PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

LE CORVEC

Sondages en grappes à un ou deux degrés

Publications des séminaires de mathématiques et informatique de Rennes, 1966-1967
« Séminaires de probabilités et statistiques », , exp. n° 7, p. 1-34

<http://www.numdam.org/item?id=PSMIR_1966-1967____A7_0>

UNIVERSITE DE RENNES
FACULTE DES SCIENCES

SEMINAIRE de PROBABILITES et STATISTIQUES

SONDAGES EN GRAPPES A UN OU DEUX DEGRES

par

Monsieur LE CORVEC

Année 1966-1967

SONDAGES EN GRAPPES A UN OU DEUX DEGRES

Iè Partie :

- Définition et exemples d'échantillons en grappes à un ou deux degrés.
- Notations.

IIè Partie :

- A) Estimé d'un total
  Variance de cet estimé
  Variance relative de cet estimé.
- B) Estimé d'un quotient
  Variance de cet estimé
  Variance relative de cet estimé.

IIIè Partie :

- Comparaison avec un échantillon simple au hasard

IVè Partie :

- Utilisation pour la détermination de l'échantillon optimum.

Iè PARTIE

A) DEFINITIONS et NOTATIONS.

Le terme "unité élémentaire" est utilisé pour désigner un individu de la population sur laquelle on désire faire des mesures et pour laquelle des analyses doivent être faites à partir des résultats de l'enquête. Lors du calcul des totaux et des pourcentages, l'unité sur laquelle le total ou le pourcentage porte, est l'unité élémentaire. Quand des répartitions de fréquence sont faites, c'est l'unité élémentaire qui est placée dans des intervalles, suivant la valeur d'une de ses caractéristiques : c'est l'unité élémentaire pour laquelle des mesures telles que la médiane, sont calculées. Ainsi, si le but de l'enquête est d'estimer le revenu total ou moyen, ou le revenu pour des personnes individuelles, alors la personne est l'unité élémentaire. Si on estime le revenu total d'une famille, le revenu médian d'une famille, ou la distribution des revenus familiaux, la famille est l'unité élémentaire.

Ainsi, l'unité élémentaire dépend de l'analyse à faire et est déterminée par les buts de l'enquête et non par la structure de l'échantillon.

Dans quelques cas, le but de l'enquête n'entraîne pas la spéficication d'une unité élémentaire, comme lorsque seuls des totaux font l'objet de mesures. Il est possible que plus d'une unité élémentaire fasse l'objet d'une enquête lorsque, par exemple, on désire mesurer à l'aide d'une enquête seulement les caractéristiques des familles, et les caractéristiques de l'individu.

L'échantillonage en grappes, entraîne, à l'aide de certaines considérations, une division de la population d'unités élémentaires en groupes ou "grappes" qui servent d'unités primaires d'échantillonage. Dans quelques cas, un échantillon de telles unités primaires est choisi, et tous les membres de la population associés aux unités choisies sont inclus dans l'échantillon : c'est un échantillonage à degré unique. Dans d'autres cas, les unités

primaires sélectionnées sont divisées en unités secondaires, et il y a alors un ou plusieurs degrés de l'échantillonage. Les unités initiales définies et sélectionnées seront appelées "first stage units" ou "primary sampling units" "en abrégé "psu").

Par échantillonage simple à 1 degré, nous entendrons qu'il y a seulement un degré d'échantillonage (i.e. il n'y a pas de sous échantillonage) et qu'un échantillonage simple au hasard est utilisé pour selectionner les unités. Par échantillonage simple à deux degrés nous désignerons une structure dans laquelle :

a) Un échantillonage à 2 degrés est utilisé.

b) Les psu sont choisies par un échantillonage simple au hasard.

c) Les unités du second degré sont choisises à l'intérieur de chaque psu à l'aide d'un échantillonage simple au hasard.

d) Les fractions sondées sont constantes à l'intérieur de chaque psu.

Dans cet exposé, nous étudierons l'échantillonage simple à 1 ou 2 degrés.

M est le nombre total de psu suivant lesquelles la population a été divisée. Ainsi, si l'enquête porte sur la mesure des caractéristiques des habitations d'une cité, nous pouvons faire correspondre un quartier à chaque habitation, et utiliser le quartier comme psu. M est le nombre de tels quartiers dans la cité.

m est le nombre de psu incluses dans l'échantillon.

N_i est le nombre d'unités de liste incluses dans ième psu.

L'unité de liste est l'unité sélectionnée au dernier degré d'échantillonage : elles est appelée unité de liste car nous aurons souvent à dresser une liste de ces unités à l'intérieur de la psu et nous choisirons alors un échantillonage à 2 degrés, l'unité de liste peut être appelée l'unité d'échantillonage. Par exemple, dans un échantillon d'habitations, de gens, ou de familles, il est souvent pratique de prendre la maison particulière comme unité de liste pour de petites habitations contenant une ou quelques familles par habitation, et de prendre l'appartement comme unité de liste à l'intérieur de grans immeubles. On doit distinguer unité de liste et unité élémentaire.

Dans certains cas, unités de liste et unités élémentaires sont identiques, mais elles peuvent être différentes.

$N = \sum_{i}^{M} N_i$ est le nombre d'unités de liste dans la population.

$\bar{N} = \frac{N}{M}$ est le nombre moyen d'unités de listes par psu dans la population.

n_i est le nombre d'unités de liste extraites de la ième psu.

$n = \sum_{i=1}^{m} n_i$ est le nombre total d'unités de liste dans l'échantillon.

$\bar{n} = \frac{n}{m}$ est le nombre moyen d'unités de liste de l'échantillon par psu sélectionnée.

$f_1 = \frac{m}{M}$ est la fraction sondée au 1er degré.

Dans cet exposé, nous supposerons les fractions sondées constantes à l'intérieur de chaque psu c'est à dire :

$f_2 = \frac{n_i}{N_i}$ est constante (fraction sondée au 2° degré).

$f = f_1 f_2 = \frac{n}{N}$ est la fraction sondée totale.

X_{ij} se rapporte à la jème unité de liste de la ième psu.

Par exemple, si l'unité de liste est la maison particulière, et qu'on veuille étudier le loyer des habitants, et s'il y a deux habitants dans la ijème unité de liste, X_{ij} est la somme des loyers des deux habitants. Si l'un des deux est le propriétaire, X_{ij} est le loyer du locataire. De façon analogue, si la caractéristique à mesurer est le nombre de locataires, la valeur 0 sera affectée au propriétaire, et la valeur 1 au locataire ; s'il y a 2 locataires $X_{ij} = 2$; s'il y a 2 copropriétaires, $X_{ij} = 0$. Pour étendre cette illustration supposons que le but soit de mesurer seulement les caractéristiques des habitations occupées par des familles ayant un certain revenu, soit par exemple des familles ayant un revenu supérieur à 5000 dollars, et la caractéristique à mesurer étant le nombre de propriétaires. Dans ce cas, on affectera

la valeur 1 à la famille ayant un revenu supérieur à 5000 dollars et qui soit propriétaire, et 0 aux autres. Alors X_{ij} est la somme de ces valeurs pour toutes les habitations se trouvant dans l'unité de liste.

$X_i = \sum_{j=1}^{N_i} X_{ij}$ est le total pour la $i^{è}$ psu.

Par exemple, si le quartier est la psu, X_i désignera le nombre total de pièces dans l'ensemble des habitations du $i^{è}$ quartier, ou le nombre de maisons particulières dans le $i^{è}$ quartier.

$X = \sum_{i=1}^{M} X_i$ est la valeur total sur toutes les psu de la population.

Aussi, la suppression d'un indice désigne la sommation sur cet indice.

$\bar{X} = \frac{X}{M}$ est la valeur moyenne par psu.

Cette valeur moyenne par psu est rarement le but d'une enquête. Elle apparaîtra néanmoins dans certaines formules de la variance de l'échantillon.

$\bar{\bar{X}} = \frac{X}{\bar{N}} = \frac{\bar{X}}{\bar{N}}$ est la valeur moyenne par unité de liste.

Ainsi, $\bar{\bar{X}}$ pourrait être le nombre moyen de locataires par unité de liste, ou le loyer moyen par unité de liste, ou le nombre de personnes par unités de liste (maisons).

De façon analogue :

$\bar{\bar{X}}_i = \frac{X_i}{N_i}$ est la valeur moyenne par unité de liste pour la $i^{è}$ psu.

$R = \frac{X}{Y}$. Si Y est le nombre d'unités élémentaires de la population, alors R est une valeur moyenne par unité élémentaire.

