Odile JOINVILLE

2ème année de l'ISUP
Filière industries et services
2002

RAPPORT DE STAGE

MISE EN ŒUVRE DU LOGICIEL POULPE POUR ESTIMER LA PRECISION DE L'ENQUETE « HANDICAP - INCAPACITE - DEPENDANCE »

INSEE
DIRECTION GENERALE

Direction des statistiques démographiques et sociales
Département de la démographie
Division Enquêtes et Etudes Démographiques

Sommaire

Chapitre 3
CALCUL D'UNE ESTIMATION DE LA VARIANCE EN UTILISANT LE LOGICIEL POULPE

INTRODUCTION

Ce stage a été effectué à la Direction Générale de l'INSEE, Institut National de la Statistique et des Etudes Economiques, noyau du système statistique public français. Plus précisément à la division Enquêtes et Etudes Démographiques, sous la direction de Monsieur Pierre MORMICHE, responsable d'enquêtes.
D'une durée de cinq mois, il a commencé le 02 mai 2002 pour s'achever le 30 septembre 2002.

L'enquête « Handicaps-Incapacités-Dépendance » (HID), dont la collecte vient de s'achever et s'est déroulée sur 4 ans visait, avec des co-financements multiples, à fournir pour la première fois des données de cadrage sur la population handicapée de tous âges : effectifs, nature, sévérité, origine des handicaps, conséquences sur la vie quotidienne et la participation sociale, aides reçues et besoins...
C'est la première fois qu'une enquête nationale et publique est réalisée sur le handicap en France.

Compte tenu de la demande sociale et des nombreuses exploitations qui vont être faites à partir des données de l'enquête, des travaux d'estimation de la précision des données obtenues s'avèrent nécessaires.
Les utilisateurs gestionnaires du système d'aide aux personnes handicapées ont besoin d'informations précises pour évaluer le coût des actions envisagées.
Les équipes de recherche ont besoin d'estimer la précision de leurs résultats comme c'est la règle pour la publication dans les revues médicales et scientifiques.

Le plan de sondage de l'enquête HID étant très complexe, la variance ne peut être estimée à partir des formules classiques de la théorie des sondages.

Mon travail consistait à mettre en œuvre le logiciel d'estimation de précision Poulpe pour proposer une estimation des intervalles de confiance. Après avoir recueilli toutes les informations sur l'échantillonnage de l'enquête, il a fallu proposer une modélisation du plan de sondage de HID, de construire les fichiers nécessaires à l'application du logiciel et réaliser les premiers calculs d'estimation de la précision de l'enquête.

Ce projet avait également un objectif méthodologique. Les précédentes applications du logiciel concernaient des enquêtes relevant d'un plan de sondage particulier. Donc la mise en œuvre du logiciel pour HID permettrait de mettre en évidence les éventuelles difficultés ou limitations dans l'application du logiciel à des plans de sondages très complexes, afin de proposer des améliorations.

Mon stage était encadré conjointement par Pascal Ardilly de l'Unité de Méthodologie Statistique (UMS) de l'INSEE, par ailleurs professeur de sondages à l'ENSAE, pour la partie statistique, et par Pierre Mormiche, responsable de HID, pour la partie références à l'enquête, au plan de sondage et définition des objectifs.

PRESENTATION DE L'INSEE

Créé en 1946, l'Institut National de la Statistique et des Etudes Economiques (INSEE) a pour rôle principal la collecte, le traitement, l'analyse et la diffusion de données statistiques sur l'économie et la société française, afin que tous les acteurs intéressés (administration, entreprises, chercheurs, médias, enseignants, particuliers) puissent les utiliser pour effectuer des études, faire des prévisions et prendre des décisions. L'INSEE respecte une règle absolue : le secret statistique.

Au sein du ministère de l'économie, des finances et de l'industrie, l'INSEE coordonne le système statistique public français. Il participe aux travaux menés par les organismes internationaux, notamment Eurostat (Office Statistique des Communautés Européennes).

L'institut assure également une fonction d'enseignement supérieur et de recherche par l'intermédiaire du Groupe des Ecoles Nationales d'Economie et Statistique (Genes).

Les six missions essentielles de l'INSEE.

1. Collecter et produire :

L'INSEE mène des opérations de grande envergure, telles que le recensement de la population et l'inventaire des équipements et services disponibles dans les communes.
Plus fréquemment, il réalise des enquêtes régulières auprès des ménages (emploi, revenus, conditions de vie, logement…) et auprès des entreprises (chiffres d'affaires, prix de ventes, services…).
Le recensement de la population est la plus vaste de toutes les opérations statistiques. Opération menée conjointement avec les 36600 communes, elle nécessite le recrutement de 115000 agents recenseurs pendant 2 mois. Le recensement permet de connaître la population exacte de chaque commune, canton, département... Il est aussi la base de toute statistique démographique et de nombreuses études socio-économiques.
L'enquête emploi est une enquête capitale. Chaque année, 1000 enquêteurs rendent visite à 75000 ménages en l'espace d'un mois ! Les données recueillies permettent d'évaluer la population active, la nature de l'emploi et les caractéristiques du chômage.

Par ailleurs, l'INSEE exploite des fichiers administratifs pour obtenir des informations, en particulier sur les salaires, les entreprises et l'emploi public. Il exploite également les bulletins d'état civil afin d'établir des statistiques sur la natalité, la mortalité et la nuptialité.

L'INSEE est responsable du calcul et de l'analyse des indices les plus courants : indice des prix à la consommation, indice du coût de la construction, indice de la production industrielle...
Il établit les comptes de la Nation : comptes de l'agriculture, du commerce, des transports...

Cette comptabilité trimestrielle et annuelle donne une vision synthétique et quantitative de l'évolution de l'économie nationale.
L'indice des prix à la consommation est un indice fondamental de l'économie nationale. L'élaboration de cet indice occupe environ 300 personnes dont 150 enquêteurs. Ils se rendent chaque mois dans 27 000 points de vente et relèvent près de 150 000 prix. Cet indice, qui mesure l'évolution des prix, joue un rôle clé dans l'analyse de la conjoncture.

Enfin, l'INSEE gère des bases de données, notamment les répertoires des entreprises, le fichier électoral. L'information collectée est en partie organisée en banques de données macro-économiques ou en bases de données infra-communales géolocalisées au niveau des quartiers, afin d'en faciliter l'utilisation et la diffusion.

2. Analyser :

Une fois la collecte réalisée, l'INSEE exploite les données, afin de décrire la situation économique et sociale. Il réalise des études sur :

- l'économie générale : conjoncture, grands équilibres économiques et financiers ;
- le système productif : situation des entreprises, leur comportement ;
- la situation démographique et sociale : naissances, migrations, salaires, emploi, retraites... ;
- l'organisation spatiale : localisation des hommes et des activités, échanges entre territoires.

