,
- partager, à l'aide des nombres au hasard, cet échantillon en deux portions 1 et 2,
- faire porter d'abord l'enquête sur la portion 1 seulement, lorsqu'elle est presque terminée, calculer la variance de quelques unes de ses principales caractéristiques et décider si oui ou non, cette première partie semble donner la précision souhaitée. S'il en est ainsi on achévera l'enquête de la première portion et le sondage sera terminé, sinon on fera porter l'enquête sur la seconde portion et on estimera la variance à l'aide de l'ensemble des résultats.

Cette méthode est quelquefois appelée sondage progressif.

Elle n'est utilisable que si l'enquête est réalisée par un petit nombre d'experts qui restent sur le terrain aussi longtemps qu'il est nécessaire : elle ne peut être employée si on doit à l'avance recruter un personnel d'enquête pour une durée définie à l'avance.

D - ETUDES ANALYTIQUES

Alors que dans ce qui précède (problèmes de dénombrement), il s'agissait de répartir l'échantillon de manière à obtenir la meilleure précision pour un total X, une moyenne $\bar{x}$ (ou une fréquence f) nous envisageons maintenant des études dont le but est de découvrir les causes ou sources de variations, de savoir si un traitement ou une condition d'environnement produit un effet et, s'il en est ainsi, d'en mesurer l'importance.

La formule générale de la variance de la différence entre deux moyennes $\bar{x}$ et $\bar{y}$ déduites d'échantillons indépendants d'effectif n_x et n_y tirés au hasard sans stratification des deux groupes (A) et (B) est

$$V(\bar{x} - \bar{y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} \tag{31}$$

Pour de telles études la répartition optimale compte tenu des coûts est définie par :

$$\frac{n_x}{n_y} = \frac{\sigma_x}{\sigma_y} \sqrt{\frac{c_B}{c_A}} \tag{32}$$

où c_A, c_B sont les coûts unitaires respectifs des deux enquêtes.

On peut noter que les effectifs des groupes (A) et (B) n'interviennent pas dans cette formule et que cette répartition optimale est différente de celle envisagée dans les problèmes précédents.

En poursuivant une étude comportant simultanément les deux caractères (dénombrement et comparaison), il peut arriver que l'échantillon optimal pour l'un des problèmes ne le sera pas pour l'autre : un compromis sera généralement nécessaire dans une étude qui se propose ces deux objectifs.

Si les écarts-types σ_x, σ_y ne sont pas très différents, ni les coûts c_A, c_B, la formule (32) se réduit à :

$$n_x = n_y,$$

mêmes nombres d'observations dans chacun des deux groupes, quel qu'en soit l'effectif total.

Pour :

$$n_x = n_y = \frac{n}{2}, \qquad \sigma_x = \sigma_y = \sigma,$$

on aura :

$$V(\bar{x} - \bar{y}) = \frac{4\sigma^2}{n} \tag{33}$$

ce qui pour une précision fixée permettra de déterminera n.

Evidemment si on dispose, sans avoir à organiser l'enquête, des résultats correspondant à des effectifs n_x et n_y, on les utilisera intégralement.

S'il s'agit d'estimer la différence constatée entre deux proportions p_x et p_y et si le sondage est binomial dans les deux strates, donnant lieu aux estimations $\hat{p}_x$ et $\hat{p}_y$, on pourra utiliser :

$$V(\hat{p}_x - \hat{p}_y) = \frac{\hat{p}_x\hat{q}_x}{n_x} + \frac{\hat{p}_y\hat{q}_y}{n_y}$$

Dans la plupart des cas, où l'enquête est organisée pour décider si p_x et p_y sont vraiment différents dans les deux populations, on pourra écrire :

$$V(\hat{p}_x - \hat{p}_y) = \hat{p}\hat{q}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)$$

avec

$$\hat{p} = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}, \qquad \hat{q} = 1 - \hat{p}$$

VI - ESTIMATIONS PAR LA REGRESSION

VI.1 - BUT DE CES METHODES

Nous avons déjà vu dans les sondages stratifiés l'emploi d'informations supplémentaires fournies par la base ; même dans le sondage élémentaire l'estimation $X = N\bar{x}$, du total A dans la base, utilise N, nombre des unités d'échantillonnage dans la base.

Nous allons maintenant envisager les moyens d'utiliser d'autres informations supplémentaires, nous introduirons une forme générale qui comprend l'estimation directe de $\bar{x}$ et l'estimation par le quotient comme cas particuliers, mais qui conduit à d'autres estimations utiles.