Les buts de l'enquête consistent souvent à estimer la valeur R pour différentes caractéristiques.

Notons bien que $R = \frac{X}{Y} \neq \frac{X}{N}$ car N représente par exemple le nombre d'appartements contenant au plus 2 familles et Y représente le nombre de bâtiments, maisons et H.L.M.

x_{ij} est la valeur pour la $j^{è}$ unité de liste dans l'échantillon, extraite de la $i^{è}$ psu de l'échantillon.

$x = \sum_{i=1}^{m} x_i$ est le total pour toutes les unités de liste de l'échantillon.

La barre et la double barre ont la même signification pour les moyennes d'échantillon que pour les moyennes de population.

$$\text{Ainsi } \bar{\bar{x}} = \frac{x}{n} \quad \bar{\bar{x}}_i = \frac{x_i}{n_i} \quad \bar{x} = \frac{x}{m}$$

x' et x'_i sont des estimés non biaisés de X et X_i, déduits de l'échantillon.

$r = \frac{x}{y} = \frac{x'}{y'}$ est un estimé de R.

Le terme "dernière grappe" ("ultimate duster") est utilisé pour désigner l'ensemble des unités incluses dans l'échantillon à partir d'une psu. La somme totale d'une caractéristique pour la $i^{è}$ grappe est x_i, la taille de la $i^{è}$ dernière grappe est n_i, et la taille moyenne des m dernières grappes de l'échantillon est $\bar{n} = n/m$. Cette définition de "dernière grappe" est valable pour un nombre quelconque de degrés d'échantillonnage.

Remarque : Chaque fois qu'une lettre capitale (X ou Y) sert d'indice à une variance ou a une covariance, ceci signifie que le symbole auquel il est adjoint, porte sur une caractéristique de la population qui ne dépend pas de la taille de l'échantillon, comme par exemple, la covariance d'une population.

Les petites lettres seront utilisées comme indices pour indiquer les variances ou covariances des estimés.

Ainsi :

$$S^2_{2X} = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j}^{N_j} (X_{ij} - \bar{\bar{X}}_i)^2$$

est la variance de la population constituée d'unités du 2ème degré dans une psu, et

$$x^2_{2X} = \frac{1}{n} \sum_{i}^{m} \frac{n_i}{n_i - i} \sum_{j}^{n_j} (x_{ij} - \bar{\bar{x}}_i)^2$$

est un estimé de S^2_{2X}.

On doit les distinguer de :

σ_x^2 = variance de la moyenne de l'échantillon x

$\sigma_{\bar{x}}^2$ = variance de l'estimé de $\bar{x}$

$\sigma_{x'}^2$ = variance de l'estimé de X, x'

S_x^2 , $S_{\bar{x}}^2$, $S_{x'}^2$ sont des estimés de σ_x^2, $\sigma_{\bar{x}}^2$, $\sigma_{x'}^2$

B) Méthode d'Echantillonage en Grappes Illustrée à l'Aide d'un Echantillonage Aréolaire.

L'échantillonage en grappes fournit souvent un moyen pratique et peu coutêux pour fixer, avant même l'enquête, la probabilité d'inclure chaque membre de la population dans l'échantillon. Ainsi, il rend souvent possible l'usage d'un plan d'échantillonage rentable dans des situations où l'échantillonage au hasard d'éléments de la population pourrait être très coûteux. Par exemple, il ne serait pas possible, dans de nombreux cas, de tirer un échantillon au hasard de personnes de la population des Etats-Unis, ou d'une grand population pour laquelle aucune liste complète et à jour, ne soit valable. Cependant, des listes complètes de grappes (une liste de comtés ou de petites zônes des Etats-Unis) sont souvent valables et peuvent être préparées sans grand effort, et d'une telle liste, des grappes peuvent être rapidement échantillonées avec des probabilités connues. De plus, l'usage des grappes peut avoir une grande répercussion sur d'autres coûts.

Comme illustration de l'échantillonage en grappes, supposons que nous nous intéressions à certaines caractéristiques des habitations d'une ville. Par exemple, nous souhaitons connaître le nombre total d'habitations, la proportion vacante, les proportions d'autres caractéristiques, et même les loyers totaux de types variés d'unités.

Un échantillon des logements d'une cité peut être obtenu en dressant une liste de tels logements et en faisant soit un échantillon au hasard à partir de cette liste, soit un échantillon stratifié. De telles méthodes peuvent fournir un échantillon efficace du seul point de vue de la taille de

la taille de l'échantillon. Pour des buts pratiques, d'autres plans d'échantillonage peuvent fournir la même précision. Pourquoi alors considerer d'autres plans ? La raison est que le coût peut n'avoir aucun lien avec la taille de l'échantillon. Supposons qu'une liste de la cité ne soit pas valable. Alors pour sélectionner un échantillon au hasard, il faudrait aller soi-même faire le recensement de toutes les maisons, ce qui reviendrait à 30 000 dollars pour une ville de 300 000 habitants. De plus, si on recueille les informations sur une grande zône, il peut être ruineux d'interroger un échantillon très dispersé au cas où l'échantillon a été tiré d'une liste valable. Le sondage aréolaire fournit souvent un moyen pratique d'identifier la population à des grappes que l'on considère comme unités d'échantillonage.

Dans l'échantillonage aréolaire, toute la zône habitée, est divisée est plus petites zônes, et chaque unité élémentaire (ici chaque habitation) est incluse dans une et seule zône, par exemple la petite zône sur laquelle l'habitation est située.

On peut montrer que si une liste de zônes est valable et si un échantillonage simple est fait sur les zônes, et si la population comprise dans ces zônes est complètement énumérée, alors la probabilité d'appartenir à l'échantillon est la même pour chaque élément de la population.

Remarque : On peut montrer que, dans un échantillon à deux degrés, la probabilité pour qu'un individu appartienne à l'échantillon, est égal à l'inverse du produit des fractions d'échantillonage à chaque degré d'échantillonage.

Soit :

$$f_1 f_2 = \frac{n_i}{N_i} \cdot \frac{m}{M}$$

avec $\frac{n_i}{N_i}$ constant quelque soit i.

C.) Estimé d'un Quotient à Dénominateur Variable.

L'usage d'estimés de quotients de variables aléatoires prend une plus grande importance dans l'échantillonage en grappes que dans l'échantillonage

simple au hasard, car, lors de l'échantillonage en grappes, on est rarement intéressé par la valeur moyenne par psu. Ainsi, si on tire un échantillon de quartiers pour estimer les caractéristiques des magasins, on aura rarement intérêt à étudier les chiffre d'affaire par quartier, mais les chiffre d'affaires moyen par magasin présentera plus d'intérêt. Ou, dans une population, le nombre moyen de travailleurs ou de chômeurs par quartier présentera moins d'intérêt que le pourcentage de la population ayant une caractéristique donnée. Dans les échantillonages en grappes, les estimés de telles moyennes ou pourcentages sont des quotients de variables aléatoires. La précision de tels quotients doit être évaluée à partir de la formule de la variance du quotient de 2 variables aléatoires.

D) Quelques Procédés Simples de Sélection d'Echantillons en Grappes, où l'on désire ne prendre que 1% des maisons.

1° méthode : On fait une partition de la ville en quartiers.

démarche 1 : numéroter les quartiers

démarche 2 :

a) échantillonage simple au hasard des quartiers où l'on ne prend que 1% des quartiers.

b) ou échantillonage symétrique.

démarche 3 : Collecter l'information désirée à partir des habitations se trouvant dans les quartiers sélectionnés (ici on a un échantillonage à degré unique).

2° méthode : Echantillonage de zônes avec sous échantillonage de plus petites zônes.

On veut 1% des habitations.

On prendra 1/25 des quartiers, et dans chaque quatier 1/4 des maisons.

démarche 1 : numéroter les quartiers

démarche 2 : faire comme dans la méthode 1 mais en ne tirant que 1/25 des quartiers

démarche 3 : diviser les quartiers de l'échantillon en 4 zônes.

démarche 4 : tirer à l'aide d'un échantillonage simple au hasard 1 zône dans chaque quartier et prendre toutes les maisons se trouvant dans les zônes.

Ce procédé est plus coûteux que la lère méthode à cause des déplacements.

3° méthode :

démarche 1 : numéroter les quartiers

démarche 2 : échantillonage simple au hasard (prendre 1/25 des quartiers).

démarche 3 : lister toutes les unités en considérant

1 unité si moins de 4 appartements

2 unités de 4 à 8 appartements

3 unités de 8 à 12 appartements etc...

démarche 4 : les numéroter

démarche 5 : à l'intérieur de chaque quartier prendre 1 unité de liste sur 4 à l'aide :

- soit d'un échantillon simple au hasard
- soit d'un échantillon symétrique.