Il élabore et diffuse des diagnostics conjoncturels à très court terme (six mois).
Il développe également des modèles macro-économiques et des modèles démographiques pour établir des prévisions à moyen et long terme.

3. Diffuser :

L'INSEE diffuse des données et des études économiques et sociales à tout utilisateur : que ce dernier représente une entreprise, une administration ou une collectivité, qu'il soit chercheur, enseignant, journaliste ou simple citoyen, il a accès à l'ensemble de l'information économique et sociale produite par l'INSEE.
Les principaux indicateurs de l'économie sont mis à disposition gratuitement sur internet.

L'INSEE diffuse des informations en veillant à respecter le secret statistique, notion stipulée dans la loi du 7 juin 1951 : il lui est interdit de communiquer à quiconque les informations individuelles issues des enquêtes statistiques ou des fichiers administratifs utilisés dans l'élaboration des statistiques.
Une deuxième loi assure un secret absolu : la loi du 6 janvier 1978 a pour objectif d'empêcher que les traitements informatiques de données concernant des personnes physiques, facilités par le recours à l'informatique, puissent porter atteinte à la vie privée et aux libertés individuelles. Par exemple, l'INSEE n'a pas le droit d'effectuer une enquête sur la séropositivité en ayant les moyens d'identifier les individus, il faut que ces individus restent

complètement anonymes. La Commission nationale de l'informatique et des libertés veille à son application.

4. Coordonner :

Chaque ministère mène des travaux statistiques dans les domaines de sa compétence. Pour assurer la cohérence de l'ensemble, l'INSEE coordonne le système statistique public en assurant le secrétariat général du Conseil National de l'Information Statistique (Cnis).
Le Cnis a pour mission d'examiner chacune des enquêtes du système statistique public et établit un programme annuel comprenant l'ensemble des enquêtes publiques.
Les services producteurs de statistiques présentent leurs projets (enquêtes, recensement, répertoires, panels, exploitations de fichiers administratifs) aux partenaires économiques et sociaux représentés au Cnis. Ces derniers les examinent du point de vue de leur finalité, de leur place dans le dispositif d'information et de leur caractère prioritaire ou non. Car il faut que toute opération corresponde bien à un besoin d'information d'intérêt général.

L'INSEE joue un rôle moteur dans le système statistique public également en élaborant et en harmonisant les concepts, les définitions, les nomenclatures pour l'ensemble de la production statistique**.**

Enfin, il intervient auprès des différents organismes de la statistique publique en mettant à leur disposition plus de 500 cadres statisticiens.

5. Travailler avec les services statistiques étrangers et internationaux :

Au-delà du territoire national, l'INSEE collabore aux travaux statistiques menés par les organismes internationaux. Il entretient des relations avec les principales organisations internationales dont Eurostat et tous ses homologues des autres pays. Il apporte son concours notamment aux services statistiques des pays de l'Est et forme les statisticiens dans les pays en voie de développement.

6. Enseigner et faire de la recherche :

Au travers du Genes, l'INSEE assure l'enseignement et la formation de spécialistes de la statistique, de l'économie et du traitement de l'information pour les entreprises, les administrations et les organismes publics.
Le Genes comprend cinq écoles dont l'ensae et l'ensai.
L'Ensae forme les statisticiens économistes et les administrateurs de l'INSEE. Les administrateurs de l'INSEE sont chargés de la conception, de la direction et de la coordination du système public d'information statistique.
L'Ensai forme les spécialistes de l'ingénierie statistique et les attachés de l'INSEE. Les attachés de l'INSEE sont chargés d'assurer les activités telles que : conception des travaux de production statistique, encadrement de l'exécution, analyse et diffusion des résultats à l'INSEE.

La structure de l'INSEE

L'INSEE est organisée en :

- **une direction générale** qui siège à Paris et qui assure les six principales missions de l'INSEE ;

- **24 directions régionales** et une direction inter-régionale pour la Guadeloupe, la Martinique et la Guyane, qui sont chargées de la collecte et du traitement des données locales d'une part, et de la définition des besoins locaux d'autre part ;

- **cinq centres informatiques nationaux** qui assurent le traitement et le stockage des statistiques produites, la maintenance et l'amélioration des programmes informatiques existants et la mise en place de nouveaux programmes répondant à des besoins inédits.

La direction générale est organisée en six directions reprenant les différents pôles d'action de l'INSEE. Chacune de ces directions comprend plusieurs départements, qui eux-mêmes sont organisés en divisions.

Organigramme de la direction générale de l'INSEE :

DIRECTION GENERALE

Paul CHAMPSUR

Coordination statistique et relations internationales	Statistiques d'entreprises	Statistiques démographiques et sociales	Etudes et synthèses économiques	Diffusion et action régionale	Enseignement et recherche
Jean-Pierre PUIG	Michel HEBERT	Michel GLAUDE	Guy LAROQUE	Michel JACOD	Alain TROGNON

Chacune des six directions ci-dessus est divisée en plusieurs départements.

La direction des statistiques démographiques et sociales comprend trois départements et une unité ayant rang de département:

- département de la démographie ;
- département emploi et revenus d'activité ;
- département des prix à la consommation, des ressources et des conditions de vie des ménages ;
- unité des méthodes statistiques.

Chaque département est divisé en plusieurs divisions ou cellules qui vont assurer des fonctions précises, tout en travaillant en coordination.