Ces estimations, comme l'estimation du rapport X/Y, exigent que l'on puisse obtenir des informations valables et indépendantes sur une seconde variable (Y).

Il est important de noter que toutes ces estimations qui procèdent d'informations supplémentaires peuvent présenter d'importants avantages sur la simple estimation $\bar{x}$, si le coefficient de corrélation ρ entre les variables x et y est élevé, mais cette condition n'est pas toujours suffisante.

Nous envisagerons les estimations résumées dans le tableau ci-après et qui sont toutes de la forme générale

$$\bar{x}_i = \bar{x} + m_i\,(\bar{b} - \bar{y}) \qquad (1)$$

dans laquelle $\bar{b} = E\bar{y}$ est la valeur moyenne de y par unité d'échantillonnage connue indépendamment par d'autres sources.

Nous considérerons ici 4 cas (i = 1, 2, 3, 4), parmi divers cas possibles largement étudiés par Hansen, Hurwitz et Madow (Wiley 1953).

Ultérieurement nous tiendrons compte des circonstances dans lesquelles $\bar{b}$ a été estimé à partir d'un échantillon plus important ou d'un échantillon indépendant.

L'estimation $\bar{x}_1 = \bar{x}$, sans biais, est celle qui a déjà été envisagée ci-dessus : elle figure ici pour comparaison, sa variance $\sigma^2_{\bar{x}}$ apparaissant dans les autres variances.

L'estimation $\bar{x}_3$ par la méthode des moindres carrés sera nettement supérieure à $\bar{x}$ pour de grands échantillons lorsque la corrélation entre x et y sera élevée. Dans certains cas, par exemple lorsque le coefficient de régression $\beta = \rho\,\frac{\sigma_x}{\sigma_y}$ et le rapport $\bar{x}/\bar{y}$ sont nettement différents, elle sera aussi nettement meilleure que l'estimation par le quotient.

Cas n°	m_i	Equations	Remarques
1	0	$\bar{x}_1 = \bar{x}$	Ce choix de m_i conduit à la valeur $\bar{x}$ précédemment obtenue et ne fait pas usage d'information supplémentaire
2	m_2	$\bar{x}_2 = \bar{x} + m_2(\bar{b} - \bar{y})$	Ici m_2 est un coefficient non déduit de l'échantillon.
3	$m_3 = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_y} = \rho\frac{\hat{\sigma}_x}{\hat{\sigma}_y}$	$\bar{x}_3 = \bar{x} + m_3(\bar{b} - \bar{y})$	Estimation par la méthode des moindres carrés.
4	$m_4 = \frac{\bar{x}}{\bar{y}} = f$	$\bar{x}_4 = f\bar{b}$	Estimation par le quotient.

L'estimation $\bar{x}_2$ exige que l'on dispose a priori (par exemple à l'aide d'une enquête précédente de même type) d'une approximation valable m_2 du coefficient de régression $\beta = \rho \frac{\sigma_x}{\sigma_y}$.

Les estimations x_3 et x_4 sont affectées de biais qui sont généralement trop faibles pour causer quelque inquiétude mais qui peuvent être importants lorsque le nombre d'unités d'échantillonnage est petit.

VI.2 - VARIANCES

Les formules ci-après supposent que n est petit par rapport à N.

- Estimation $\bar{x}_1$:

On a vu que

$$V(\bar{x}_1) = \sigma^2_{\bar{x}} \quad \text{de même que} \quad V(\bar{y}_1) = \sigma^2_{\bar{y}}$$

- Estimation $\bar{x}_2$

$$V(\bar{x}_2) = \sigma^2_{\bar{x}} (1 - \rho^2) + \sigma^2_{\bar{y}} (m_2 - \beta)^2$$

$$= \sigma^2_{\bar{x}} (1 - \rho^2 + \rho^2 e^2)$$

avec

$$\beta = \rho\frac{\sigma_{\bar{x}}}{\sigma_{\bar{y}}} , \qquad e = \frac{m_2 - \beta}{\beta}$$

(Noter que

$$\rho e = (m_2 - \beta)\frac{\sigma_{\bar{y}}}{\sigma_{\bar{x}}} , \text{ même si } \rho = 0$$

- Estimation $\bar{x}_3$

$$V(\bar{x}_3) = \sigma^2_{\bar{x}} (1 - \rho^2) + R$$

- Estimation x_4

$$V(\bar{x}_4) = \sigma_{\bar{x}}^2 \left(1 + \frac{C_{\bar{y}}^2}{C_{\bar{x}}^2} - 2\rho \frac{C_{\bar{y}}}{C_{\bar{x}}}\right) + R'$$

R et R' étant les restes de développements en séries de Taylon (termes en $1/n^2$ et de degrés plus élevés, négligeable si n est grand).