Définition de la "Structure Optimum" :

c'est la structure qui fournit l'exactitude requise à moinsdres frais, ou la structure qui fournit la meilleure précision avec les fonds et ressources dont on dispose.

IIè PARTIE

Pour un échantillonage simple à 1 ou 2 degrés, un estime, basé sur l'échantillon, du total pour une caractéristique quelconque, comme, par exemple, la valeur totale d'habitations dans une zône, la production totale d'une forme, ou le nombre total de personnes ayant une certaine caractéristique, est donné par :

$$x' = \frac{1}{f} x \qquad (1)$$

où x est la valeur totale d'une caractéristique pour les unités de liste incluses dans l'échantillon, et f est la fraction globale échantillonée. L'estimé donné plus haut est un estimé sans biais ; et cet estimé, obtenu en multipliant le total de l'échantillon par l'inverse de la fraction sondée, est appelé : "estimé simple non biaisé".

L'estimé de la valeur moyenne par unité élémentaire d'une caractéristique comme, par exemple, la valeur moyenne du loyer par habitation dans une zône, ou le pourcentage de propriétaires est donné par le quotient :

$$r = \frac{x}{y} = \frac{x'}{y'} \qquad (2)$$

où x et x' sont respectivement, le total de l'échantillon, et son estimé pour la population quant à la caractéristique observée, et où y et y' sont respectivement le nombre d'unités élémentaires dans l'échantillon, et l'estimé du nombre total d'unités élémentaires dans la population (i.e. $y' = \frac{1}{f} y$).

L'équation (2) est aussi utilisée pour obtenir un estimé du quotient de la valeur moyenne d'une certaine caractéristique par celle d'une autre caractéristique tel que le rapport loyer sur le revenu.

Les équations (1) et (2) sont valables si on utilise un échantillonage à 1 ou 2 degrés. L'estimé (2) est un quotient de deux variables aléatoires, et il est donc sujet à un biais qui devient insignifiant dès que le nombre de psu augmente, et qui pourra donc être négligé pour des échantillons relativement grands.

Remarques et Notations.

$\sigma^2_{x'}$ désignera la variance de l'estimé de X donc :

$$\sigma^2_{x'} = E\left[x' - E(x')\right]^2 = E\left[x' - X\right]^2$$

et représente donc la moyenne du carré de l'erreur absolue. C'est pourquoi on attachera une certaine importance au calcul de la variance de l'estimé.

$V^2_{x'}$ désignera la variance relative de x' qui est par définition :

$$V^2_{x'} = \frac{\sigma^2_{x'}}{\left[E(x')\right]^2}$$

Si, comme il arrive assez souvent lorsque la taille de l'échantillon est grande, x' est pratiquement sans biais, alors E(x') = X et

$$V^2_{x'} = \frac{E(x' - X)^2}{X^2} = E\left(\frac{x' - X}{X}\right)^2$$

et $V^2_{x'}$ représentera la moyenne du carré de l'erreur relative. Aussi, attachera-t-on une grande importance à la valeur de $V^2_{x'}$ qui nous donnera une idée de la précision. Lorsque $V^2_{x'}$ sera faible, x' sera donc un bon estimé.

Lorsqu'on rencontrera N_i, pour savoir si N_i désigne le nombre d'éléments de la i$^{\text{è}}$ psu de l'échantillon, ou de la i$^{\text{è}}$ psu de la population, il suffit de regarder le signe de sommation Σ qui précède.

Si on a : $\sum_{i=1}^{m} N_i$ on est dans le 1er cas

Si on a : $\sum_{i=1}^{M} N_i$ on est dans le 2è cas

Même remarque pour les n_i.

A) ESTIME d'un TOTAL

L'estimé x' = 1/f x est sans biais. En effet

$x' = \frac{M}{m} \sum_{i=1}^{m} x'_i$ $\quad x'_i$ étant l'estimé du total de la i$^{\text{è}}$ psu de l'échantillon

$= \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$

Donc : $E\,x' = \frac{M}{m} \sum_{i=1}^{m} E(x_i')$

Or : $E(x_i') = E\left[E\, x_i'/b^* = B_k\right]$

où $E(x_i'/b^* = B_k) = E_k\, x_i'$ signifie l'espérance conditionnelle de x_i' sachant que la $i^{è}$ psu de l'échantillon est la $k^{è}$ psu de la population.

$$E\,x' = \frac{M}{m} \sum_{i=1}^{m} E\, x_i' = \frac{M}{n} \sum_{i=1}^{m} E\left[E_k\, x_i'\right]$$

$$= \frac{M}{m} \sum_{i=1}^{m} E\left[E_k \frac{N_i}{n_i} \sum_{j=1}^{n_i} x_{ij}\right] = \frac{M}{n} \sum_{i=1}^{m} E \frac{N_i}{n_i} \left[E_k \sum_{j=1}^{n_i} x_{ij}\right]$$

$$= \frac{M}{n} \sum_{i=1}^{m} E \frac{N_i}{n_i} \frac{n_k}{N_k} \sum_{i=1}^{N_k} X'_{kj}$$

Or : $\frac{n_i}{N_i} = \frac{n_k}{N_k}$

$$E\,x' = \frac{M}{n} \sum_{i=1}^{m} E \sum_{j=1}^{N_k} X'_{kj} = \frac{M}{n} \sum_{i=1}^{m} \frac{1}{m} \sum_{j=1}^{M} X_j$$

$$E\,x' = M\,\bar{X} = X$$

VARIANCE et VARIANCE RELATIVE de l'ESTIME

Nous allons montrer que :

$$\sigma^2_{x'} = M^2 \frac{M-m}{M} \frac{S^2_{1X}}{n} + M^2 \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{S^2_{2X}}{m\bar{n}}$$

$$V^2_{x'} = \frac{M-m}{M} \frac{B^2_X}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{W^2_X}{m\bar{n}}$$

avec

$$S^2_{1X} = \frac{\sum_{i=1}^{M} (X_i - \bar{X})^2}{M-1}$$

$$B^2_X = \frac{S^2_{1X}}{\bar{X}^2}$$

où S_{1X}^2 est la variance et B_X^2 la variance relative entre les grappes et :

$$\left\| \begin{array}{l} S_{2X}^2 = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}}_i)^2 \\ W_X^2 = \frac{S_{2X}^2}{\bar{\bar{X}}^2} \end{array} \right.$$

où S_{2X}^2 est la variance moyenne, et W_X^2 la variance relative moyenne entre unités de liste à l'intérieur des psu.

Pour calculer $\sigma_{x'}^2$, on procèdera de la façon suivante :

$$\sigma_{x'}^2 = E\, \sigma_{x'/1}^2 + \sigma^2 E(x'/1)$$

où $\sigma_{x'/1}^2$ est la variance conditionnelle de x' en tenant fixe "l'ensemble des m psu", et E(x'/1) l'espérance conditionnelle de x' en tenant fixe "l'ensemble des m psu". $\sigma^2 E(x'/1)$ est la variance de ces espérances conditionnelles sur "tous les ensemb-les possibles de m psu".

La variance de x' pour un ensemble fixé d'unités primaires dans l'échantillon est la variance de (M/m x") où x" est l'estimé du total pour l'ensemble fixe des m psu (i.e. x" estimé $\sum_{i=1}^{m} X_i$).