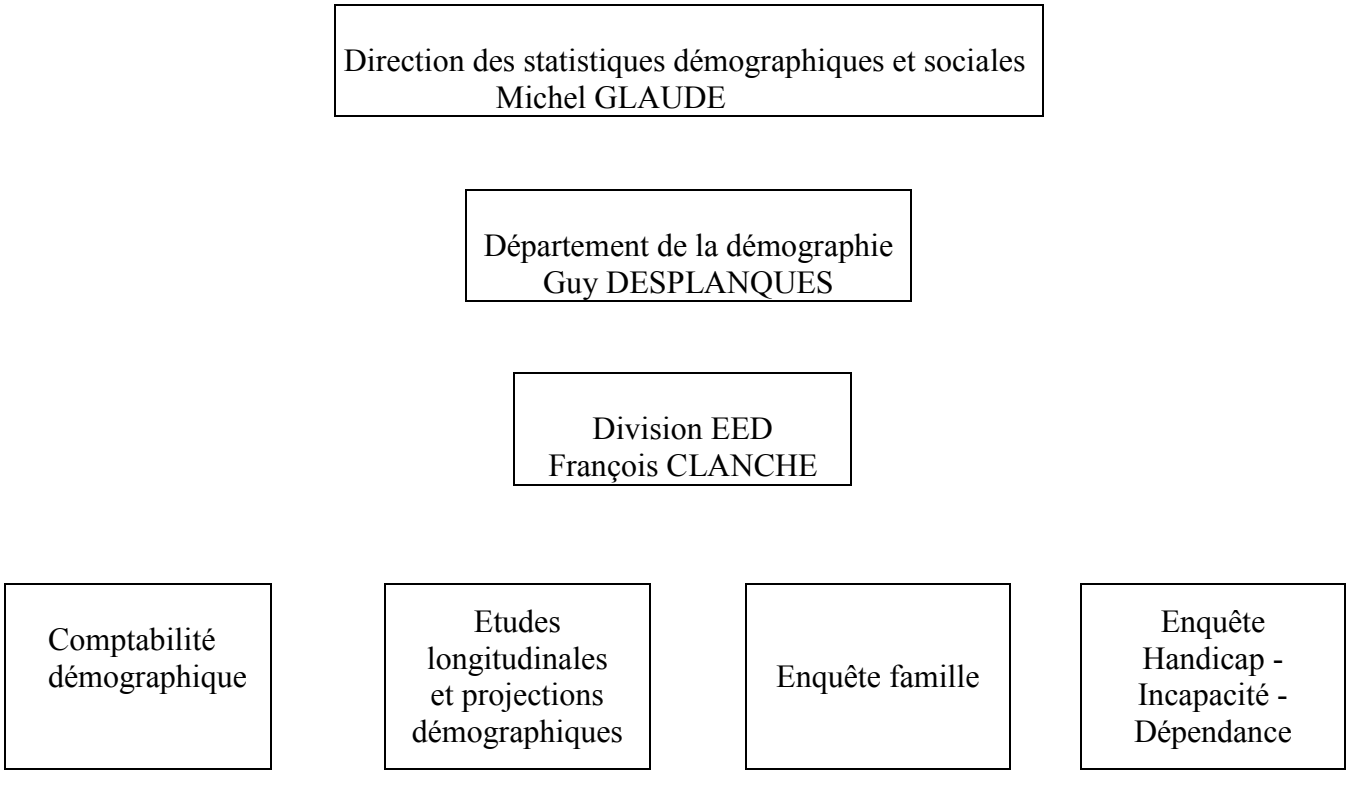
Le département de la démographie est chargé de l'ensemble des travaux et des études démographiques sous le double aspect national et local, en relation étroite avec les directions régionales. Il comprend :

- la cellule Statistiques et études sur l'immigration ;
- la division Répertoires et mouvement de la population ;
- la division Enquêtes et études démographiques ;
- la division Recensement de la population ;
- le pôle Infrastructures géographiques.

La division enquêtes et études démographiques

Au sein du département de la démographie, **La division enquêtes et études démographiques** est responsable de l'enquête famille associée à chaque recensement et d'enquêtes démographiques menées souvent en collaboration avec l'Ined (Institut National des Etudes Démographiques) ainsi que du calcul des principaux indicateurs démographiques. Le bilan démographique annuel, les évaluations et les projections de population sont à la charge de la division. Cette division regroupe 16 agents dirigés par Mr François CLANCHE.
Le nouveau chantier sur les incapacités et la dépendance vient enrichir les approches habituelles de la mortalité et du vieillissement.

Le travail de la division est réparti en quatre sections :



La section « comptabilité démographique » regroupe cinq agents chargés du suivi de la conjoncture en matière de naissances, mariages et décès, des évaluations de population, de la production des principaux indicateurs démographiques et de la publication des données.
La section exploite les fichiers de l'état civil afin de produire des estimations mensuelles, trimestrielles ainsi que des statistiques exhaustives annuelles.

La section « études longitudinales et projections démographiques » regroupe trois agents. Cette section à la charge des études longitudinales à caractère sociodémographique et des projections démographiques. Elle met à jour les décès et assure la constitution des échantillons de mortalité afin d'étudier les variations de la mortalité par sexe et par couche sociale. Enfin, la section réalise les projections démographiques à long terme, après chaque recensement.

L'enquête famille est réalisée à l'occasion de chaque recensement. L'« Étude de l'histoire familiale » de 1999 s'intéresse aux situations « de fait » et à l'histoire familiale des hommes et des femmes, depuis le départ du domicile parental jusqu'aux arrière-petits-enfants, en passant par les périodes de vie en couple marié ou non, les enfants et beaux enfants. Une partie de l'enquête décrit la transmission familiale des langues.
Cinq agents travaillent sur cette enquête.

L'enquête « Handicaps-incapacités-dépendance » permettra de mesurer la fréquence de différents types de handicaps, et l'évolution des situations individuelles grâce à un suivi sur deux ans. La collecte, réalisée auprès de personnes vivant en ménages ou en institutions, s'est étalée sur quatre ans, de 1998 à 2001.
Le responsable de la section est Mr Pierre MORMICHE, il travaille en collaboration avec trois autres agents :
- Mme Christine COUET,
- Mlle Catherine GOILLOT,
- Mme Christelle ROINEAU.

Actuellement, l'équipe de travail achève l'édition de statistiques issues de l'enquête, ainsi que l'édition d'un dictionnaire des codes, afin de permettre à des organismes de recherches d'effectuer des études. La section mène également des travaux d'estimation de la précision du sondage.

PRESENTATION DU SUJET DU STAGE

L'enquête « Handicaps - Incapacités - Dépendance » (HID)

Pourquoi une enquête sur le handicap ?

Les spécialistes du domaine du handicap représentent les liens entre les maladies et les handicaps par le schéma dit « schéma de WOOD » :

Maladies ⟶ Déficience ⟶ Incapacité ⟶ Désavantage

Prenons l'exemple d'un homme affecté par la poliomyélite.

Poliomyélite ⟶ Atrophie des jambes ⟶ ne peut marcher ⟶ besoin d'un poste de travail adapté

Les « maladies » comprennent également les accidents et autres traumatismes moraux et physiques, elles relèvent du diagnostic médical.

Les « déficiences » sont les pertes ou les disfonctionnements des diverses parties du corps ou du cerveau. Elles sont en général les conséquences d'une maladie ou d'un traumatisme.

Les « incapacités » sont les difficultés ou impossibilités de réaliser certains actes de la vie quotidienne : se déplacer, s'habiller, se servir d'un téléphone, lire, mémoriser par exemple.

Les « désavantages », terme préféré à « handicap », désignent les difficultés ou impossibilités que rencontre une personne à remplir les rôles sociaux auxquels elle peut aspirer ou que la société attend d'elle : accomplir un travail rémunéré, communiquer, suivre des cours scolaires par exemple.

