PANEL SOCIO-ECONOMIQUE

"LIEWEN ZU LETZEBUERG"

Document PSELL N° 25

DISPOSITIF DES PONDERATIONS INDIVIDUELLES ET DES PONDERATIONS DES MENAGES de 1985 à 1987

B. Gailly
P. Hausman

Document produit par le

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

C.E.P.S./INSTEAD

B.P. 65 L-7201 Walferdange
Tél. (352) 33 32 33 - 1

Président: Gaston Schaber

---

1 9 9 0

PANEL SOCIO-ECONOMIQUE

"LIEWEN ZU LETZEBUERG"

Document PSELL N° 25

DISPOSITIF DES PONDERATIONS INDIVIDUELLES ET DES PONDERATIONS DES MENAGES de 1985 à 1987

B. Gailly
P. Hausman

Document produit par le

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

C.E.P.S./INSTEAD

B.P. 65 L-7201 Walferdange
Tél. (352) 33 32 33 - 1

Président: Gaston Schaber

---

1 9 9 0

DISPOSITIF DES PONDÉRATIONS INDIVIDUELLES ET DES PONDÉRATIONS DES MÉNAGES de 1985 à 1987

Ce document s'adresse aux utilisateurs des données collectées par le Panel Socio-Economique Luxembourgeois et à tous ceux qui seront appelés à pondérer les échantillons successifs de cette étude.

Il a été réalisé grâce à la collaboration de A. Wagner, S. Breulheid, R. De Wever, P. Hausman, F. Hentges, A. Kerger, G. Schmaus et J.J. Wester.

Nos remerciements s'adressent particulièrement à Greg DUNCAN. Il n'a ménagé ni son temps, ni sa peine, pour nous introduire dans les arcanes du dispositif de pondération du P.S.I.D. (Ann Arbor, Michigan). A quelques détails près, les techniques de pondération appliquées dans le P.S.E.L.L. sont empruntées au nouveau dispositif adopté par le P.S.I.D.

DISPOSITIF DES PONDERATIONS INDIVIDUELLES ET DES PONDERATIONS DES MENAGES de 1985 à 1987

Après un bref rappel des différentes raisons qui justifient l'emploi de facteurs de pondération dans l'analyse des données, ce document présente l'organisation du fichier dans lequel les principales variables de pondération sont répertoriées.

La troisième partie du document est consacrée à la présentation de ces variables. Elles sont décrites systématiquement sous trois aspects: leur fonction, leur contenu et les catégories de personnes éligibles.

Ces variables ont des fonctions différentes dans l'ensemble du système de pondération. Les unes ont un rôle "opérationnel": elles peuvent être utilisées pour corriger les échantillons. Les autres sont des variables "instrumentales" qui ont permis ou permettront de calculer des variables "opérationnelles".

Le "contenu" de ces variables correspond à leur utilité dans le système de pondération; il résulte des opérations qui ont permis de les calculer.

Les catégories de personnes "éligibles" définissent les limites de la pertinence de chaque variable.

Le fichier individuel longitudinal réunit l'ensemble des variables de pondération et l'ensemble des personnes interrogées au moins une fois dans le cadre du panel. La sélection des sous-échantillons détermine le choix des variables de pondération.

Cette version a été structurée de telle façon que les versions suivantes, apportant des éléments de documentation complémentaires ou

Chapitre 1

DEUX RAISONS DE PONDERER L'ECHANTILLON

Le Panel Socio-Economique Luxembourgeois (PSELL) construit une base de données individuelles temporelles. Ces données décrivent l'état et l'évolution des conditions d'existence des personnes et des ménages résidant au Luxembourg.

La population de référence de l'étude

- **inclut**
  - tous les individus résidant au Luxembourg et bénéficiaires de la Sécurité Sociale ou de la Protection sociale,
  - tous les individus membres du ménage de ces bénéficiaires de la Sécurité sociale ou de la Protection sociale à condition qu'ils soient:
    - présents dans le ménage au moment de l'enquête
    - absents temporairement au moment de l'enquête (hospitalisation, prison,...);
- **exclut**
  - les individus bénéficiaires de la Sécurité sociale luxembourgeoise, mais résidant à l'étranger
  - les résidents étrangers non assurés à la Sécurité sociale luxembourgeoise et non bénéficiaires de la Protection sociale luxembourgeoise
  - les individus installés définitivement dans des ménages collectifs.

Cette base de données offre le choix entre plusieurs unités d'analyse. Les personnes et les ménages seront sans doute les unités les plus couramment utilisées mais la base de données permet de saisir également d'autres unités d'analyse.

Ces unités sont observées annuellement et les données recueillies au cours de chaque vague d'enquête permettent d'effectuer deux types d'analyse. D'une part, les données annuelles peuvent être analysées dans une perspective synchronique; l'analyse vise alors à refléter aussi fidèlement que possible la situation actuelle dans le pays. D'autre part, les échantillons sont liés progressivement les uns aux autres afin de constituer des séries temporelles individuelles. Ce dispositif permet de procéder à des analyses longitudinales. L'objectif est, alors, de progresser dans la connaissance des processus socio-économiques plutôt que de "refléter" la situation générale de la population de référence.

Chacune de ces approches pose des problèmes techniques et méthodologiques particuliers.

L'analyse synchronique pose des problèmes liés à la nature même de ses objectifs. En particulier, elle exige que l'échantillon ne soit pas affecté par des biais systématiques qui invalideraient les estimateurs statistiques et ne permettraient plus d'établir la marge d'erreur qui encadre toute valeur observée ou estimée sur la base de l'échantillon. L'échantillon doit être "représentatif" de la population-cible.