Donc :

$$x' = \frac{M}{m} x'' \Longrightarrow \sigma^2(x'/1) = \frac{M^2}{m^2} \sigma^2(x'')$$

Or :

$$x'' = \frac{N_i}{n_i} x = \frac{N_i}{n_i} n \bar{x}$$

Donc :

$$\sigma^2(x'/1) = \frac{M^2}{m^2} \sigma^2(x'') = \frac{M^2}{m^2} \frac{N_i^2}{n_i^2} n^2 \sigma^2(\bar{x})$$

Or les m psu peuvent être considérées comme strates. Aussi pour trouver $\sigma^2(\bar{x})$, nous appliquerons le théorème suivant :

THEOREME : Supposons qu'on ait m strates, N_i éléments dans la iè strates, $D = \sum_{i=1}^{m} N_i$ éléments en tout, et soit :

$$\bar{\bar{x}}_i = \frac{x_i}{n_i} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

si :

$$\alpha = \sum_{i=1}^{m} \frac{N_i \bar{\bar{x}}_i}{D}$$

alors :

$$\sigma^2(\alpha) = \frac{1}{D^2} \sum_{i=1}^{m} N_i^2 \frac{1-f_i}{n_i} \sum_{j=1}^{N_i} \frac{(X_{ij}-\bar{\bar{X}}_i)^2}{N_i-1} = \frac{1}{D^2} \sum_{i=1}^{m} N_i^2 \frac{1-f_i}{n_i} S_{2iX}^2$$

Or ici :

$$\bar{x} = \sum_{i=1}^{m} \frac{N_i (n_i/N_i) \bar{\bar{x}}_i}{n/D\ D} = \frac{D}{n} \frac{n_i}{N_i} \alpha$$

$$\sigma^2(\bar{x}) = \frac{D^2}{n^2} \left(\frac{n_i}{N_i}\right)^2 \sigma^2(\alpha)$$

$$\sigma^2(x'/1) = \frac{M^2}{m^2} \sigma^2(x'') = \frac{M^2}{m^2} \left(\frac{N_i^2}{n_i^2}\right) n^2 \sigma^2(\bar{x})$$

$$= \frac{M^2}{m^2} \frac{N_i^2}{n_i^2} n^2 \frac{D^2}{n^2} \left(\frac{n_i}{N_i}\right)^2 \frac{1}{D^2} \sum_{i=1}^{m} N_i^2 \frac{1-f_i}{n_i} S_{2iX}^2$$

et finalement :

$$\sigma^2(x'/1) = \frac{M^2}{m^2} \sum_{i=1}^{m} N_i^2 \frac{N_i - n_i}{n_i} S_{2iX}^2$$

où :

$$S_{2iX}^2 = \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{\bar{X}}_i)^2}{N_i - 1}$$

et :

$$\bar{\bar{X}}_i = \frac{X_i}{N_i}$$

Donc :

$$E\,\sigma^2(x'/1) = \frac{M^2}{m^2} E \sum_{i=1}^{m} N_i^2 \frac{N_i - n_i}{n_i N_i} S_{2iX}^2$$

$$= \frac{M^2}{m^2} m\, E\left[N_i^2 \frac{N_i - n_i}{n_i N_i} S_{2iX}^2 \right]$$

$$= \frac{M^2}{m^2} m \frac{1}{M} \sum_{i=1}^{M} N_i^2 \frac{N_i - n_i}{n_i N_i} S_{2iX}^2$$

car la variable aléatoire :

$$\mu_i = N_i^2 \frac{N_i - n_i}{n_i N_i} S_{2iX}^2$$

a M valeurs possibles, chacune avec la probabilité 1/M. Donc :

$$E\,\sigma^2(x'/1) = \frac{M}{m} \sum_{i=1}^{M} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{2iX}^2$$

toujours avec :

$$S_{2iX}^2 = \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{\bar{X}}_i)^2}{N_i - 1}$$

qui n'est rien d'autre que la variance à l'intérieur de la i$^{\text{è}}$ psu. Or, on sait de plus que $n_i/N_i = f_2 = C^{te} \quad \forall i$.

Montrons que $f_2 = \bar{n}/\bar{N}$.

En effet :

$$f_2 = \frac{f_1 f_2}{f_1} = \frac{n}{N} : \frac{m}{M} = \frac{n}{m} \frac{M}{N} = \frac{\bar{n}}{\bar{N}} \qquad (1)$$

Donc, en posant :

$$S_{2X}^2 = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}}_i)^2 = \frac{1}{N} \sum_{i=1}^{M} N_i S_{2iX}^2$$

On a donc :

$$E\,\sigma^2(x'/1) = \frac{M}{m} \sum_{i=1}^{M} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{2iX}^2$$

$$= \frac{M}{m} \sum_{i=1}^{M} N_i \frac{\bar{N}}{\bar{n}} \left(1 - \frac{\bar{n}}{\bar{N}}\right) S_{2iX}^2$$

$$= \frac{M}{m} \frac{\bar{N}}{\bar{n}} \frac{\bar{N} - \bar{n}}{\bar{N}} \sum_{i=1}^{M} N_i S_{2iX}^2$$

Or, d'après (1) $M \frac{\bar{N}}{\bar{n}} = N \frac{m}{n}$ (évident)

$$E\,\sigma^2(x'/1) = \frac{1}{m} N \frac{m}{n} \frac{\bar{N} - \bar{n}}{\bar{N}} N S_{2x}^2$$

$$\boxed{E\,\sigma^2(x'/1) = N^2 \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{S_{2X}^2}{m\bar{n}}}$$

Evaluons maintenant $\sigma^2 E(x'/1)$

$$E(x'/1) = E\left(\frac{M}{m}\sum_{i=1}^{m} x'_i/1\right) = \frac{M}{m} E\left(\sum_{i=1}^{m} N_i \bar{\bar{n}}_i/1\right)$$

$$= \frac{M}{m}\sum_{i=1}^{m} N_i \bar{\bar{X}}'_i \quad \text{car l'ensemble des m psu est fixé et peuvent donc être considérées comme strates}$$

$$= \frac{M}{m}\sum_{i=1}^{m} X'_i \quad \text{les primes étant utilisées pour indiquer que } \bar{\bar{X}}'_i \text{ et } X'_i \text{ sont des v.a.}$$

$$\sigma^2 E(x'/1) = \sigma^2\left(\frac{M}{m}\sum_{i=1}^{m} X'_i\right) = M^2 \frac{M-m}{Mm} \sum_{i=1}^{M} \frac{(X_i - \bar{X})^2}{M-1}$$

d'après les résultats sur les échantillons simples

$$\boxed{\sigma^2 E(x'/1) = M^2 \frac{M-m}{M} \frac{S^2_{1X}}{m}}$$

Donc, finalement, comme

$$\sigma^2_{x'} = E\,\sigma^2(x'/1) + \sigma^2 E(x'/1)$$

on a bien :

$$\boxed{\sigma^2(x') = M^2 \frac{M-m}{M} \frac{S^2_{1X}}{m} + N^2 \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{S^2_{2x}}{m\bar{n}}}$$

$$V^2(x') = \frac{M-m}{M} \frac{M^2}{X^2} \frac{S^2_{1X}}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{N^2}{X^2} \frac{S^2_{2X}}{m\bar{n}}$$

ou

$$\boxed{V^2(x') = \frac{M-m}{M} \frac{B^2_X}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{W^2_X}{m\bar{n}}}$$

avec :

$$S_{1X}^2 = \sum_{i=1}^{M} \frac{(X_i - \bar{X})^2}{M - 1}$$

$$B_X^2 = \frac{S_{1X}^2}{\bar{X}^2}$$

$$S_{2X}^2 = \frac{1}{\bar{N}} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}}_i)^2$$

$$W_x^2 = \frac{S_{2X}^2}{\bar{X}^2}$$

La variance de x' est exprimée en fonction des contributions à la variance de chaque degré d'échantillonage. Le premier terme représente la contribution à la variance de x' due au 1er degré, et est appelé la composante entre psu de la variance. C'est la variance de x' basée sur un échantillon de m psu sans sous échantillonage, et alors la variation à l'intérieur des psu n'intervient pas dans cette expression. On peut le voir en faisant $\bar{n} = \bar{N}$ dans $\sigma^2(x')$ ce qui aurait lieu si les N_i unités élémentaires de la $i^{\text{è}}$ psu selectionnée étaient toutes incluses dans l'échantillon.

Le second terme représente la contribution à la variance de x' due aux unités du second degré et est appelé la composante intra psu de la variance. C'est, exactement, à un facteur près, la variance de x' pour un échantillon stratifié à fraction sondée constante de n éléments, les psu servant de strates. La variation entre psu n'intervient pas dans cette expression. On peut s'en apercevoir en faisant m = M dans $\sigma^2(x')$, ce qui aurait lieu si un échantillon était sélectionné à partir de toutes les psu.