On évalue qu'entre 6 millions et 9 millions de personnes de tous âges souffrent d'un handicap ou d'une gène permanente liée à leur état de santé, et notamment à leur âge. Une part importante d'entre elles ne peut effectuer sans aide les activités courantes de la vie quotidienne.

La majorité des personnes concernées résident en domicile ordinaire (par opposition à la vie en institution, en communauté). De très nombreux organismes, nationaux, locaux, privés ou associatifs, apportent une aide financière ou humaine aux personnes concernées.
On évalue à 2.5 millions le nombre de personnes percevant une allocation en raison d'un handicap, dont 250 000 vivant en institutions.

Compte tenu des enjeux économiques et sociaux liés à la population concernée par le handicap et de l'insuffisance des informations actuellement disponibles en France sur le handicap et les déficiences en général, l'INSEE a décidé de réaliser une enquête sur les conséquences des problèmes de santé sur la vie quotidienne des personnes.

Les objectifs de l'enquête.

L'enquête HID répond à trois objectifs :

1. **mesurer le nombre** des personnes handicapées ou dépendantes ;
2. **évaluer les flux** d'entrée et de sortie en incapacité ;
3. relever la nature, la quantité, et les fournisseurs d'**aides** existantes, ainsi que les **besoins** non satisfaits.

L'enquête couvrait l'ensemble de la population, de tous âges, vivant en domicile ordinaire et en institutions. La méthode consistait à interroger un échantillon représentatif de la population pour compter la fréquence de tous les types de difficultés liés à la santé. Le questionnaire passe donc en revue les gestes de la vie quotidienne et les activités sociales des personnes.

Il a été décidé de réaliser l'enquête auprès d'un échantillon de 15 000 pensionnaires des institutions spécialisées (établissements pour personnes âgées, foyers pour handicapés, institutions psychiatriques) et de 20 000 personnes vivant en domicile ordinaire.
En ce qui concerne la vague ménages, l'INSEE a décidé de filtrer la population afin de localiser les personnes concernées par le handicap et les déficiences en général. Pour réaliser le filtrage, un court questionnaire (Vie Quotidienne et Santé) a été remis à un gros effectif (400 000 personnes) au moment du recensement de la population de mars 1999.

L'enquête s'est déroulée sur quatre ans, d'octobre 1998 à fin 2001. A la fin de l'année 1998, près de 15 000 pensionnaires des institutions ont été interrogés par HID. L'enquête de filtrage a été réalisé pendant le recensement, près de 360 000 réponses exploitables ont été recueillies. L'enquête HID a été menée fin 1999 auprès de 20 000 personnes sélectionnées parmi celles ayant répondu à l'enquête VQS.

L'enquête devant aussi servir de base aux prévisions, il était nécessaire également de relever l'ancienneté des difficultés ainsi que leur évolution dans le temps. C'est pourquoi une deuxième interview des personnes a eu lieu deux ans plus tard, fin 2000 pour les personnes vivant en institutions et fin 2001 pour la vague ménages.

Ce projet, inédit en France a été financé par de nombreux organismes : le ministère de l'emploi et de la solidarité, l'association chargée de gérer les fonds d'insertion des travailleurs handicapés (AGEFIPH), les caisses nationales de sécurité sociale, le monde des mutuelles et assurances, les caisses de retraite complémentaire, ainsi que l'association des paralysés de France (APF).

L'enquête HID a aussi été réalisée auprès d'un échantillon de détenus.

La structure du questionnaire de l'enquête auprès des personnes vivant en domicile ordinaire est présentée en annexe 1.

Les objectifs du stage.

Pourquoi établir des intervalles de confiance ?

L'INSEE a déjà reçu (et va encore recevoir) une forte demande d'exploitation des résultats de l'enquête par des organismes de recherche privées et publiques. Des études sur les conditions de vie des handicapés sont menées dans divers domaines, par exemple : « la scolarité et le handicap », « les handicapés vieillissants ». On s'intéresse beaucoup aux prévalences. On peut regarder s'il existe des différences entre les hommes et les femmes, entre les couches sociales, entre les tranches d'âge, s'il existe une sous-population qui se démarque des autres, de façon à définir les cibles privilégiées de la politique du handicap. Cependant, seule la confrontation des intervalles de confiance permettra de conclure à une différence significative.
Les équipes de recherche auront donc besoin d'assurer la fiabilité de leurs résultats pour être autorisées à les **publier**.

L'enquête HID a pour but de permettre aux acteurs de la politique sociale et de santé de savoir sur quels critères agir pour réduire le champ du handicap et comment répondre aux besoins des personnes concernées. Ces actions peuvent être :

- la recherche et les soins médicaux pour guérir ou prévenir la maladie ;
- la mise au point et la mise en place de prothèses pour réduire une déficience ;
- la diffusion d'aide technique ou l'apport d'aide humaine ;
- une action environnementale : aménager les rues, les transports, les postes de travail ...

Supposons que l'Etat décide de mettre en place une allocation pour aider les personnes souffrant de telle déficience et ayant tel besoin. Grâce aux données de HID, les acteurs pourront savoir combien elles sont. Cependant, pour des **raisons budgétaires**, il est nécessaire d'avoir un encadrement relativement sûr de l'effectif.

Les objectifs du stage

Compte tenu de la complexité du plan de sondage de l'enquête HID, les formules classiques d'estimation de la variance ne peuvent être appliquées pour estimer la précision de l'enquête. La Division EED a donc décidé d'utiliser le logiciel de calcul de précision Poulpe, conçu par l'INSEE.

La mise en œuvre de ce logiciel demande un long travail préparatoire qui passe par une description fine du plan de sondage et par la construction de plusieurs fichiers de travail.
La première partie de mon travail a donc constitué en une documentation complète sur l'échantillonnage de l'enquête, afin de soumettre au logiciel une modélisation du plan de sondage.
Après l'étude des pondérations de l'enquête, le travail de préparation à la mise en œuvre du logiciel a été effectué. Les premiers calculs d'estimation de variance ont été réalisés sur une dizaine de variables.
Les estimations de variance étudiée ici ne concernent que la vague ménages de l'enquête.

Le but de ce travail est de fournir le logiciel accompagné des fichiers de travail aux équipes de recherche, afin qu'elles puissent mener des calculs de variance sur les variables de leurs choix.

Ce projet a également un objectif méthodologique. L'application de Poulpe à HID est en quelque sorte un essai. Les enquêtes pour lesquelles des estimations de précision ont été réalisées en utilisant Poulpe avaient un plan de sondage particulier. Les tirages avaient été effectués à partir de « l'échantillon-maitre », et les échantillons étaient de tailles plus limitées que la pré enquête VQS.

Le fait d'appliquer Poulpe à HID permettra de savoir si le logiciel est suffisamment adapté aux plans de sondage très complexes, s'il existe des difficultés ou des limitations dans l'utilisation du logiciel et si certaines des modalités ont besoin d'être revues.