Pour n grand et $C_{\bar{y}} \# C_{\bar{x}}$, on a :

$$V(\bar{x}_4) \sim 2\, \sigma_{\bar{x}}^2 \, (1 - \rho)$$

Il en résulte que :

1/ $$\frac{V(\bar{x}_4)}{V(\bar{x}_2)} = \frac{1 + C_{\bar{y}}^2/C_{\bar{x}}^2 - 2\rho \; C_{\bar{y}}/C_{\bar{x}} + R'}{1 - \rho^2 + (m_2 - \beta)^2 \, \sigma_{\bar{y}}^2/\sigma_{\bar{x}}^2}$$

$$\sim \frac{2}{1 + \rho}$$

(si n grand, $m_2 \# \beta$, $C_{\bar{x}} \# C_{\bar{y}}$)

2/ $$\frac{V(\bar{x}_2)}{V(\bar{x}_3)} = \frac{\sigma_{\bar{x}}^2 \, (1 - \rho^2) + \sigma_{\bar{y}}^2 \, (m_2 - \beta)^2}{\sigma_{\bar{x}}^2 \, (1 - \rho^2) + R}$$

$$\sim 1$$

(si n grand, $m_2 \# \beta$).

3/ $$\frac{V(\bar{x}_4)}{V(\bar{x}_3)} = \frac{1 + C_{\bar{y}}^2/C_{\bar{x}}^2 - 2\rho \; C_{\bar{y}}/C_{\bar{x}} + R'}{1 - \rho^2 + R}$$

$$\sim \frac{2}{1 + \rho}$$

(si n grand et $C_{\bar{y}} \# C_{\bar{x}}$).

VI.3 - CHOIX DE L'ESTIMATION

La comparaison des variances est importante mais l'importance des calculs exigés par les estimations elles-mêmes est aussi un facteur du choix de l'estimation.

L'estimation $\bar{x}_3$ exige beaucoup plus de calculs que les autres. Ceci peut être de peu d'importance si on dispose d'un calculateur électronique ou si n est petit et s'il n'y a qu'un petit nombre d'estimations à calculer.

Les estimations $\bar{x}_2$ et $\bar{x}_4$ sont très simples à calculer, ne demandant guère plus d'effort que $\bar{x}$.

Les variances de $\bar{x}_3$ et $\bar{x}_4$ sont sensiblement égales si on dispose pour m_2 d'une valeur acceptable fournie par une étude antérieure : il suffira pratiquement que m_2 ne diffère pas de β de plus de 30 à 40 % (c'est-à-dire que $e < 0,3$ ou $0,4$).

Dans une suite continue d'enquêtes, on pourra souvent adopter pour m_2 l'estimation fournie par la méthode des moindres carrés à partir des précédentes estimations du coefficient de régression.

Le choix de l'estimation dépend en fait de divers arguments concurrents liés à la connaissance du domaine d'étude, des facilités d'enquête et de calculs, éléments qu'il n'est pas possible de placer dans un cadre théorique général.

- Si la corrélation ρ entre x et y est modérée ou forte, mais si la ligne de régression de x en y passe loin de l'origine, l'estimation $\bar{x}_3$ sera nettement supérieure à $\bar{x}_4$ et à $\bar{x}_1$.
- Si la dispersion relative des y est nettement plus grande que celle des x (c'est-à-dire si $C_{\bar{y}}$ est nettement supérieure à $C_{\bar{x}}$), l'estimation x_4 par le quotient est beaucoup moins précise que la simple estimation $\bar{x}_1 = \bar{x}$, même si ρ est élevé et plus particulièrement encore si la ligne de régression passe loin de l'origine.
- D'autre part, si la ligne de régression de x en y passe par l'origine ($\rho\ C_{\bar{x}} = C_{\bar{y}}$) ou au voisinage, $\bar{x}_3$ et $\bar{x}_4$ auront sensiblement la même variance mais $\bar{x}_4$ peut être beaucoup plus facile à calculer.