Mais l'apparition de tels biais est pratiquement inévitable lorsque l'observation est diachronique et porte sur un échantillon composé de personnes interrogées à maintes reprises. En effet, il est peu probable que deux propriétés élémentaires de l'échantillon soient respectées d'une période d'observation à l'autre :

- tous les éléments de l'échantillon doivent conserver la même probabilité de sélection d'une observation à l'autre

- tous les sous-ensembles d'éléments de l'échantillon (toutes les catégories de population) doivent conserver la même probabilité de sélection d'une observation à l'autre.

En outre, les catégories de la population que l'on souhaite analyser avec une attention particulière doivent conserver une taille suffisante au fil des années (G. DUNCAN - 1989).

L'usage des pondérations vise à limiter les effets de cette usure de l'échantillon :

- en détectant les sources des biais

- en estimant l'ampleur des biais

- en compensant les déficits qui ont pu être repérés.

Deux types de phénomènes favorisent l'apparition de biais, lorsque le même échantillon est observé à plusieurs reprises:

- les premiers sont le fait de l'échantillonnage:
  - le mode de tirage de l'échantillon
  - les refus de participer à l'enquête
  - l'entrée de nouvelles personnes dans l'échantillon;
- les seconds correspondent à des événements démographiques; ils s'observent également dans la population générale:
  - des naissances
  - des décès
  - des mouvements migratoires.

Tous ces phénomènes contribuent à modifier la structure de l'échantillon. Ils doivent être pris en compte dans le calcul des facteurs de pondération.

Il est utile de décrire, en premier lieu, l'organisation du fichier individuel où les variables de pondération sont et seront répertoriées.

ORGANISATION

DU FICHIER

Chapitre 2

ORGANISATION DU FICHIER

Le répertoire des variables de pondération réunit l'ensemble des PERSONNES interrogées au moins une fois au cours du panel et l'ensemble des VARIABLES de pondération.

1. Différentes catégories de personnes peuvent être distinguées en fonction de leur trajectoire au sein du panel. Toutes ces trajectoires peuvent être décrites à partir d'une matrice élémentaire basée sur une double distinction: les membres et les non-membres d'un échantillon annuel, d'une part, les membres et les non-membres du panel, d'autre part.

2. Des distinctions supplémentaires doivent être introduites. Certaines catégories de personnes jouent un rôle particulier dans l'évolution de l'échantillon et dans le calcul des variables de pondération. Elles doivent donc être identifiables dans le fichier.

3. Les variables "DIA-gnostic" permettent précisément de sélectionner ces sous-échantillons et de reconstituer les échantillons qui feront l'objet des analyses.

Chaque fois qu'une nouvelle vague d'enquêtes est introduite dans le fichier, une nouvelle variable "diagnostic" est insérée. Elle permet de partitionner l'ensemble du fichier en sous-échantillons. Ces sous-échantillons correspondent à l'ensemble des trajectoires individuelles observées entre la première vague d'enquête et la dernière vague introduite dans le fichier.

4. Toutes les variables de pondération répertoriées dans le fichier n'ont pas la même fonction. On distinguera des variables "opérationnelles" et des variables "instrumentales" en précisant les indications de leur usage.

2.1. LES PERSONNES

2.1.1. Les membres et les non-membres d'un échantillon

Le fichier total est un fichier évolutif et cumulatif. Chaque année les résultats d'une nouvelle enquête viennent s'ajouter aux données enregistrées au cours des années précédentes. Toutes les personnes présentes dans un échantillon ne sont plus nécessairement présentes dans l'échantillon suivant. Inversement, chaque année, de nouvelles personnes participent à l'étude.

Le fichier est évolutif:

* des nouvelles personnes entrent dans l'échantillon et dans le fichier à l'occasion de chaque vague d'enquêtes. Elles ont rejoint, selon des modalités variables, un ménage ou une personne présente dans l'échantillon précédent;

* des personnes quittent l'échantillon provisoirement ou définitivement pour des raisons très diverses (absence du ménage, refus de répondre, décès, émigration, ...).

Le fichier est cumulatif:

* il recense systématiquement toutes les personnes contactées, au moins une fois au fil des vagues d'enquêtes.

Par conséquent,

* le nombre de personnes recensées augmente chaque année,

* le nombre de personnes présentes dans le fichier ne diminuera jamais,

* toutes les personnes recensées dans le fichier n'auront pas nécessairement des "valeurs" ou observations enregistrées pour chaque année . Le numéro d'identification de TOUTES les personnes reste néanmoins enregistré parce que certaines personnes peuvent rejoindre l'échantillon après une absence de plus ou moins longue durée.

Le fichier est donc formé de deux grandes catégories de personnes:
- les membres d'un échantillon; ils ont été interrogés au cours d'une vague d'enquêtes donnée;
- les non-membres d'un échantillon; ils n'ont pas été interrogés au cours de cette vague d'enquêtes.

2.1.2. Les membres et non-membres du panel

Toute personne répertoriée dans le fichier n'est pas nécessairement membre du panel.

Une personne est **membre du panel** lorsqu'elle remplit l'une des deux conditions suivantes.

- Elle appartient à l'échantillon initial tiré en 1985. Les 6110 personnes appartenant à cet échantillon sont des membres du panel.

- Flle est entrée dans un échantillon du panel après la première vague d'enquêtes; elle descend en ligne directe d'un membre du panel (père et/ou mère); elle est entrée simultanément dans l'échantillon et dans la population.