Chapitre 1
Description du plan de sondage de l'enquête HID

Le plan de sondage mis en œuvre par l'INSEE pour réaliser l'enquête HID (Handicap- Incapacité-Dépendance) est un plan très complexe. En effet, le sondage a été effectué en deux temps :

- D'abord une enquête de filtrage VQS (Vie Quotidienne et Santé) selon un mode de tirage stratifié, à deux degrés et aréolaire.

- Puis l'enquête HID elle-même auprès d'un sous- échantillon des répondants à VQS, selon un mode de tirage stratifié et à allocation non proportionnelle.

Il s'agit donc d'un tirage en deux phases avec post - stratification.

I.Le plan de sondage VQS.

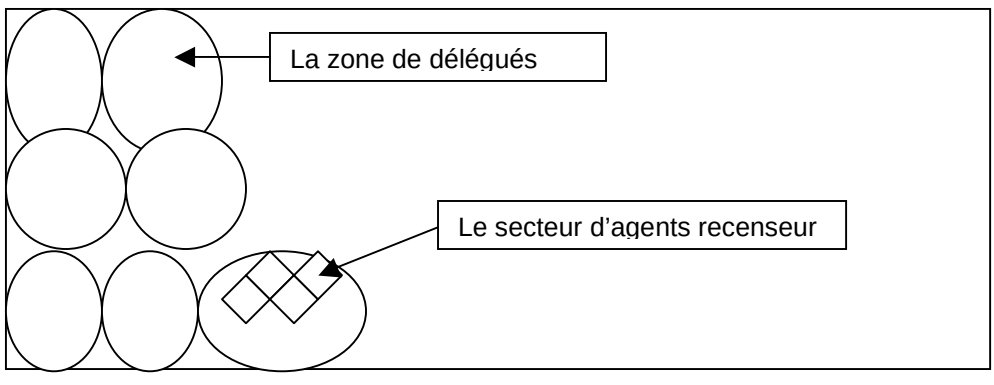
L'enquête VQS devait à l'origine être effectuée auprès d'un échantillon de 123 000 ménages (tous les individus du ménage seraient interrogés). Certains départements ayant demandé des estimations locales, on a effectué des extensions d'échantillon sur ces départements, avec au total un échantillon de 180 000 ménages, ce qui représentait 443 000 individus.

La population française a été découpée en strates géographiques, de façon à permettre une bonne représentativité pour les grandes et les moyennes régions. Le tirage de l'échantillon des individus (plus précisément des ménages) pour le sondage VQS a été réalisé en deux étapes à l'intérieur de chaque strate :

- dans un premier temps, on a effectué le tirage de regroupements de districts appelés ***zones de délégués*** ;
- puis dans un deuxième temps, on a tiré ***des secteurs d'agents recenseurs*** à l'intérieur de chaque zone de délégués sélectionnés ;
- enfin, l'agent recenseur enquête tous les habitants de son secteur.

Le tirage pour une autre enquête a été en quelque sorte associé à celui du sondage VQS : l'Etude de l'Histoire Familiale (EHF). En fait, le tirage des zones de délégués est commun aux deux enquêtes, mais comme l'échantillon EHF était plus grand, on a tiré plus d'agents recenseurs pour EHF dans les zones de délégués sélectionnées.

La strate géographique



I.1.La stratification.

Le découpage du territoire métropolitain en strates s'est effectué selon divers aspects.
Au début, une strate représentait une région administrative.
Mais pour tenir compte des extensions demandées par certains départements, le but était de tirer une plus grande proportion d'individus dans ces départements; il a donc fallu séparer le département en question du reste de sa région ; il constituait à lui seul une strate, les autres départements de la région constituant une deuxième strate.
Ensuite, certaines fractions de département ayant demandé une extension « langues » dans le cadre de l'enquête EHF, cette partie du département a constitué une strate, le reste du département et les autres départements de la région ont constitué une deuxième strate.

Prenons le cas de la Loire (département 42) qui a décidé de financer une extension d'échantillon pour le sondage VQS. Une première strate est constituée par ce département et une deuxième strate est constituée par les départements 01, 07, 26, 38, 69, 73 et 74.

Ou encore le département des Pyrénées Atlantiques (département 64) qui a décidé de financer une extension « langues ». La partie du département parlant le basque a constituée une strate et le reste du département 64 ainsi que les départements 47, 40, 33 et 24 ont constitué une deuxième strate.

En ce qui concerne l'enquête VQS, 7 départements et une région ont demandé une extension :
- région 23 (Haute-Normandie) avec les départements 27 (Eure) et 76 (Seine-Maritime),
- dép. 77 (Seine-et-Marne) rattaché à la région de collecte 21 (Champagne-Ardenne),
- dép. 95 (Val-d'Oise) rattaché à la région de collecte 23 (Haute-Normandie),
- dép. 62 (Pas-de-Calais) de la région 31 (Nord-Pas-de-Calais),
- dép. 35 (Ille-et-Vilaine) de la région 53 (Bretagne),
- dép. 42 (Loire) de la région 82 (Rhône-Alpes),
- dép. 34 (Hérault) de la région 91 (Languedoc-Roussillon),
- dép. 13 (Bouches-du-Rhône) de la région 93 (Provence - Alpes - Côtes d'Azur).

Les départements concernés par une extension « langue » sont :

- dép. 59 (Nord) de la région 31 (Nord-Pas-de-Calais),
- dep. 62 (Pas-de-Calais) de la région 31 (Nord-Pas-de-Calais),
- dep. 57 (La Mozelle) de la région 41 (Lorraine),
- dep. 64 (Les Pyrénées Atlantiques) de la région 72 (Aquitaine),
- dep. 66 (Les Pyrénées Orientales) de la région 91 (Languedoc-Roussillon).

Le territoire métropolitain a donc été divisé en 36 strates, de taille régionale ou infra-régionale.

	Région de collecte	Départements	particularités
strate 01	11	75,78,92,93 et 94	
strate 02	21	8,10,51,52	
strate 03	21	77	extension VQS
strate 04	22	2,60,80	
strate 05	23	27	extension VQS
strate 06	23	76	extension VQS
strate 07	23	95	extension VQS
strate 08	24	18,28,36,37,41,45	
strate 09	24	91	rattaché à la région de collecte 24
strate 10	25	14,50,61	
strate 11	26	21,71,89	
strate 12	31	59	reste de la région 31
strate 13	31	59	extension langues
strate 14	31	62	extension VQS
strate 15	31	62	extension langues
strate 16	41	54,55,57	
strate 17	41	57	extension langues
strate 18	42	67,68	
strate 19	43	25,39,70	
strate 20	52	44,49,53,85	
strate 21	53	22,29,56	
strate 22	53	35	extension VQS
strate 23	54	16,17,79,86	
strate 24	72	24,33,40,47,64,	
strate 25	72	64	extension langues
strate 26	73	12,31,32,46,81,82	
strate 27	74	19,23,87	
strate 28	82	1,7,26,38	
strate 29	82	42	extension VQS
strate 30	83	3,15,63	
strate 31	91	30,48	
strate 32	91	34	extension VQS
strate 33	91	66	extension langues
strate 34	93	4,5,6,83,84	
strate 35	93	13	extension VQS
strate 36	94	20	

I.2.Le tirage des zones de délégués (ZD).