B) ESTIME d'un QUOTIENT

Variance et Variance Relative de l'Estimé

Soit $r = \frac{x'}{y'} = \frac{x}{y}$

On se contentera de calculer V_r^2

car $\sigma_r^2 = V_r^2 \cdot E(r)^2 = V_r^2 \cdot r^2$ si r est sans biais. On montre qu'une approximation satisfaisante de V_r^2 est donnée par :

$$V_r^2 = V_{x'}^2 + V_{y'}^2 - 2V\, x'\, y'$$

avec :

$$V_{x'}^2 = \frac{\sigma_{x'}^2}{X^2} \quad V_{y'}^2 = \frac{\sigma_{y'}^2}{Y^2} \quad V_{x'y'} = \frac{\sigma_{x'y'}}{XY}$$

D'après la partie A on connait $V_{x'}^2$ et $V_{y'}^2$. Reste à calculer $V_{x'y'}$, ou $\sigma(x'y')$ qui représente la covariance de x' et y'

$$\sigma(x'\, y') = E(x' - X)\,(y' - Y)$$

En utilisant le théorème :

$$\sigma_{uv} = E\,\sigma(uv/b^*) + \sigma\left[E(u/b^*)\,E(v/b^*)\right]$$

et en faisant une démonstration analogue à celle faite dans la partie A pour le calcul de $\sigma_{x'}^2$, on démontre que :

$$\sigma_{(x'y')} = \frac{M^2}{m}\,\frac{M-m}{M}\,S_{1XY} + \frac{M}{m}\sum_{i=1}^{M}\frac{N_i^2}{n_i}\,\frac{N_i - n_i}{N_i}\,S_{2iXY}$$

où

$$S_{1XY} = \sum_{i=1}^{M}\frac{(X_i - \bar{X})(Y_i - \bar{Y})}{M-1}$$

et

$$S_{2iXY} = \sum_{j=1}^{N_i}\frac{(X_{ij} - \bar{\bar{X}}_i)(Y_{ij} - \bar{\bar{Y}}_i)}{N_i - 1}$$

Si on fait X = Y dans $\sigma(x'y')$ on retrouve $\sigma^2(x')$. Si n_i/N_i = constante = $\bar{n}/\bar{N}$

$$\boxed{\sigma(x'y') = M^2\,\frac{M-m}{Mm}\,S_{iXY} + N^2\,\frac{\bar{N}-\bar{n}}{\bar{N}m\bar{n}}\,S_{2XY}}$$

où S_{1XY} et S_{2XY} sont définis plus loin et :

$$V_{x'y'} = \frac{\sigma(x'y')}{X^2\,Y^2} = \frac{M-m}{Mm}\,\frac{S_{iXY}}{\bar{X}\bar{Y}} + \frac{\bar{N}-\bar{n}}{\bar{N}m\bar{n}}\,\frac{S_{2XY}}{\bar{X}\bar{Y}}$$

et

$$\boxed{V_{x'y'} = \frac{M - m}{M} \frac{B_{XY}}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W_{XY}}{m\bar{n}}}$$

où B_{XY} et W_{XY} sont définis plus loin .

NOTATIONS

$$S_{1XY} = \sum_{i=1}^{M} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{M - 1}$$

$$S_{2XY} = \frac{1}{\bar{N}} \sum_{i=1}^{M} N_i \, S_{2iXY} = \frac{1}{\bar{N}} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}}_i) (Y_{ij} - \bar{\bar{Y}}_i)$$

$$S_{1X}^2 = \sum_{i=1}^{M} \frac{(X_i - \bar{X})^2}{M - 1}$$

$$S_{2X}^2 = \frac{1}{\bar{N}} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}}_i)^2$$

$$B_X^2 = \frac{S_{1X}^2}{\bar{X}^2} \qquad W_X^2 = \frac{S_{2X}^2}{\bar{\bar{X}}^2}$$

$$B_{XY} = \frac{S_{1XY}}{\bar{X}\bar{Y}} \qquad W_{XY} = \frac{S_{2XY}}{\bar{\bar{X}}\bar{\bar{Y}}}$$

$$S_1^2 = S_{1X}^2 + R^2 \, S_{1Y}^2 - 2R \, S_{1XY}$$

$$S_2^2 = S_{2X}^2 + R^2 \, S_{2Y}^2 - 2R \, S_{2XY}$$

$$B^2 = B_X^2 + B_Y^2 - 2 \, B_{XY}$$

$$W^2 = W_X^2 + W_Y^2 - 2 \, W_{XY}$$

Compte tenu de ces notations, et comme il a été vu plus haut :

$$V_{x'}^2 = \frac{M - m}{M} \frac{B_X^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W_X^2}{m\bar{n}}$$

$$V^2_{x'} = \frac{M - m}{M} \frac{B^2_Y}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W^2_Y}{m\bar{n}}$$

$$2V_{x'y'} = 2 \frac{M - m}{M} \frac{B_{XY}}{m} + 2 \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W_{XY}}{m\bar{n}}$$

Comme :

$$V^2_r = V^2_{x'} + V^2_{y'} - 2 V_{x'y'}$$

$$V^2_r = \frac{M - m}{M} \frac{B^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}}$$

où le premier terme est dû au 1er degré d'échantillonage seulement, et le 2è terme est dû au 2è degré d'échantillonage.

IIIè PARTIE

Bien que la décision d'échantillonner des unités de liste au hasard ne soit pas un choix pratique dans les problèmes courants, l'échantillonage simple au hasard fournit un moyen de sélectionner des choix raisonnables, et de discuter des variances dues à ces différents choix. Pour cette raison il sera souvent pratique d'exprimer la variance d'échantillons en grappes en fonction de la variance des échantillons au hasard affectés des coefficients reflétant l'effet des l'échantillonage en grappes sur la variance.

A) APPROXIMATION de la VARIANCE RELATIVE

Nous avions vu dans la 2ème partie que :

$$V_r^2 = \frac{M - m}{M} \frac{B^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}} \qquad (0)$$

Soit, si m est relativement faible devant M

$$V_r^2 = \frac{M - 1}{M} \frac{B^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}} \qquad (1)$$

Posons :

$$\hat{V}^2 = \frac{M - 1}{M} B^2 + \frac{\bar{N} - 1}{\bar{N}} W^2 \qquad (2)$$

et :

$$\delta = \frac{(M-1)/M\ B^2 - \hat{V}^2/N}{(\bar{N}-1)\ V^2/\bar{N}} \qquad (3)$$

de (3) on tire :

$$\frac{M - 1}{M} B^2 = \frac{\hat{V}^2}{\bar{N}} \left[1 + \delta(\bar{N}-1)\right] \qquad (4)$$

qui, à un coefficient près, figure dans l'expression de V_r^2 que nous voulons établir . Ajoutons $\frac{\bar{N}-1}{\bar{N}} W^2$ aux deux membres de (4). En tenant compte de (2) il vient :

$$\hat{V}^2 = \frac{\hat{V}^2}{\bar{N}} \left[1 + \delta(\bar{N}-1)\right] + \frac{\bar{N} - 1}{\bar{N}} W^2$$

$$\frac{\bar{N}-1}{\bar{N}} W^2 = \frac{\hat{V}^2}{\bar{N}} \left[\bar{N} - 1 - \delta N + \delta\right]$$

$$W^2 = \frac{\bar{N}}{\bar{N}-1} \hat{V}^2 \frac{\bar{N}-1}{\bar{N}} (1-\delta)$$

$$W^2 = \hat{V}^2 (1-\delta) \qquad (5)$$

Substituons (4) et (5) dans (1)

$$V_r^2 = \frac{M-1}{M} B^2 + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{W^2}{m\bar{n}}$$

$$V_r^2 = \frac{\hat{V}^2}{m\bar{N}} \left[1 + \delta(N-1)\right] + \frac{\bar{N}-\bar{n}}{\bar{N}} \frac{1}{m\bar{n}} \hat{V}^2(1-\delta)$$

$$V_r^2 = \frac{\hat{V}^2}{m\bar{n}} \left[\frac{\bar{n} + \bar{n}\bar{N}\delta - \bar{n}\delta + \bar{N} - \bar{N}\delta - \bar{n} + \bar{n}\delta}{\bar{N}}\right]$$

$$V_r^2 = \frac{\hat{V}^2}{m\bar{n}} \left[1 + \delta(\bar{n}-1)\right] \qquad (6)$$

On montre que :

$$V_r^2 = \frac{1-f}{m\bar{n}} \hat{V}^2 \left[1 + \delta(\bar{n}-1)\right] \qquad (7)$$

est une meilleure approximation de l'équation (0)

En effet :

$$eq(7) = eq(0) = \frac{B^2}{M} \frac{m-1}{m} \qquad (8)$$

$$eq(6) - eq(0) = \frac{\bar{N}-\bar{n}}{\bar{N}M} \left(B^2 - \frac{W^2}{\bar{N}}\right) - \frac{1-f}{m-1} \left(\frac{B^2}{M} - \frac{m-1}{m}\right) \qquad (9)$$

et en valeur absolue, l'expression (9) est plus grande que l'expression (8), pourvu que $m > 1$ et $(B^2 - W^2/\bar{N}) > 0$, ce qui est communément le cas.