  (Il faut toutefois signaler une exception: quelques personnes sont entrées dans l'échantillon au cours de la deuxième vague d'enquêtes, au titre de fils ou de fille d'un membre du panel. Elles sont rentrées dans leur famille après une brève absence qui correspondait au moment de l'enquête. Elles ont été considérées comme membres du panel bien qu'elles ne soient pas entrées simultanément dans l'échantillon et dans la population).

  En 1986, 78 personnes sont entrées dans l'échantillon au titre de membres du panel.

  En 1987, 65 personnes sont entrées dans l'échantillon au titre de membres du panel: elles sont nées pendant l'année qui sépare les deux vagues d'enquête.

Toute autre personne est définie comme **non-membre du panel**. Elle peut être l'ascendant direct, le conjoint, l'oncle ou l'ami d'un membre du panel. Elle entre dans l'échantillon parce qu'elle se joint à un ménage ou à une personne présente au préalable.

Cette définition est très large: elle n'exclut pas qu'une personne non-membre du panel se sépare du ménage ou de la personne qui l'a accueillie et accueille à son tour une nouvelle personne non-membre du panel. Des ménages non-membres du panel peuvent alnsi se former. Ils continueront à participer à l'enquête et seront toujours répertoriés dans le fichier.

Leur rôle dans le calcul des pondérations annuelles devra être précisé par la suite.

2.1.3. Eléments des trajectoires individuelles

Le fichier est organisé à partir de quatre éléments. La combinaison de ces éléments constitue des trajectoires individuelles. Ces trajectoires retracent la carrière des personnes au fil des échantillons successifs. Elles déterminent leur rôle dans la procédure de pondération.

SCHEMA I.

Eléments des trajectoires individuelles

	statut dans l'échantillon	
statut dans le panel	membre	non-membre
membre	A	B
non-membre	C	D

Le schéma I. permet de distinguer quatre catégories de personnes. Celles-ci sont définies en fonction de leur statut dans le panel et dans un échantillon donné.

La trajectoire d'un individu dans le fichier (et dans le panel) correspond à la série des états qu'il occupe successivement dans les échantillons annuels.

Chacun de ces états définit un ensemble de personnes ayant un rôle particulier à jouer dans le calcul des pondérations.

Case A

Cette catégorie rassemble les personnes membres du panel lorsqu'elles sont présentes dans un échantillon.

Lorsqu'elles occupent cette position, elles sont prises en compte dans la procédure de pondération.

Ce sont les seules personnes qui reçoivent une valeur de pondération. En effet, ce sont les seules personnes présentes dans l'échantillon. Leur probabilité de sélection initiale est connue. Il est donc possible de calculer leur nouvelle probabilité de sélection, pour

une année donnée: cette nouvelle probabilité dépend du taux de réponses observé cette année-là pour l'ensemble des membres du panel.

Ce taux de réponses est calculé chaque année par rapport à l'ensemble des membres du panel. En d'autres termes, il n'est pas seulement calculé par référence aux membres du panel appartenant à l'échantillon initial; ce taux de réponses prend également en compte les membres du panel entrés dans l'échantillon au cours des vagues d'enquête successives.

* Les taux de réponses doivent être calculés par référence à *l'état de l'échantillon de la première année d'observation*; cette précaution permet de réajuster l'échantillon observé sur le profil de l'échantillon aléatoire initial; elle permet également de prendre en compte chaque année des personnes qui rejoignent l'échantillon après une absence plus ou moins prolongée.

Cette mobilité des personnes ne pourrait plus être prise en compte, si les taux de réponses étaient calculés par référence à l'état de l'échantillon de l'année précédente. Toute personne ayant quitté l'échantillon, à un moment donné, serait définitivement écartée de l'étude, puisqu'il serait impossible de calculer la probabilité de répondre au temps t d'une personne absente au temps t-1.

Les effets liés à l'usure de l'échantillon peuvent être limités, en prenant l'échantillon initial comme base de calcul des taux de réponses.

* L'état de l'échantillon de la première année d'observation doit être modifié chaque année; *les nouveaux membres du panel doivent être insérés dans cet ensemble de référence.*

Une probabilité de sélection peut être attribuée aux nouveaux membres du panel. Leur entrée dans l'échantillon dépend directement de la probabilité de sélection de leurs parents. Les nouveaux-nés reçoivent donc un poids équivalent à la moyenne des poids calculés pour leurs parents au cours de l'année de la naissance. Ils reçoivent donc le même poids que leurs parents lorsque le père et la mère ont des poids identiques. Ils reçoivent le poids de l'ascendant présent dans l'échantillon lorsque l'un des conjoints est absent.

La probabilité de sélection de ces nouveaux membres est ensuite calculée chaque année selon la procédure générale appliquée à l'ensemble des membres du panel.

Cette modification de l'échantillon de référence permet de prendre en compte l'évolution démographique de la population de référence.

Ces personnes situées dans la case A, ne passeront jamais à l'état C ni à l'état D au cours de leur carrière. Les membres du panel

conserveront ce statut définitivement. Par contre, il est possible que certains membres du panel quittent l'échantillon (provisoirement ou définitivement): la seule transition possible conduirait ces personnes vers la case B.

Case B

Cette catégorie rassemble les personnes membres du panel lorsqu'elles sont absentes d'un échantillon annuel.

Trois cas peuvent se présenter.

* Ces personnes refusent de répondre, sont absentes ou introuvables. Dans ce premier cas, elles sont prises en compte dans le calcul des taux de réponses qui détermineront toute la procédure de calcul des pondérations. La probabilité de sélection initiale de ces personnes étant connue, il est possible d'évaluer l'impact de leur disparition sur le profil de l'échantillon.