VI.4 - ERREURS D'ECHANTILLONNAGE DANS L'ESTIMATION DE $\bar{b}$

Il arrive souvent que la population des y par unité d'échantillonnage n'est pas connue avec la précision donnée par un recensement mais est fournie elle-même par un autre échantillon : ceci introduit un terme additionnel dans les variances.

Soit n la taille de l'échantillon que l'on étudie, n' celle de l'autre échantillon qui donne une estimation de $\bar{b}$ avec une variance $n\,\sigma^2_{\bar{y}}/n'$.

Nous distinguerons deux cas :

1/ L'échantillon de taille n est un sous-échantillon d'un échantillon de taille n'.

2/ Les deux échantillons sont indépendants.

La table ci-après donne les variances relatives dans les deux cas.

Estimation	Cas n°1 Echantillon de taille n sous échantillon de l'échantillon de taille n'.	Cas n° 2 Les deux échantillons de tailles n et n' sont indépendants.
$\bar{x}_1$	$C^2_{\bar{x}}$	$C^2_{\bar{x}}$
$\bar{x}_2$	$C^2_{\bar{x}}\left[1 - \rho^2(1 - e^2)\left(1 - \frac{n}{n'}\right)\right]$	$C^2_{\bar{x}}\left[1 - \rho^2(1 - e^2) + \rho^2(1 + e)^2\frac{n}{n'}\right]$
$\bar{x}_3$	$C^2_{\bar{x}}\left[1 - \rho^2\left(1 - \frac{n}{n'}\right)\right]$	La même que dans le cas n° 1
$\bar{x}_4$	$C^2_{\bar{x}} - \left(2\rho\ C_{\bar{x}}C_{\bar{y}} - C^2_{\bar{y}}\right)\left(1 - \frac{n}{n'}\right)$	$C^2_{\bar{x}} - (2\rho\ C_x C_{\bar{y}} - C^2_{\bar{y}})\ (1 - \frac{n}{n'})$ $+ (2C_{\bar{y}}\ \frac{n}{n'})\ (C_{\bar{y}} - \rho C_{\bar{x}})$

Le tableau ci-après montre une comparaison numérique pour les estimations 1, 3 et 4 dans le cas n° 1 pour $C_{\bar{x}} = C_{\bar{y}}$.

On observera que lorsque $C_{\bar{x}} \neq C_{\bar{y}}$, et que ρ est élevé, les variances des estimations 3 et 4 sont presque égales et que ces deux estimations sont nettement plus précises que l'estimation 1.

Ce bénéfice est particulièrement frappant quand n' est grand par rapport à n.

La différence entre les estimations 1 et 4 s'annule lorsque $\rho = 0,5$, mais la précision de l'estimation 4 diminue rapidement lorque $\rho < 0,5$.

D'autre part l'estimation 3 reste supérieure à l'estimation 1, même pour ρ petit, (compte non tenu des calculs supplémentaires qu'elle exige).

(L'estimation 2 n'est pas envisagée dans ce tableau en raison de la large latitude de choix pour m_2, qui dépend essentiellement des informations locales disponibles).

Rapports des variances de $\bar{x}_3$ et $\bar{x}_4$ à la variance de $\bar{x}_1$.

(Cas n° 1, l'échantillon de taille n est un sous-échantillon de n')

Hypothèse $C_{\bar{x}} = C_{\bar{y}}$

ρ	Estimation	n'/n = 5	n'/n = 10	n'/n ∞
0,95	3	0,278	0,188	0,097
	4	0,280	0,190	0,100
0,9	3	0,352	0,271	0,190
	4	0,360	0,260	0,200
0,8	3	0,488	0,424	0,360
	4	0,520	0,460	0,400
0,7	3	0,606	0,559	0,510
	4	0,680	0,640	0,600
0,5	3	0,800	0,775	0,750
	4	1,000	1,000	1,000
0,3	3	0,928	0,919	0,910
	4	1,320	1,360	1,400
0,1	3	0,992	0,991	0,990
	4	1,640	1,720	1,800

VI.5 - EXEMPLE

Des études antérieures ont montré que le coefficient de corrélation ρ entre deux variables particulières x et y n'était presque jamais inférieur à 0,8 et souvent beaucoup plus élevé.

Soient $f = \bar{x}/\bar{y}$, le rapport estimé dans un sous-échantillon d'effectif n, et $\bar{b}$ l'estimation de la valeur moyenne de y, déjà fournie par un échantillon principal d'effectif n'.