B) APPLICATION à l'ETUDE d'une MOYENNE par UNITE de LISTE

On cherche à estimer la valeur moyenne par unité élémentaire c'est à dire $\bar{\bar{X}} = X/N$. Nous supposerons que les psu sont de taille égale i.e. $N_i = \bar{N}$

(si bien que chaque psu contient $\bar{N} = N/M$ unités élémentaires).

Partant de l'expression :

$$V_r^2 = \frac{1-f}{m\bar{n}} \hat{V}^2 \left[1 + \delta(\bar{n} - 1)\right]$$

nous allons établir que :

$$V_{\bar{\bar{x}}}^2 = \frac{1-f}{m\bar{n}} V^2 \left[1 + \delta(\bar{n} - 1)\right]$$

avec :

$$V^2 = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{\bar{X}})^2}{N \bar{\bar{X}}^2}$$

$$\delta = \frac{\sigma_b^2 - \sigma^2/\bar{N}}{(\bar{N}-1)\sigma^2/\bar{N}}$$

$$\sigma_b^2 = \frac{\sum_{i=1}^{M} (\bar{\bar{X}}_i - \bar{\bar{X}})^2}{M}$$

$$\sigma^2 = \frac{\sum_{i=1}^{M} \sum_{j=1}^{\bar{N}} (X_{ij} - \bar{\bar{X}})^2}{N}$$

Or ici :

$$r = \bar{\bar{x}} = \frac{x}{y} = \frac{x}{m\bar{n}}$$

avec :

$$x = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_{ij} \qquad y = m\,\bar{n}$$

$N_i = \bar{N} = N/M$ signifie qu'il y a le même nombre d'éléments dans chaque grappe de la population. Donc comme $n_i = f_2 N_i$ n_i = constante :

1°) $Y_i = \bar{Y}$ car $Y_i = N_i = \bar{N}$ et $\bar{Y} = \bar{N}$

2°) De plus $Y_{ij} = 1$ et $\bar{\bar{Y}}_i = Y_i/N_i$ car $Y_i = N_i$

Or :

$$\hat{V}^2 = \frac{M-1}{M} B^2 + \frac{\bar{N}-1}{N} W^2$$

compte tenu des expressions de B^2 et W^2 données à la fin de la 2ème partie

$$\hat{V}^2 = \frac{M-1}{M}\left[\sum_{i=1}^{M}\frac{(X_i-\bar{X})^2}{(M-1)\,\bar{X}^2}+\sum_{i=1}^{M}\frac{(Y_i-\bar{Y})^2}{(M-1)\,\bar{Y}^2}-2\sum_{i=1}^{M}\frac{(X_i-\bar{X})(Y_i-\bar{Y})}{(M-1)\,\bar{X}\,\bar{Y}}\right]$$

$$+\frac{\bar{N}-1}{\bar{N}}\left[\frac{1}{\bar{N}}\sum_{i=1}^{M}\frac{N_i}{N_i-1}\sum_{j=1}^{N_i}\frac{(X_{ij}-\bar{\bar{X}}_i)^2}{\bar{\bar{X}}^2}+\frac{1}{\bar{N}}\sum_{i=1}^{M}\frac{N_i}{N_i-1}\sum_{j=1}^{N_i}\frac{(Y_{ij}-\bar{\bar{Y}}_i)^2}{\bar{\bar{Y}}^2}\right.$$

$$\left.-2\,\frac{1}{\bar{N}}\sum_{i=1}^{M}\frac{N_i}{N_i-1}\sum_{j=1}^{N_i}\frac{(X_{ij}-\bar{\bar{X}}_i)(Y_{ij}-\bar{\bar{Y}}_i)}{\bar{\bar{X}}\,\bar{\bar{Y}}}\right]$$

Après simplifications, et compte tenu du fait que :

$$Y_i = \bar{Y} \qquad Y_{ij} = \bar{\bar{Y}}_i \qquad N_i = \bar{N}$$

$$\hat{V}^2 = \frac{1}{M}\sum_{i=1}^{M}\frac{(X_i-\bar{X})^2}{\bar{X}^2}+\frac{1}{\bar{N}}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\frac{(X_{ij}-\bar{\bar{X}}_i)^2}{\bar{\bar{X}}^2}$$

en divisant le ler terme par N_i^2 au numérateur et au dénominateur :

$$\hat{V}^2 = \frac{1}{M}\sum_{i=1}^{M}\frac{(\bar{\bar{X}}_i-\bar{\bar{X}})^2}{\bar{\bar{X}}^2}+\frac{1}{\bar{N}}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\frac{(X_{ij}-\bar{\bar{X}}_i)^2}{\bar{\bar{X}}^2}$$

or :

$$(X_{ij}-\bar{\bar{X}}_i)^2 = (X_{ij}-\bar{\bar{X}})^2+(\bar{\bar{X}}-\bar{\bar{X}}_i)^2+2(X_{ij}-\bar{\bar{X}})(\bar{\bar{X}}-\bar{\bar{X}}_i)$$

$$\sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}}_i)^2 = \sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}})^2+N_i(\bar{\bar{X}}-\bar{\bar{X}}_i)^2+2(\bar{\bar{X}}-\bar{\bar{X}}_i)(X_i-N_i\,\bar{\bar{X}})$$

$$\sum_{i=1}^{M}\sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}}_i)^2 = \sum_{i=1}^{M}\sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}})^2+N_i\sum_{i=1}^{M}(\bar{\bar{X}}-\bar{\bar{X}}_i)^2+2\sum_{i=1}^{M}(\bar{\bar{X}}-\bar{\bar{X}}_i)(X_i-N_i\bar{\bar{X}})$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}})^2+N_i\sum_{i=1}^{M}(\bar{\bar{X}}-\bar{\bar{X}}_i)^2-2N_i\sum_{i=1}^{M}(\bar{\bar{X}}-\bar{\bar{X}}_i)^2$$

$$= \sum_{i=1}^{M}\sum_{j=1}^{N_i}(X_{ij}-\bar{\bar{X}})^2-N_i\sum_{i=1}^{M}(\bar{\bar{X}}-\bar{\bar{X}}_i)^2$$

En tenant compte du fait que dans l'expression $\hat{V}^2$ on a $\bar{N} = +N_i$ et en remplaçant on obtient :

$$\hat{V}^2 = \sum_{i=1}^{M}\sum_{j=1}^{N_i}\frac{(X_{ij}-\bar{\bar{X}})^2}{N\,\bar{\bar{X}}^2} = V^2$$

$$\delta = \frac{(M-1)/M\ B^2 - \hat{V}^2/N}{(\bar{N}-1)\ \hat{V}^2/\bar{N}}$$

comme on l'a vu plus haut

$$\frac{M-1}{M} B^2 \text{ se réduit à } \frac{M-1}{M} \sum_{i=1}^{M} \frac{(X_i - \bar{X})^2}{(M-1)\ \bar{X}^2}$$

Donc :

$$\frac{M-1}{M} B^2 = \sum_{i=1}^{M} \frac{(X_i - \bar{X})^2}{M\ \bar{X}^2} = \sum_{i=1}^{M} \frac{(X_i/N_i - \bar{X}/N_i)^2}{M(\bar{X}/N_i)^2}$$

$$\frac{M-1}{M} B^2 = \frac{\sigma_b^2}{\bar{\bar{X}}^2} = \sum_{i=1}^{M} \frac{(\bar{\bar{X}}_i - \bar{\bar{X}})^2}{M\ \bar{\bar{X}}^2}$$

$$\frac{\hat{V}^2}{\bar{N}} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}})^2}{\bar{N}\ N\ \bar{\bar{X}}^2} = \frac{M}{\bar{N}} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{\bar{X}})^2}{N\ \bar{\bar{X}}^2}$$

$$\frac{\hat{V}^2}{\bar{N}} = \frac{M}{\bar{N}} \frac{\sigma^2}{\bar{\bar{X}}^2} = \frac{\sigma^2}{\bar{N}\ \bar{\bar{X}}^2}$$

D'où :

$$\delta = \frac{\sigma_b^2/\bar{\bar{X}}^2 - \sigma^2/\bar{N}^2\bar{\bar{X}}^2}{(\bar{N}-1)\ \sigma^2/\bar{N}}$$

$$\delta = \frac{\sigma_b^2 - \sigma^2/\bar{N}}{(\bar{N}-1)\sigma^2/\bar{N}}$$

Récapitulons :

$$V_{\bar{\bar{x}}}^2 = \frac{1-f}{m\ \bar{n}}\ V^2 \left[1 + \delta(\bar{n} - 1)\right]$$

$$V^2 = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \frac{(X_{ij} - \bar{\bar{X}})^2}{N\ \bar{\bar{X}}^2}$$

$$\delta = \frac{\sigma_b^2 - \sigma^2/\bar{N}}{(\bar{N}-1)\sigma^2/\bar{N}}$$

$$\sigma_b^2 = \frac{\sum_{i=1}^{M} (\bar{\bar{X}}_i - \bar{\bar{X}})^2}{M}$$

$$\sigma_W^2 = \frac{\sum_{i=1}^{M} \sum_{j=1}^{\bar{N}_i} (X_{ij} - \bar{\bar{X}}_i)^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{M} \sum_{j=1}^{\bar{N}} (X_{ij} - \bar{\bar{X}})^2}{N}$$

On démontre aisément que $\sigma^2 = \sigma_b^2 + \sigma_W^2$

a) Si $\sigma_W^2 = 0$ i.e. si $X_{ij} = X_i/N_i$, alors tous les éléments à l'intérieur d'une grappe sont semblables du point de vue de la caractéristique étudiée ; par suite il n'y a pas de variation entre unités à l'intérieur des grappes, i.e. il y a homogénéité..