Ces personnes peuvent rejoindre l'échantillon au cours d'une vague d'enquête ultérieure. Elles seront à nouveau membres de l'échantillon et reçevront à nouveau un poids relatif. Elles rejoignent les membres de la case A.

* Ces personnes ont émigré. Dans ce deuxième cas, elles ont quitté simultanément la population de référence et l'échantillon. Elles ne sont pas prises en compte dans le calcul des taux de réponses. Leur départ ne modifie pas le rapport entre l'échantillon et la population de référence.

Ces personnes peuvent revenir au Luxembourg et rejoindre l'échantillon après une ou plusieurs années. Dans ce cas, elles rentrent dans la case A. Elles sont à nouveau prises en compte dans le calcul des taux de réponses. Elles reçoivent un poids relatif.

* Ces personnes sont décédées: elles ont quitté la population de référence et l'échantillon simultanément. Leur départ n'influence pas la représentativité de l'échantillon. Toutefois, les taux de mortalité cumulés dans l'échantillon et dans la population peuvent diverger sensiblement après quelques années. Dans ce cas, l'échantillon sera ajusté à intervalles réguliers.

Les deux premières catégories de personnes ne passeront jamais vers l'état C ou vers l'état D. Elles resteront définitivement membres du panel. La seule transition possible les conduirait à rejoindre la catégorie A et à réduire le taux d'usure de l'échantillon.

Case C

Cette catégorie correspond aux personnes non-membres du panel, lorsqu'elles sont présentes dans un échantillon.

Toutes ces personnes sont des nouveaux membres. Par définition, aucune d'entr'elles n'appartient au premier échantillon alors qu'elles appartenaient à la population de référence au moment du tirage de l'échantillon initial (hormis les personnes immigrées après 1985 et les enfants nouveaux-nés issus de deux parents membres de l'échantillon mais non-membres du panel).

Leur probabilité de sélection, au cours d'une vague d'enquête, ne peut être établie que sur la base d'hypothèses invérifiables. Elles ne sont donc pas prises en compte dans le calcul des taux de réponses.

Ces personnes reçoivent un poids de '0',lorsqu'elles sont présentes dans l'échantillon.

Cette valeur de pondération n'exerce aucune influence sur le profil de l'échantillon mais elle provoque quatre effets.

* Ces personnes ne sont pas prises en compte lorsque l'analyse porte sur l'échantillon pondéré.

* Ces personnes sont prises en compte lorsque l'analyse porte sur l'échantillon non-pondéré.

* Ces personnes sont prises en compte lorsque l'analyse porte sur des données temporelles, à condition qu'elles appartiennent aux échantillons couverts par la période analysée. (A notre connaissance, les analyses de données individuelles temporelles s'effectuent sur des échantillons non-pondérés).

* Le "poids relatif " de ces personnes intervient d'une part, au moment de calculer le poids relatif des nouveaux membres du panel (moyenne du poids des parents) (2.2.1.6.) et d'autre part, au moment de calculer le poids relatif des ménages (moyenne des poids des membres du ménage. Voir 2.2.).

La seule transition possible pour les membres de cette catégorie les conduirait dans la case D. Ils ne seront jamais des membres du panel. Ils n'interviendront jamais dans le calcul des taux de réponses annuels. Leur trajectoire individuelle sera entièrement définie par leur stabilité dans la case C ou par des mouvements plus ou moins fréquents entre les états C et D.

que ces personnes ont occupés au cours de la période considérée.

ex. membre panel 85/ émigré 86 / membre panel 87 / décédé 88
---- / nouv.mem.86/ membre pan.87 / mem.pan.88

Des modalités supplémentaires apparaissent chaque fois qu'un nouvel échantillon annuel est inséré dans le fichier: l'ensemble des trajectoires se diversifie.

Cette variable risque de devenir rapidement complexe et d'un maniement difficile. Une seconde variable propose chaque année une version simplifiée (à partir de la troisième vague d'enquête): cette variable simplifiée reprend uniquement les éléments définissant le statut des personnes dans le système de pondération.

En 1986, la version simplifiée et la version élargie coïncident.

A partir de 1987, la version simplifiée ne prend plus en compte que les termes extrêmes des trajectoires:

- à quel titre les personnes sont-elles entrées dans le fichier?

- quel est leur statut dans l'échantillon choisi?

Ces variables "DIA-gnostic" permettent de gérer aisément le fichier, de sélectionner les échantillons et les sous-échantillons que l'on souhaite analyser voire pondérer.

Dans l'exemple proposé:

* la sélection des groupes #0, #1 et #4 réunit l'ensemble des membres de l'échantillon interrogés en 1985, soit 6110 personnes;

* la sélection des groupes #1, #2 et #3 réunit l'ensemble des personnes interrogées en 1986;

* la sélection des vecteurs #0 et #1 permet de calculer les taux de réponses des membres du panel en 1986 et d'élaborer le dispositif de pondération de l'échantillon observé en 1986;

* la sélection du groupe #1 permet de saisir l'ensemble des personnes interrogées à deux reprises: ce type d'échantillon se prêtera aux analyses de données individuelles temporelles;

* la sélection des groupes #0, #1 et #4 permet de reconstituer l'ensemble des personnes interrogées en 1985 et de changer de niveau d'analyse. Chaque personne est désignée, dans le fichier, par un numéro d'identification. Ce numéro est strictement individuel et reste identique d'années en années. Chaque personne reçoit également un numéro correspondant au numéro

d'identification du ménage auquel elle appartient. Ce numéro d'identification du ménage est identique pour toutes les personnes appartenant au même ménage. Il est donc possible de procéder à l'agrégation des valeurs individuelles au niveau du ménage et, inversement, de ventiler des valeurs observées directement au niveau du ménage sur chacun des membres de ce ménage.