L'examen de la table montre que, par exemple, pour $\rho = 0,8$ et $\frac{n'}{n} = 10$, si on prélève un sous-échantillon d'effectif $n = \frac{n'}{10}$ et si on calcule

$$\bar{x}_4 = f\bar{b} = \bar{b}\left(\frac{\bar{x}}{\bar{y}}\right)$$

on aura

$$V(\bar{x}_4) = 0,460\ V(\bar{x}_1)$$

Il en résulte que l'erreur type de $\bar{x}_4$ sera seulement $\sqrt{10 \times 0,46} = 2,1$ fois supérieur à celle qui l'obtiendrait en utilisant $\bar{x}_1$, estimée à partir de l'échantillon principal d'effectif n'.

La table montre aussi que $V(\bar{x}_3)$ serait inférieur d'environ 10 % à $V(\bar{x}_4)$ pour un calcul plus important.

Dans cet exemple, on a supposé que l'échantillon principal n' était déjà utilisé pour d'autres buts, l'étude actuellement envisagée en étant un sous-produit.

En conséquence, le seul problème était de fixer une fraction $\frac{n}{n'}$ de sous échantillonnage donnant une précision suffisante pour le but envisagé.

Si le plan de sondage devait être entièrement établi pour estimer $\bar{x}$, l'information supplémentaire relative à y ne pouvant pas être obtenue d'une autre source, on devrait évidemment tenir compte des coûts d'enquête relatifs à x et y et, aussi bien pour n que pour n', déterminer les tailles optimales.

Les couts devraient entrer en ligne de compte dans les formules d'une manière analogue à celle déjà envisagée dans les sondages avec stratification.

BIBLIOGRAPHIE

I. REFERENCES CITEES

WILLIAM G. COCHRAN - *Sampling Techniques* (Wiley, 1953, 1963). "Relative accuracy of systematic and stratified random samples for a certain class of populations" Annals of Mathematical Statistics, vol. 17 - 1946 - p. 164-177

W. EDWARDS DEMING - *Some Theory of Sampling* (Wiley, 1950). "On a probability mechanism to attain an economic balance between the résultant error of response and the bias of non response", Journal of the American Statistical Association, vol. 48 - 1953, p. 732/772. *Sample Design in Business Research* (Wiley, 1960)

JAMES DURBIN - "A note on the application of Quenouille's method of bias reduction to the estimation of ratios", Biometrika Vol. 46, 1959 - p.477/80.

Sir RONALD FISHER - "On the mathematical fondations of theoretical statistics", Philosophical Transactions of the Royal Society, vol. 222 A, p. 309/368. (Reproduced in Fisher, *Contributions to Mathematical Statistics*, Wiley, 1950) ; *Statistical Methods and Scientific Inference* (Oliver and Boyd, 1956)

KARL FRIEDERICH GAUSS - *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Pars Posterior* (Göttingen, 1923). Reprinted in his *Werke* Vol. 4, 1876, art. 38.

FRANK E. GRUBBS and CHALMERS L. WEAVER - "The best unbiased estimate of population standard deviation based on group-ranges, Journal of the American Statistical Association, vol. 42, 1947 - p. 224/241.

MORRIS H. HANSEN and WILLIAM N. HURWITZ - "On the theory of sampling from finite populations", Annals of Mathematical statistics, vol. 14 - 1943 - p. 333/362. "The problem of nonresponse in sample-surveys", Journal of the American Statistical Association, vol. 41, 1946 - p. 517/529.

MORRIS H. HANSEN - WILLIAM N. HURWITZ and WILLIAM G. MADOW, *Sample Survey Methods and Theory*, (Wiley, 1953).

SIEGFRIED KOLLER - "Replicated sampling", Bulletin of the International Statistical Institute, vol. 36 - part 3, 1958. p. 135/143. Kontrol der Auswahl, pages 86 and 87 in *Stichproben der Amtlichen Statistik* (Wiesbaden: Statistiches Bundesamt. W. Kohlhammer, Stuttgart, 1960).