Dans ce cas $\sigma^2 = \sigma_b^2$ et $\delta = 1$

b) si $\sigma_b^2 = 0$ i.e. si $\bar{\bar{X}} = X_i/N_i$ alors les moyennes de toutes les grappes sont identiques, alors $\sigma^2 = \sigma_W^2$ i.e. σ_W^2 est élevé, et il y a donc hétérogénéité à l'intérieur de chaque psu.

Dans ce cas $\delta = \frac{-1}{\bar{N} - 1}$

On vient donc de voir que la corrélation interclasse δ est voisine de 1 si la contribution σ_b^2 entre grappes est forte dans la variance σ^2 et que δ sera faible voire même négative si σ_b^2 est faible.

Ainsi, δ mesure le degré d'homogénéité ou d'hétérogénéité à l'intérieur des grappes. Quand les unités élémentaires à l'intérieur des grappes sont relativement homogènes i.e. quand elles sont semblables du point de vue d'une caractéristique, la corrélation interclasse entre unités pour cette caractéristique sera élevé. Réciproquement, si les unités élémentaires à l'intérieur des grappes sont relativement hétérogènes, la corrélation interclasse sera faible ou même négative.

Or $V_{\bar{\bar{x}}}^2$ mesure en quelque sorte la précision de l'estimé $\bar{\bar{x}}$ de $\bar{\bar{X}}$. Le nombre total d'unités élémentaires dans l'échantillon est $m\bar{n}$ et $\frac{1-f}{m\bar{n}} V^2$ est la variance de l'estimé dans un échantillon au hasard de $m\bar{n}$ unités élémentaires. Cela entraîne que $1 + \delta(\bar{n}-1)$ représente dans $V_{\bar{\bar{x}}}^2$ le facteur par elquel la variance relative d'un échantillon au hasard doit être multipliée pour obtenir

la variance relative due à l'échantillonage en grappes (toujours pour des grappes de même taille). Donc si $\bar{n} = 1$, i.e. on tire un élément dans chaque grappe, $1 + \delta(\bar{n}-1) = 1$ et la variance relative est la même pour un échantillon au hasard et pour un échantillon en grappes. De plus, si $\delta > 0$, tout augmentation de $\bar{n}$ augmente la variance relative dans le cas de l'échantillonage en grappes, d'où diminution de précision. Donc, dans le cas de grappes homogènes, i.e. $\delta > 0$, on aura plutôt intérêt à utiliser un échantillon au hasard. Inversement, dans le cas de grappes hétérogènes, i.e. $\delta \leq 0$, on aura plutôt intérêt à utiliser l'échantillonage en grappes car $V^2_{\bar{x}}$ diminuera, i.e. la précision sera accrue.

Remarque 1 : illustration

Supposons que l'on veuille évaluer le nombre moyen d'habitants par appartement dans une ville. On divise la ville en quartiers qui seront les grappes. Les appartements seront les unités de liste.

12			14
	1	2	
	3	4	
13			15

1er cas : Si l'échantillon ne comporte que 4 grappes, et si on tire les grappes 1, 2, 3, 4 : ce sont des quartiers centraux, où sont situés les bureaux, etc... où le nombre d'habitants par appartement est faible. Ces grappes sont homogènes, et pourtant on aura une sous estimation du nombre d'habitants par appartement.

2ème cas : On tire les grappes 12, 13, 14, 15 : ce sont des quartiers périphériques donc populeux. Ces grappes sont homogènes et pourtant on aura une surestimation du nombre d'habitants par appartement.

Remarque 2 :

On pourrait confondre grappes et strates. De fait, les grappes sont des strates particulières car on a dit qu'un individu appartient à une grappe et une seule, et que chaque individu appartient à une grappe. Aussi, on pourrait se demander si, lorsque δ est grand i.e. les grappes sont homogènes, i.e. l'échantillon simple est meilleur que l'échantillon en grappes, on ne pourrait pas choisir un autre système de grappes telles que δ soit faible et par suite l'échantillon en grappes soit meilleur que l'échantillon simple.

Mais en général, le choix des grappes est systématique et s'impose de lui-même (par exemple il correspond à des divisions géographiques pour que le coût de l'enquête soit moindre) alors que la stratification est définie par une relation d'équivalence qui fait que les éléments peuvent être dispersés. Aussi, si on transformait les grappes en strates, ce qu'on gagnerait en précision en diminuant ainsi δ, serait largement perdu du point de bue du coût.

C) GENERALISATION

Revenons au cas général où :

$$V_r^2 = \frac{1 - f}{m\, n} \hat{V}^2 \left[1 + \delta(\bar{n} - 1)\right]$$

que nous écrirons :

$$V_r^2 = \left(\frac{1 - f}{m\, \bar{n}} V^2\right) \frac{\hat{V}^2}{V^2} \left[1 + \delta(\bar{n} - 1)\right]$$

où

$$V^2 = V_X^2 + V_Y^2 - 2\, V_{XY}$$

avec :

$$V_{XY} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}})(Y_{ij} - \bar{\bar{Y}})}{(N - 1)\, \bar{\bar{X}}\, \bar{\bar{Y}}}$$

et :

$$V_X^2 = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} (X_{ij} - \bar{\bar{X}})^2}{(N-1)\, \bar{\bar{X}}^2}$$

On cherche à estimer $R = X/Y$ à l'aide de $r = x/y$.

V_r^2 est composé de trois facteurs :

a) Le facteur $\frac{1 - f}{n \bar{n}} V^2$ est la variance relative due à un échantillon simple, comme dans le cas B. Cependance, ici, il représente la variance relative d'un quotient de variables aléatoires, alors que plus haut, il représentait la variance relative d'une moyenne par unité de liste où le dénominateur n'était pas une variable aléatoire.

b) Le second facteur $\hat{V}^2/V^2$ n'apparaissait pas dans le cas particulier. Il est du au fait que maintenant le nombre d'unités de liste par psu est variable. Dans le cas particulier où $N_i = \bar{N}$, il est alors égal à 1. Habituellement $\hat{V}^2$ est supérieur à V^2, bien que, si les totaux X_i et Y_i, pour chaque grappe, tendent à être proportionnels aux N_i, alors la variation de taille des psu, n'augment pas beaucoup $\hat{V}^2$ par rapport à V^2. Aussi, $\hat{V}^2$ et V^2 seront voisins dans ce nombreux cas. Mais si X_i est proportionnel à N_i, et qu'il n'en est pas de même pour Y_i, alors $\hat{V}^2/V^2$ peut être nettement supérieur à 1, et il peut donc y avoir un grand désavantage à utiliser de telles grappes dans la sélection de l'échantillon, quelle que soit la taille moyenne des "ultimate clusters". C'est ce qui arrive en général lorsqu'on cherche un estimé non biaisé d'un total.

Le troisième facteur $1 + \delta(\bar{n}-1)$ est comparable au même terme étudié plus haut. δ mesure ici aussi l'homogénéité des grappes. Comme dans le cas précédent, si les grappes sont homogènes i.e. $\delta \simeq 1$, comme $\bar{n}$ est supérieur à 1, V_r^2 augmente, donc la précision diminue ; on a donc intérêt à prendre un échantillon simple au hasard.

Si les grappes sont hétérogènes, i.e. δ faible voir même négatif, V_r^2 sera faible, donc la précision accrue, et par suite on a intérêt à prendre un échantillon en grappes.