2.2. LES VARIABLES DE PONDERATION

Chaque *chapitre* présente l'ensemble du système de pondération appliqué à *une vague d'enquête.*

Tous les chapitres sont structurés de la même manière. L'exposé suit la logique du calcul des pondérations: chaque *étape* de ce processus correspond à la création d'une *variable intermédiaire.* Toutes les variables ou étapes du processus de calcul sont insérées dans le fichier chaque année.

Les règles élémentaires du calcul des pondérations d'un échantillon probabiliste ne seront pas rappelées chaque fois qu'elles seront mises en oeuvre.

Par contre, le caractère longitudinal de l'étude exige que certaines règles d'ajustement de l'échantillon annuel soient précisées. L'échantillon est soumis à des déformations qui n'apparaissent pas dans une étude synchronique. Ces règles seront évoquées régulièrement. Cette précaution s'impose parce que toutes les études longitudinales n'appliquent pas nécessairement les mêmes règles.

- Les procédures permettant de gérer les *sources de biais spécifiques* sont donc décrites dans cette introduction. Ces règles sont, bien entendu, rappelées régulièrement lorsqu'elles sont mises en oeuvre (2.2.1.).

- Un schéma, placé en tête de chaque chapitre, récapitule l'ensemble de la procédure et permet à l'utilisateur de trouver rapidement la variable qui l'intéresse en consultant uniquement ce schéma (2.2.2.).

- Ces variables sont dites "*instrumentales*" ou "*opérationnelles*" en fonction de leur rôle dans le système de pondération (2.2.3.).

- Enfin, les *poids calculés pour les ménages* ne peuvent pas être utilisés sans quelques manipulations préalables. Ces poids ont été reventilés sur les membres du ménage afin de rassembler toutes les valeurs de pondération dans un seul fichier individuel (2.2.4.).

2.2.1. Processus de calcul: étapes et aspects spécifiques

2.2.1.1. *Première étape*

La première étape de cette démarche consiste à calculer les probabilités de sélection des unités d'analyse. La première année ces probabilités dépendent essentiellement de la position des unités d'analyse dans la *procédure d'échantillonnage*. Les vagues d'observation des années suivantes sont soumises à d'autres aléas. Les probabilités de sélection des unités d'analyse sont modifiées par la *répartition inégale des taux de réponses* dans l'échantillon.

Particularité de l'échantillon initial

L'échantillon initial de cette étude, est un échantillon probabiliste: chaque unité d'analyse a une chance connue et différente de zéro d'apparaître dans l'échantillon. La théorie statistique permet d'établir les propriétés des estimations effectuées dans ce type d'échantillon. Il est donc possible de procéder à des inférences statistiques.

Les unités de tirage sont sélectionnées par la méthode de l'échantillonage aléatoire simple: chaque élément de la population a la même probabilité d'être tiré; chaque combinaison d'éléments a la même probabilité d'être tiré.

Les estimations statistiques n'en sont pas pour autant à l'abri de certains biais.

Le mode de tirage de l'échantillon initial pose un problème à cet égard: l'unité de *tirage*, l'unité d'*observation* (le ménage) et les différents types d'unités d'*analyse* utilisables (individus, ménages, groupes de revenus) ne sont pas identiques.

L'échantillon des unités de tirage ne présente pas de biais systématique mais les ménages et les membres des ménages observés ne forment pas des échantillons aléatoires simples.

Ce problème peut être résolu.

En effet, il est possible d'identifier les liens qui existent entre l'unité de tirage, l'unité d'observation (le ménage) et les unités d'analyse (ménages, individus).

L'identification de ces liens permet de dériver les probabilités de sélection des unités d'analyse.

Ces probabilités étant connues, l'équiprobabilité des unités d'analyse peut être restaurée.

La pondération des unités d'analyse produit exactement cet effet: elle consiste à accorder à chaque unité d'analyse un poids relatif. Toutes les unités ne sont plus équivalentes à une unité. Chaque unité vaut une fraction ou un multiple de l'unité. Cette valeur s'obtient en divisant chaque unité par sa probabilité de sélection.

Ces différents aspects sont traités de manière détaillée dans le chapitre 3 (3.1.).

Généralisation aux vagues suivantes

A partir de la deuxième vague d'enquête, les déformations de l'échantillon sont liées à deux facteurs. Le premier est propre à la vie de l'échantillon: les refus de répondre. Le second reproduit dans l'échantillon les conditions d'évolution de la population de référence (naissances, émigration, décès).

Dans un premier temps, toutes les personnes, entrées à un moment quelconque dans l'étude, doivent être classées en fonction de leur position par rapport à ces deux facteurs.

* Les facteurs démographiques ajustent automatiquement le profil de l'échantillon au profil de la population de référence: les personnes émigrées et les personnes décédées peuvent donc être tenues à l'écart du calcul des nouvelles probabilités de sélection.

* La probabilité de sélection des non-membres du panel est inconnue: ces personnes ne sont donc pas prises en compte dans l'analyse de l'évolution de l'échantillon.

* Les enfants nés de parents membres du panel sont définis comme étant des membres du panel. Ils forment une catégorie particulière. Ils recevront, au terme de la procédure, un poids égal à la moyenne des poids qui auront été attribués à leurs parents pour l'année considérée.