Il y a au total 3553 zones de délégués, dont le nombre par région dépend bien-sûr de la taille de la région, avec un maximum de 464 pour l' Ile de France et un minimum de 31 pour la Corse.

Au cours de ce premier tirage, on a décidé de sélectionner environ 1 ZD sur 10 si la strate représentait une région ou le complément régional de départements en cas d'extension VQS, et un peu plus si la strate correspondait à un département (afin de traiter les extensions VQS).

En fait pour pouvoir réaliser de bonnes estimations locales, le but était de tirer un échantillon d'au moins 20 000 personnes, en tous cas tirer au moins 8 ZD ; donc en moyenne, dans les départements demendant une extension VQS, on a tiré environ 1 ZD sur 4, avec un maximum pour les plus petits départements.

Pour la réalisation de ce tirage, on disposait du nombre de ZD dans chaque région et on a déterminé la quantité qu'il fallait sélectionner, c'est à dire environ un dixième.
Pour chaque département, on a déterminé le nombre de ZD à tirer dans le cas où le département demanderait une estimation locale, c'est à dire un quart du nombre de ZD du département avec un minimum de 8 ZD. Il est à préciser qu'au moment où a été effectué le tirage des ZD, la liste des départements intéressés par l'extension n'était pas encore connue.
Afin de pallier ce problème, on a décidé d'effectuer le tirage des zones de délégués d'abord au niveau départemental et de faire comme si tous les départements demandaient une extension. Ensuite on a regroupé les ZD ainsi sélectionnées par région et parmi elles, on a tiré les ZD qu'il fallait dans le cas où il n'y a pas d'extension. Ainsi, quand on connaîtrait la liste des départements intéressés par l'extension, il suffirait de conserver toutes les autres ZD sélectionnées au cours du premier tirage dans le département concerné par l'extension VQS.

Pour réaliser l'extension langue sur une fraction de département, deux mesures ont été prises. D'abord, étant donné que la sélection des ZD a été effectuée par région (ou par département en cas d'extension VQS), rien ne garantissait que des ZD soient tirées dans la fraction de département concernée. C'est pourquoi a t-on rajouté des régions « fictives » constituées par les fractions de département intéressées par l'extension langue d'une même région. Par exemple, dans la région 31, la partie du département 59 et la partie du département 62 qui ont bénéficié d'une extension langue ont formé la région 30. Ces fractions de département sont devenues respectivement les départements 3A et 3B. Donc les strates 13 et 15 sont constituées respectivement des départements 3A et 3B et forment la région 30.

Au total, quatre régions fictives ont été rajoutées aux 22 régions de collecte :

- région 30 constituée des départements 3A et 3B ;
- région 40 constituée de la fraction du département 57 (région 41), elle est donc constituée de la strate 17 ; cette partie du département était appelée département 4A ;
- région 70 constituée de la fraction du département 64 (région 72), elle est donc constituée de la strate 25 ; cette partie du département était appelée département 7A ;
- région 90 constituée du département 66 (région 91), elle est donc constituée de la strate 33 .

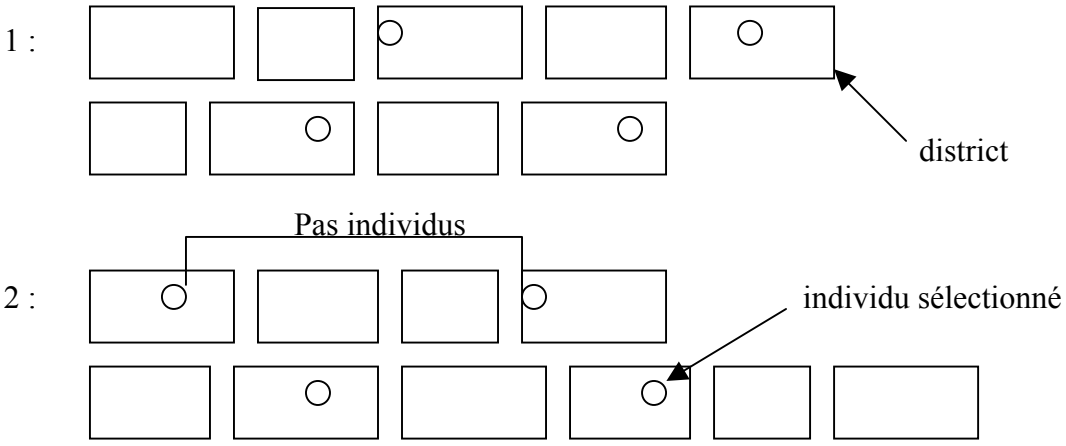
La deuxième mesure prise pour réaliser l'extension langue était de sélectionner plus de ZD dans la fraction de département concerné. Donc dans les 4 régions fictives, on a tiré une proportion plus grande en ZD que celle de 1 ZD sur 10.

En pratique, on a utilisé le fichier des districts établi après le recensement de la population en 1990. Un district représente environ 180 habitants. Pour chaque district, on a recherché la structure socio-économique ainsi que la structure familiale. A l'intérieur de chaque strate (donc ici le département), on a classé les districts selon une nomenclature socio-économique.

Soient NDVQS le nombre de ZD à tirer dans la strate et Xsp_{90} la population de la strate au recensement de la population de 1990.On choisit de tirer deux ZD supplémentaires pour les réserves, on constitue ainsi un matelas (de deux zones) par strate.
L'algorithme de tirage employé est celui du tirage systématique.
On calcule le pas en nombre d'individus qui vaut :

Pas = int (Xsp_{90} / NDVQS +2) où int (a) désigne la partie entière du réel positif a.

Le premier individu est tiré au hasard en début de fichier de la manière suivante :
on détermine un nombre au hasard entre 0 et 1 appelé aléa et son rang dans le fichier vaut : aléa * Pas.
Partant de celui-ci, le principe consiste à descendre ensuite le long du fichier en retenant un individu tous les Pas individus.
Ensuite on a recherché dans quels districts étaient les individus tirés et ces districts ont constitués des « points d'entrée », les ZD retenues étant celles incluant ces districts.