P.C. MAHALANOBIS - "Studies in educational tests, n° 1, The reliability of a group-test of intelligence in Bengali", Sankhyà, vol. 1 - 1933, p. 25/49. "A sample-surveys of the acreage under jute in Bengal", Sankhyà vol. 4 ; 1940 - p. 511/529. "On large-scale sample-surveys", Philosophical Transactions of the Royal Society, Series B, vol. 231, 1944 - p. 329/451. "The National Sample Survey : general report n° 1", Sankhyà vol. 13, 1954 - p. 47/214. Report N° 5 - Sankhyà vol. 14, 1954, p. 264/316. "Recent experiences in statistical sampling in the Indian Statistical Institute". Journal of the Royal Statistical Society - vol. cix, 1946 - p. 325/378.

NATHAN MANTEL - "On a rapid estimation of standard errors for the means of small samples", The American Statistician vol. 5, 1951 - p. 26/27.

J. NEYMAN - "On the two different aspects of the representative method. Journal of the Royal Statistical Sociéty, vol. xcvii, 1934 - p. 558/606 "Contribution to the theory of sampling human populations", journal of the American Statistical Association, vol. 33 - 1938, p. 101 - 116. Eq. 49 on page 110 in particular.

KARL PEARSON - "Contributions to the mathematical theory of évolution II. Skew variation". Philosophical transactions of the Royal Society of London, Series A, vol. 185, 1894 - p. 71/110.

M.H. QUENOUILLE - "Approximate tests of correlation in time-series", Journal of the Royal Statistical Society, Series B, vol. 11, 1949, p. 68-84. "Notes on bias in estimation", Biometrika, vol. 43, 1956 - p. 533/560. "Rapid Statistical Calculations", (Hafner, 1959). Pages 5, 6, 7, show estimates of the standard deviation by use of the range.

F.E. SATTERTHWAITE - "An approximate distribution of estimates of variance components" Biometrics, vol. 2 - 1946 - p. 110/114.

WALTER A. SHEWHART - *Economic Control of Quality of Manufactured Product* (Van Nostrand, 1931). *Statistical Method from the Viewpoint of Quality Control.* (The graduate School, Department of Agriculture, Washington, 1939).

F.F. STEPHAN and Philip J. Mc. CARTHY - *Sampling Opinions* (Wiley, 1958).

L.H.C. TIPPETT - *The Methods of Statistics* (Williamq and Norgate, and Wiley, 1931, 1952).

JOHN W. TUKEY - "Bias and confidence in not-quite large samples", Annals of Mathematical Statistics, vol. 29, 1958 - p. 614.

FRANK YATES - *Sampling Methods for Census and Survey* (Griffin, 1950). Highly recommanded.

II - AUTRES OUVRAGES CLASSIQUES

CHURCHILL EISENHART - MILLARD W. HASTAY - W. ALLEN WALLIS - *Selected Techniques of Statistical Analysis* (Mc. GRAW-HILL, 1947).

J. NEYMAN - *Lectures and Conferences on Mathematical Statistics and Probability* (The graduate School, Department of Agriculture, Washington, 1939, 1952).

P.V. SUKHATME - *Sampling Theory of Surveys, with Applications* (Iowa State College Press, and Indian Society of Agricultural Statistics, New-Delhi, 1953).

OSKAR ANDERSON - *Probleme der Statistischen Methodenlehre in den Sozialwissenschaften* (Physica-Verlag, Würzburg, 1957).

TORE DALENIUS - *Sampling in Sweden* (Almquist et Wiksell, Stockholm, 1957). In English.

ERNEST KURNOW - GERALD GLASSER - F.R. OTTMAN - *Statistics for Business Decisions* (Irwin, 1959)

R. CLAY SPROWLS - *Elementary Statistics for Students of Social Science and Business* (Mc. GRAW-HILL, 1955)

W. ALLEN WALLIS and HARRY V. ROBERTS - *Statistics : a New Approach* (Mc. MILLAN, 1956).

III - AUTRES RERERENCES D'ORIGINE FRANCAISE (N.d.I.R.)

P. THIONET - Application des Méthodes de sondage aux enquêtes statistiques. I.N.S.E.E. 1953.

P. THIONET - La Théorie des sondages. I.N.S.E.E. 1954.

P. THIONET - Méthodes statistiques modernes des Administrations fédérales aux Etats-Unis. Hermann, 1946.

DESABIE - Théorie et pratique des sondages (Cours de l'Ecole Nationale de la Statistique et de l'Administration Economique) 2 Vol. I.N.S.E.E., 1960

MATTHIS - THIONET et LEVY BRUHL - Manuel des enquêteurs par sondage. I.N.S.E.E. 1955.