* Un seul facteur reste à mesurer: l'importance relative des refus de répondre.

  Le taux de réponses correspond au rapport entre

  - le nombre de membres du panel présents dans l'échantillon

  - le nombre total des membres du panel.

  Les nouveaux membres du panel entrés dans l'échantillon au cours de la vague d'enquête considérée n'entrent pas dans le calcul des taux de réponses.

Tous les membres du panel entrés au cours des vagues précédentes sont pris en compte, excepté: les personnes émigrées et les personnes décédées.

Tous ces membres sont déjà affectés d'un poids: le poids relatif qui leur a été attribué au moment de leur entrée dans le panel (2.1.3. Case A). Cette précaution s'impose étant donné le biais de sélection lié au mode de tirage de l'échantillon initial.

Si le taux de réponses était uniforme, quelle que soit la manière de partitionner l'échantillon, il faudrait conclure que les refus de répondre n'introduisent aucun biais dans l'échantillon et ne modifient en rien les probabilités de sélection des membres du panel. Il n'y aurait pas lieu de calculer de nouvelles pondérations.

C'est précisément ce qu'il faut vérifier.

Dans un second temps, il convient donc de rechercher des variables, caractéristiques des membres du panel, dont les modalités présentent des taux de réponses très contrastés. Lorsque le contraste entre les taux est significatif, l'échantillon est affecté d'un biais qu'il y a lieu de corriger.

L'identification des variables les plus discriminantes peut être effectuée par différentes procédures (analyse discriminante, "Automatic Interaction Detector", Probit ou toute autre procédure adaptée au fait que la variable dépendante est dichotomique: présence/absence des personnes dans l'échantillon).

Suivant les conseils et l'expérience du P.S.I.D., le panel luxembourgeois a opté pour une procédure qui permet d'éviter d'une part, les contraintes de la linéarité et d'autre part, l'effet des "petits groupes". En évitant d'isoler des trop petits groupes de personnes, cette procédure évite que des taux de réponses extrêmes et propres à des groupes très marginaux influencent le calcul des pondérations.

La procédure adoptée s'apparente à l'A.I.D. mais elle présente deux caractéristiques particulières.

Elle réintroduit à chaque étape de l'analyse TOUTES les variables mises en concurrence. Une variable peut donc intervenir à plusieurs reprises dans le processus d'analyse.

Elle permet d'opérer un choix entre deux variables dont les effets sont approximativement équivalents. L'examen des résultats des discriminations qui suivraient chacune des options possibles fournit des résultats anticipés. Ces résultats peuvent s'avérer très utiles à la décision. La procédure en sens inverse permet de modifier une décision antérieure et d'obtenir une solution plus satisfaisante.

Cette procédure sera illustrée dans chaque chapitre à partir de la deuxième année. Elle peut être décrite de la manière suivante.

* Les membres du panel présents dans l'échantillon reçoivent un code ' 1 '. Les membres du panel absents de l'échantillon reçoivent un code ' 0 ' (les personnes émigrées et les personnes décédées ne sont pas prises en compte).

* Le taux de réponses correspond à la proportion de codes 1 observée dans l'ensemble de l'échantillon ou dans une catégorie déterminée de population.

* L'échantillon est pondéré en fonction de la probabilité de sélection initiale des membres. Cette variable de pondération doit être actualisée chaque année.

  En 1986, année de la seconde vague d'enquête, cette variable correspond exactement à la variable de pondération de l'échantillon initial. En effet, les premiers nouveaux membres du panel apparaissent au cours de cette seconde vague mais ils ne sont pas encore pris en compte dans le calcul des taux de pondération.

  A partir de la troisième vague, la variable de pondération est actualisée chaque année par l'adjonction des poids des nouveaux membres entrés dans l'échantillon au cours de la vague d'enquête précédente.

  Cette première pondération corrige a priori l'inégalité des probabilités de sélection initiales des membres. Dès lors, l'analyse des taux de réponses permet d'identifier *les biais d'échantillonnage liés spécifiquement à la vague d'enquête considérée.*

* Les membres du premier échantillon sont décrits par des caractéristiques qu'ils présentaient en 1985. Les nouveaux membres sont décrits par les caractéristiques qu'ils présentaient au moment ou ils sont entrés dans l'échantillon.

* Ces variables doivent donc être actualisées chaque année par l'adjonction des caractéristiques des nouveaux membres entrés dans l'échantillon au cours de la vague d'enquête précédente.

* Les variables choisies décrivent des caractéristiques du ménage, du chef de ménage et de la personne elle-même.

CARACTERISTIQUES DU MENAGE	DU CHEF DE MENAGE	DE LA PERSONNE
Canton de résidence	Age	Age
Type d'habitat	Sexe	Sexe
Taille du ménage	Etat civil	Etat civil
Nombre d'adultes	Emploi (oui/non)	Formation
Nombre d'enfants		Emploi (O/N)
Nombre d'emplois		Adulte/Enfant
		Lien avec le CM

Toutes ces caractéristiques sont attribuées aux personnes. Les caractéristiques du ménage et du chef de ménage décrivent l'environnement de la personne.

* Ces variables permettent de comparer les taux de réponses de différentes catégories de personnes et d'identifier les catégories de population sous-représentées ou sur-représentées.