……

1 : signifie que les districts ont une structure socio-économique de classe 1

Enfin, on a déterminé au hasard les deux ZD qui seraient matelas parmi celles ainsi tirées.

Le fait d'avoir trié le fichier des districts a assuré une bonne représentativité de l'échantillon selon les caractéristiques sociales.
On montrera dans le chapitre suivant que les zones de délégués n'étant pas toutes de même taille, les probabilités de chacune d'être sélectionnée à l'intérieur d'une même strate diffèrent.

Pour se ramener à la strate régionale, dans laquelle il fallait tirer une ZD sur 10, on regroupe par région les « points d'entrée » tirés précédemment, en excluant les zones matelas, puis on les a trié selon une nomenclature de structure familiale.

Soient NDEHF le nombre de ZD à tirer dans la strate et Xsp_{90} sa population au recensement de la population de 1990. On rajoute un matelas de deux ZD par strate.
On a tiré les NDEHF zones de délégués parmi les NDVQS zones de délégués (moins les zones matelas) que l'on a sélectionné par le premier tirage.
La procédure de tirage est la même que précédemment, le pas de tirage en nombre d'individus étant égal à

Pas2 = int (Xsp_{90} / NDEHF + 2).

Dans cette étape du tirage, on affecte aux districts un poids différent de celui qu'ils avaient au premier tirage. En effet, au cours du tirage départemental, on a tiré des individus dans le département, chaque district étant constitué de sa population de 1990 ; ainsi, le total des populations des districts était égal à la population du département en 1990.
Au cours du deuxième tirage, on affecte à chacun des NDVQS districts tirés dans le

département une population qui vaut : $\frac{\text{population du département}}{\text{NDVQS}}$, ainsi le total des populations donne la population de la région en 1990.

Et comme pour le premier tirage, le premier individu est sélectionné au hasard en début de fichier, puis on descend le long du fichier en sélectionnant un individu tous les Pas2 individus.
Les matelas sont également sélectionnés au hasard.

Une fois que les départements voulant une extension pour l'enquête VQS ont été connus, on a conservé toutes leurs ZD sélectionnées au premier tirage ; ainsi on avait la configuration d'une ZD sur 10 par strate et au moins 8 ZD en cas d'extension.

On a ainsi tiré un échantillon de 425 zones de délégués.
Le fait d'avoir trié le fichier des districts a assuré à l'échantillon une bonne représentativité selon les caractéristiques sociales et démographiques de la population française.

I.3.Le tirage des secteurs d'agents recenseurs (AR).

Un agent recenseur est responsable en moyenne d'une zone de 530 habitants. Il est chargé de distribuer et de relever les questionnaires VQS à l'ensemble de la population vivant en domicile ordinaire dans sa zone de recensement.
Une ZD comporte en moyenne 30.8 AR .

Comme on l'a expliqué au paragraphe précédent, le tirage des zones de délégués pour l'enquête EHF est commun à celui de l'enquête VQS.
Dans une ZD sélectionnée, on a décidé de tirer 7.7 AR sur les 30.8 pour les deux enquêtes, mais chaque AR tiré ne s'occupera que d'une seule enquête.

Le nombre d'AR à sélectionner pour l'enquête VQS dépendait de la nature de la zone de délégués, à savoir s'il s'agissait :

1. d'une zone sans aucune extension ;
2. d'une zone avec extension VQS ;
3. d'une zone avec extension « langue » ;
4. d'une zone avec extension VQS et extension « langue » (cas du département 62).

Dans le premier cas, sur les 7.7 AR à sélectionner, 1.6 s'occupaient de l'enquête VQS.
Dans le deuxième cas, la part VQS était plus grande afin d'avoir un échantillon plus important et surtout de taille suffisante. Et dans le troisième cas, moins de secteurs d' AR pour VQS ont été sélectionnés afin d'interroger plus de personnes dans l'enquête EHF.

On a calculé les différentes probabilités txv de tirage des secteurs d'AR ; ce calcul sera présenté au chapitre suivant. txv prend des valeurs qui vont de 0.02 à 0.17 avec bien-sûr un maximum pour les zones de délégués tirées dans un département demandant une extension VQS.

Pour chaque ZD, on connaissait le nombre exact N de secteurs d'AR au moment du tirage ;on a donc calculé le nombre n de secteurs d'AR à tirer pour l'enquête VQS en appliquant le taux de sondage txv correspondant à la nature de la zone.
Le tirage des n secteurs d'AR parmi les N de la ZD a été effectué par sondage aléatoire simple.

Toutefois, il a été décidé qu'en raison de la taille de la strate 13 (fraction du département 59 concernée par l'extension langue), aucun secteur d'AR ne serait tiré pour l'enquête VQS. Donc on ne prendra pas en compte les ZD de cette strate, le taux txv vaut 0 dans toutes ses ZD. Par ailleurs, à cause des effets d'arrondi lors du calcul du nombre de secteurs d'AR à tirer dans une ZD, il est arrivé qu'aucun secteur n'aient été tiré dans certaines ZD bien que le taux txv ne soit pas nul.
Au final, le nombre de ZD concernées par l'enquête VQS est de 391.

Sur le territoire métropolitain, 763 secteurs d'agents recenseurs ont ainsi été sélectionnés, 2 275 districts ont été interrogés appartenant aux 391 zones de délégués.
Les départements avec extension VQS ont enquêté beaucoup plus de secteurs que la moyenne car non seulement plus de zones de délégués faisaient partie de l'échantillon, mais plus de secteurs ont été tirés dans chacune de leurs zones. Par exemple, le département 77 avait 40 secteurs de sélectionnés alors que le département 75 en avait 18.
Tous les ménages habitant les secteurs tirés à titre de résidence principale ont donc reçu un questionnaire. L'échantillon de l'enquête VQS a ainsi été déterminé.

Dans les faits, 416 000 individus ont été concernés par le sondage VQS, parmi lesquels 359 000 ont fournis une réponse suffisamment complète pour pouvoir être exploitée pour l'enquête HID , soit un taux d'échec de 14% .

I.4. Résumé.

Le tirage systématique a assuré à chaque individu d'une strate la même probabilité de tirage. On peut donc dire que le tirage des districts s'est effectué proportionnellement à la taille de leur population en 1990. On montrera au chapitre suivant que le premier degré de tirage peut-être assimilé à un tirage de zones de délégués proportionnellement à la taille de leur population en 1990.

En résumé, la première phase du plan de sondage est un sondage stratifié à deux degrés et en grappes, le premier degré de tirage étant le tirage de zones de délégués proportionnellement à la taille de leur population en 1990, le second étant un tirage d'agents recenseurs par sondage aléatoire simple.