IVè PARTIE

A) Fonctions de coût simples.

Pour étudier le coût d'une enquête entraînant l'usage de l'échantillonage en grappes, il est nécessaire de déterminer les différentes phases de l'enquête, et de déterminer approximativement :

a) Les fraix généraux, i.e. ces frais indépendants de la manière dont la taille de l'échantillon ou l'échantillon lui même sont choisis.

b) Les frais qui dépendent d'abord du nombre d'unités primaires incluses dans l'échantillon et ensuite de la variation de ces frais en fonction de la variation du nombre de ces unités primaires.

c)Les frais qui dépendent d'abord du nombre d'unités de liste incluses dans l'échantillon, et ensuite de la variation de ces frais en fonction de la variation du nombre d'unités de liste de l'échantillon.

Supposons par exemple, q'un total de 5000 dollars soit à notre disposition pour une enquête sur les loyers des habitations dans un comté, et que de petites zônes soient les unités primaires, avec les habitations comme unités de liste, et que les frais s'élèvent de la façon décrite ci-dessous :

a) Frais totaux et frais fixes : le coût total C qui intervient dans la fonction de coût dépend seulement des coûts variables, d'où sont exclus les frais généraux constants quelle que soit l'enquête. Les frais dus au travail du centre administratif et technique, et les frais d'équipement peuvent être à peu près identiques même lorsqu'il y a des variations dans la taille de l'échantillon. Nous supposerons que, dans le cas particulièr traité les frais généraux invariants (fixes) sont estimés à 1500 dollars. Ceci réduit les fonds disponibles pour les frais variables à 5000 - 1500 = 300 dollars de telle sorte que nous avons :

$$C = 3500 \text{ dollars.}$$

b) Frais variant proportionnellement au nombre des psu dans l'échantillon.

Certains autres frais peuvent varier directement en fonction du nombre de psu incluses dans l'échantillon. Ces dépenses peuvent comprendre les frais dus au choix, au repérage des psu, aux déplacements, à la préparation d'une liste des habitations, et au travail de sous échantillonage. Disons que le coût de telles opérations est égal à C_1 par psu incluse dans l'échantillon, e et que, en nous basant sur une enquête préalable, nous pouvons estimer, dans le cas particulier présente, C_1 à 2 dollars. Ceci signifie que nous estimons que les frais d'adjonction d'une psu à l'échantillon serait de 2 dollars, sans compter les frais dus à l'interrogation et à la répertoration des données pour les familles qui peuvent être incluses dans l'échantillon issu de cette psu.

c) Frais variant proportionnellement au nombre d'unités de liste de l'échantillon.

D'autres frais sont ceux qui peuvent dépendre directement du nombre total d'unités de liste incluses dans l'échantillon. Ici, on peut inclure les frais dus aux interviews et aux examens des résultats. C_2 représentera le coût de telles opérations par unité de liste, coût qui varie directement avec le nombre d'unités de liste incluses dans l'échantillon.

Nous supposerons, pour notre cas particulier, qu'après une enquête préalable, nous avons estimé C_2 à 1 dollar.

Fonction de coût totale (cas simple) :

Si nous trouvons dans la situation simple où tous les frais de l'enquête peuvent être représentés à peu près par les types de dépenses décrits plus haut, alors C, le coût total attendu de l'enquête, sera donné par :

$$C = C_1 m + C_2 m \bar{n}$$

où m = nombre de psu sélectionnées

$m\bar{n} = n$ = nombre d'unités de liste.

Dans notre exemple :

$$C = 3500 \quad C_1 = 2 \quad C_2 = 1$$

et la fonction de coût devient :

$$3500 = 2m + m\bar{n} \qquad (1)$$

B) Utilisation des fonctions de coût pour la recherche des valeurs optimum de m et $\bar{n}$ dans le cas d'une fonction de coût simple :

Le problème consiste à déterminer les valeurs de m et $\bar{n}$ qui fournissent l'exactitude requise à moindres frais, ou celles qui fournissent la meilleure précision avec les fonds et ressources dont on dispose. On cherchera donc à minimiser la variance relative.

Une fois les unités primaires définies, nous sommes en mesure de déterminer approximativement la fraction optimum d'échantillonage du second degré, f_2, et le nombre optimum de psu à inclure dans l'échantillon en vue de l'estimation d'une certaine caractéristique.

Pour déterminer la valeur optimum de f_2, nous déterminerons d'abord la valeur optimum de $\bar{n}$, i.e. le nombre moyen d'unités élémentaires par psu et alors :

$$\text{optimum } f_2 = \frac{\text{opt. } \bar{n}}{\bar{N}}$$

Considérons à nouveau le problème de l'estimation de la valeur moyenne de l'estimation de la valeur moyenne des loyers. Nous pouvons déterminer à l'aide de l'équation (1) $3500 = 2m + m\bar{n} = m(2+\bar{n})$ tous les couples $(m,\bar{n})$ qui entraînent un coput total donné, et tirer de ces couples celui qui fournira la plus petite variance relative que nous prendrons égale à, en négligeant (1-f),

$$V_r^2 = \frac{\hat{V}^2}{m\bar{n}} \left[1 + \delta(\bar{n} - 1)\right]$$

Supposons $\delta = 0{,}25$ $\quad \hat{V}^2 = 1$

alors:

$$V_r^2 = \frac{1}{m\bar{n}} \left[1 + 0{,}25\ (\bar{n} - 1)\right]$$

Le problème consiste à minimiser V_r^2 en tenant compte de la contrainte (1).
Or $m = \frac{3500}{2+\bar{n}}$

$$V_r^2 = \frac{2 + \bar{n}}{3500\bar{n}} \left[1 + 0{,}25\ (\bar{n} - 1)\right]$$

D'où le tableau :

m	$\bar{n}$	V_r^2
1167	1	0,000857
875	2	0,000714
700	3	0,000714
583	4	0,000750
500	5	0,000800
350	8	0,000982
292	10	0,001113

On s'aperçoit que opt $\bar{n} = 3$ et opt $m = 700$. On prendra donc trois unités de liste par psu.

Formule explicite optimum $\bar{n}$ pour une fonction de coût simple :

Il s'agit de minimiser :

$$V_r^2 = (1 - \frac{m}{M}) \frac{B^2}{m} + (1 - \frac{\bar{n}}{\bar{N}}) \frac{W^2}{m\bar{n}}$$

En tenant compte de la relation $C = C_1\ m + C_2\ m\bar{n}$.

Considérons la forme Lagrangienne :

$$F = V_r^2 + \lambda(C_1\ m + C_2\ m\bar{n} - C)$$

Alors, opt $\bar{n}$ et opt m sont donnés par les équations $\frac{\partial F}{\partial m} = 0$ et $\frac{\partial F}{\partial \bar{n}} = 0$

$$\frac{\partial F}{\partial m} = - \frac{B^2}{m^2} - \frac{W^2}{m^2\ \bar{n}} + \frac{W^2}{m^2\ \bar{N}} + \lambda(C_1 + C_2\ \bar{n}) = 0 \qquad (2)$$

$$\frac{\partial F}{\partial n} = - \frac{W^2}{m\ \bar{n}^2} + \lambda\ C_2\ m = 0 \qquad (3)$$

Multiplions l'équation (2) par m :

$$-\frac{B^2}{m} - \frac{W^2}{mn} + \frac{W^2}{m\bar{N}} + \lambda (C_1 m + C_2 m\bar{n}) = 0$$

Multiplions l'équation (3) par $\bar{n}$:

$$-\frac{W^2}{mn} + \lambda C_2 m \bar{n} = 0$$

Soustrayons (3) de (2) :

$$\lambda C_1 m = \frac{B^2}{m} - \frac{W^2}{m\bar{N}}$$

$$\lambda m^2 = \frac{B^2 - W^2/\bar{N}}{C_1} \qquad (4)$$

De l'équation (3) on tire $\lambda m^2 = \dfrac{W^2}{C_2 \bar{n}^2}$ (5)

Comparons (4) et (5) :

$$\bar{n}^2 = \frac{C_1}{C_2} \frac{W^2}{B^2 - W^2/\bar{N}}$$

Finalement on obtient :

$$\text{opt } \bar{n} = \sqrt{\frac{C_1}{C_2} \frac{W^2}{B^2 - W^2/\bar{N}}} = \sqrt{\frac{C_1}{C_2} \frac{1-\delta}{\delta}}$$

et

$$\text{opt } m = \frac{C}{C_1 + C_2 \bar{n}}$$