Exemple:

Codes	AGE du Chef de ménage	Taux de réponses (86)
1.	< 24 ans	.81
2.	25 à 34 ans	.84
3.	35 à 44 ans	.77
4.	45 à 54 ans	.77
5.	55 à 64 ans	.73
6.	> 64 ans	.69
TOTAL		.77

Dans cet exemple, les personnes appartenant à un ménage dont le chef de ménage est âgé de plus de 64 ans semblent sous-représentées. Leur taux de réponses est particulièrement faible. Leur probabilité de sélection est inférieure à celle de l'échantillon. Leur comportement vis-à-vis de l'enquête a pour effet de sous-représenter dans l'échantillon la catégorie de population dont elles devraient être, théoriquement, les témoins fidèles.

D'autres distortions peuvent être plus centrales. Cette catégorie de personnes peut être absorbée entièrement ou partiellement par d'autres partitions de l'échantillon. Quelle est la partition la plus efficace? Quelle est la partition la plus discriminante?

* Avant de comparer le pouvoir discriminant des variables disponibles, celles-ci doivent être dichotomisées. Cette

précaution annule les effets statistiques liés au nombre de modalités.

Les modalités d'une variable peuvent être combinées selon différents regroupements en vue d'obtenir la dichotomie la plus discriminante.

Dans l'exemple, trois variables peuvent être construites: le groupe #6 peut être opposé à l'ensemble des autres modalités, les catégories #5 et #6 peuvent former une modalité et les groupes #1 et #2 peuvent être opposés aux autres.

* Le test du ETA permet de sélectionner la ou les variables dichotomiques qui identifient et opposent les groupes les plus homogènes (voir encadré).

ENCADRE **RAPPEL**

E^2_{yx} = 1 - (var. intra de Y / var. totale de Y)

$$= 1 - \frac{(Sy^2 - Sn_k Y^2_k)}{(Sy^2 - NY^2)}$$

où: Sy^2 est simplement la somme générale des carrés des scores (y)

Y est la moyenne générale des y.

$Sn_k Y^2_k$ est la somme sur l'ensemble des catégories k, des produits dans chaque catégorie (k) du nombre de cas (n) par le carré de la moyenne de la catégorie (Y^2)

* Généralement, ce test ne suffit pas: plusieurs variables ou plusieurs dichotomies d'une variable restent en concurrence parce que les valeurs de ETA restent très proches.

Exemple: En 1987, la meilleure dichotomie des "cantons de résidence" et la meilleure dichotomie des "âges des chefs de

ménage selon leur sexe" présentaient une valeur de ETA égale à .12.

La comparaison des moyennes des groupes (t) permet de choisir la variable qui crée le contraste le plus significatif (P).

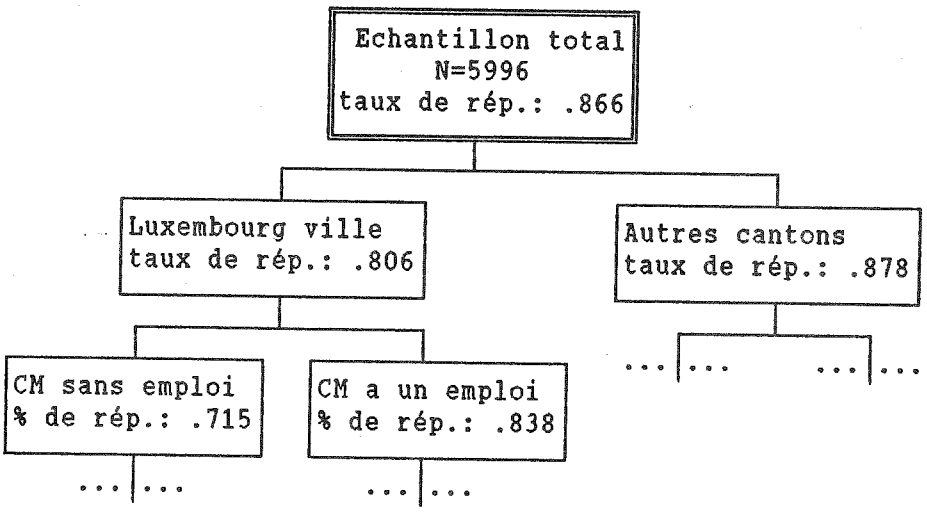
Lorsque les valeurs de t restent très proches, le nombre de cas observés dans chaque modalité peut être pris en compte. Le plus souvent, l'équilibre entre les tailles des groupes sera privilégié. Dans certains cas, l'isolement d'un groupe particulier peut s'avérer plus satisfaisant pour la suite de l'analyse.

Lorsque des hésitations peuvent encore se justifier, il est utile de pouvoir examiner les conséquences de chaque option sur les résultats ultérieurs. (Voir plus haut, les avantages de la procédure adoptée par le C.E.P.S.).

* La première étape de l'analyse permet donc d'identifier la caractéristique dichotomique qui oppose les deux groupes de personnes dont les taux de réponses sont les plus homogènes (Variance intra-groupe) et les plus contrastés (Comparaison des moyennes). Cette partition de l'échantillon est adoptée à deux conditions: (1) le contraste est significatif sur le plan statistique (le biais est significatif), (2) chaque modalité compte au minimum 200 personnes.

L'analyse est interrompue lorsque l'une de ces deux conditions ne peut être remplie. Il ne faut pas en conclure pour autant que l'échantillon est exempt de tout biais lié aux refus de répondre. Il se peut que la ou les variables pertinentes n'aient pas été prises en compte.

* La première partition scinde l'échantillon en deux catégories.



R désigne les taux de réponses observés dans les catégories typologiques de l'échantillon observé au moment t.

A partir de la deuxième vague d'observation, ce calcul doit prendre en compte les poids relatifs des membres au moment de leur entrée dans l'étude.

L'équation (1) doit intégrer ce facteur et devient:

$$W_{i(to+n)} = W_{i(to)} \times (1 / R_{ki(to+n)}) \qquad (2)$$

où: $W_{i(to+n)}$ désigne les poids des individus membres du panel, n années après leur entrée dans le panel

$W_{i(to)}$ désigne les poids des individus au moment de leur entrée dans le panel; t0 correspond à la première année de leur présence dans l'échantillon.

$k_{i(to+n)}$ désigne les groupes typologiques d'appartenance des individus, dans l'échantillon observé n années après leur entrée dans le panel.

R désigne les taux de réponses observés dans les catégories typologiques de l'échantillon observé au moment t_{o+n}.

Exemple:

En 1986: WGT85 est la variable de **pondération** des membres du panel observés en **1985**

W est la variable qui définit la **probabilité de sélection** des membres du panel, présents dans l'échantillon observé en **1986**

W86 est la variable de **pondération** des membres du panel observés en **1986**

d'où: W86 = WGT85 x (1 / W)

Cette variable, définissant le poids des membres du panel présents dans l'échantillon, portera chaque année un nom composé de deux éléments: W (weight) suivi de l'année au cours de laquelle l'échantillon a été observé:

W86 en 1986, W87 en 1987, ...

Il est évident que ces variables ne s'appliquent pas aux nouveaux membres du panel, entrés dans l'échantillon au cours de la vague d'enquête considérée. Ils n'entrent pas dans le calcul des taux de réponses (2.2.1.1.) et leur probabilité de sélection sera établie selon d'autres règles.

Ces variables ne s'appliquent qu'aux membres du panel ayant participé au moins deux fois à l'étude:

Soit: $i_{(t0+n)}$, $n > 0$

2.2.1.4. *Quatrième étape*

L'utilisation de cette variable de pondération ne se justifie que dans le cadre d'une étude qui ne prendrait pas en compte les nouveaux membres du panel entrés dans l'échantillon au cours de la vague d'enquête considérée.

Toutefois il faut noter que cette variable ne peut pas toujours être utilisée telle quelle (2.2.3.).

Cette variable utilisée à l'état "brut" aura généralement pour effet de modifier la taille de l'échantillon.

Dans un échantillon non-biaisé, tous les membres représentent une unité d'analyse. Toutes les unités d'analyse ont le même poids. Toutes les unités d'analyse ont un poids unitaire et la somme des poids individuels est égale au nombre d'unités observées.

Dans un échantillon biaisé, une catégorie de personnes est sur-représentée. La probabilité de sélection de ses membres est supérieure à la probabilité de sélection des autres personnes. Cette sur-représentation d'une fraction de l'échantillon doit être corrigée. Par exemple, si les membres de cette fraction ont deux fois plus de chances d'être sélectionnés que les autres, ces membres ne doivent représenter qu'une demi-unité d'analyse (l'inverse de leur probabilité de sélection).

Lorsque l'échantillon biaisé est pondéré par ce facteur de correction, toutes les unités d'analyse n'ont plus le même poids. Certaines unités d'analyse ne sont plus prises en compte que pour une fraction d'unité.

Par conséquent, *la somme totale des poids individuels n'est plus égale au nombre d'individus.*

La moyenne des poids individuels est inférieure à l'unité.

2.2.1.5. *Cinquième étape*

Cette étape intermédiaire ne donne pas lieu à la création d'une nouvelle variable.

La procédure d'enquête permet d'observer chaque année de nouvelles personnes: l'observation se déroule au sein des ménages. Les ménages sont des unités très instables: des personnes entrent dans le ménage, quittent le ménage, se séparent pour former de nouveaux ménages, reviennent dans le ménage initial seules, accompagnées d'un conjoint, d'un ami, d'une amie d'un ou plusieurs enfants.

L'observation des personnes qui forment le ménage conduit inévitablement à l'observation de nouvelles personnes chaque année.

Dans la majorité des cas, ces personnes appartenaient à la population avant d'entrer dans l'échantillon. Elles auraient pu entrer dans l'échantillon initial (1985). Elles n'ont pas été sélectionnées au moment du tirage de l'échantillon. Leur probabilité de sélection au moment où elles entrent dans l'échantillon est donc inconnue (à moins que des hypothèses particulières soient émises en dépit du fait que ces hypothèses sont invérifiables).

Ces personnes entrent dans l'échantillon. Elles seront suivies lors des vagues d'enquêtes suivantes, aussi longtemps qu'elles accepteront de répondre. Elles seront prises en compte dans des analyses longitudinales et dans l'analyse des données collectées au niveau des ménages. Mais elles ne seront pas prises en compte dans les analyses synchroniques: elles ne sont pas membres du panel (2.1.3. Case C).

Ces personnes reçoivent un poids de '0'. Cet artifice leur confère des propriétés illustrées plus loin (2.2.1.6.).

2.2.1.6. *Sixième étape*

Les personnes qui entrent dans l'échantillon et dans la population au cours de la même année sont prises en compte au titre de membres du panel: ce sont pour la plupart des nouveaux-nés.

Ce statut leur est attribué parce qu'ils régénèrent l'échantillon conformément à l'évolution de la population. En outre, leur probabilité de sélection initiale dépend directement de la probabilité de sélection de leurs parents *au moment ou ils naissent* et entrent dans l'échantillon (2.1.3. Case A).

Le calcul de leur poids initial suit une règle générale: *leur poids initial est égal à la moyenne des poids de leurs parents.* Leur poids