Les probabilités de tirage pour les deux degrés diffèrent entre les deux strates.

II. Le plan de sondage HID.

A l'origine, l'enquête HID devait être réalisée auprès d'un échantillon de 20 000 personnes parmi les répondants à l'enquête VQS. L'Hérault qui avait déjà demandé une extension VQS pour son département a en plus demandé une extension d'échantillon pour l'enquête HID. C'est la raison pour laquelle, on a cherché à ce que l'échantillon compte 20 000 personnes pour l'enquête nationale plus 1 800 dans le département de l'Hérault.

Pour cette enquête, on a décidé de déterminer un échantillon qui surreprésenterait fortement les personnes les plus certainement et sévèrement atteintes par un handicap, permettant ainsi d'en décrire les situations avec suffisamment de précision. L'analyse des résultats de VQS a permis de construire un indicateur synthétique des réponses en 6 modalités de handicap croissant :

Groupe 1 (78.6 %) : personnes déclarant ne souffrir d'aucune difficulté ;

Groupe 2 (6.6 %) : personnes déclarant une seule difficulté ;

Groupe 3 (4.4 %) : personnes déclarant « avoir un handicap » ou « avoir demandé une reconnaissance » ou souffrir d'une « limitation d'activité » ou dépendre d'une aide humaine ou souffrir de plusieurs autres difficultés.

Groupe 4 (2.6 %) : personnes déclarant « avoir un handicap » ou « avoir demandé une reconnaissance », et personnes déclarant souffrir d'une « limitation d'activité », déclaration appuyée par des items d'aide humaine ou technique ou plusieurs autres.

Groupe 5 (3.8 %) : personnes déclarant « avoir un handicap » ou « avoir demandé une reconnaissance », déclaration fortement appuyée par d'autres items ;

Groupe 6 (4 %) : personnes déclarant avoir obtenu une reconnaissance de leur handicap (plus, pour les moins de 16 ans : enfants et adolescents inscrits dans une classe ou un établissement spécialisé).

Le tirage des individus HID dans l'échantillon des répondants VQS est stratifié par VQS et par l'âge (en deux modalités) avec des probabilités fortement inégales . En effet, les personnes appartenant au groupe 6 ont été tirées selon un taux de sondage élevé, alors que celles du groupe 1, à plus grand effectif dans l'ensemble de la population, ont eu au contraire un taux de sondage minimal.
On a tenu à garder un groupe témoin (groupe 1) et d'y mener l'enquête détaillée même si les répondants ont déclaré n'avoir aucune difficulté à signaler.
L'éventail des probabilités de tirage de cette étape varie presque de 1 à 100.

II.1.La stratification « HID ».

Il fut intéressant de prendre en compte l'âge des personnes dans la stratification car les réponses à VQS ont permis d'estimer que dans la population totale, la répartition des individus dans les différents groupes VQS était très différente selon qu'il s'agissait d'une personne de plus de 70 ans ou de moins de 70 ans.
Voici la répartition par groupe de l'échantillon des répondants VQS et celle de la population totale obtenue par pondération.

Chez les moins de 70 ans		
	Echant. VQS	Popul. totale
Groupe 1	83.8 %	83,7
Groupe 2	6.0 %	6,1
Groupe 3	3.3 %	3,4
Groupe 4	1.3 %	1,4
Groupe 5	2.1 %	2,1
Groupe 6	3.4 %	3,4

Chez les 70 ans et plus		
	Echant. VQS	Popul. totale
Groupe 1	33.6 %	34,2
Groupe 2	12.1 %	12,1
Groupe 3	13.4 %	13,2
Groupe 4	13.6 %	13,2
Groupe 5	18.4 %	18,4
Groupe 6	9.1 %	8,8

C'est pourquoi dans l'échantillon HID, on a cherché à avoir des proportions par groupe différentes selon la classe d'âge. L'échantillon HID présentait la répartition suivante :

Chez les moins de 70 ans :

13% issus du groupe 1 ;
7% issus du groupe 2 ;
13% issus du groupe 3 ;
7% issus du groupe 4 ;
20% issus du groupe 5 ;
40% issus du groupe 6 .

Chez les 70 ans et plus :

15% issus du groupe 1 ;
24% issus du groupe 2 ;
16% issus du groupe 3 ;
11% issus du groupe 4 ;
10% issus du groupe 5 ;
25% issus du groupe 6 .

Le tirage a donc été effectué selon 10 strates croisant les six groupes VQS et un ou deux groupes d'âge selon les groupes VQS.

Strate 01 : groupe VQS 1 et âge < 70 ans ;
Strate 02 : groupe VQS 1 et âge >= 70 ans ;
Strate 03 : groupe VQS 2 et âge < 70 ans ;
Strate 04 : groupe VQS 2 et âge >= 70 ans ;
Strate 05 : groupe VQS 3 de tout âge ;
Strate 06 : groupe VQS 4 de tout âge ;
Strate 07 : groupe VQS 5 et âge < 70 ans ;
Strate 08 : groupe VQS 5 et âge >= 70 ans ;
Strate 09 : groupe VQS 6 et âge < 70 ans ;
Strate 10 : groupe VQS 6 et âge >= 70 ans .

II.2.Le tirage de l'échantillon HID.

II.2.1.Les probabilités de tirage.

Cette stratification a été effectuée au sein de chaque strate géographique de VQS ; on rappelle que ces strates sont des regroupements de départements de taille régionale ou infra - régionale. Dans cette deuxième phase du plan de sondage, on n'a pas tenu compte des spécificités propres à l'enquête EHF puisque les deux enquêtes n'avaient plus rien en commun ; le nombre de strates géographiques a donc été ramené à 31 sur l'ensemble de la métropole.

La première étape a été de déterminer les probabilités de tirage au sein de chaque strate HID (au nombre de 10), ce pour chacune des strates VQS (au nombre de 31).
En plus du fait que les individus soient surreprésentés par handicap croissant, deux autres raisons ont fait que les probabilités de tirage étaient très inégales.

Tout d'abord, selon la strate VQS.

- Les 23 strates où l'échantillon VQS a été tiré sans surreprésentation locale représentaient environ 79.5 % de la population métropolitaine habitant en domiciles ordinaires (soit 45 600 000 personnes au 1er décembre 1999, date centrale de la collecte). Elles ont regroupé 216 000 réponses exploitables à VQS, soit un taux de sondage de 1 / 211. On y a recueilli 12 436 réponses à HID, soit un taux HID / VQS de 1 / 17 et un taux global de sondage pour HID d'environ 1 / 3670.

- Les 7 strates où VQS seule a été surreprésentée (mais pas HID) comptaient environ 10 960 000 habitants en domiciles ordinaires. Elles ont regroupé environ 128 000 réponses exploitables à VQS, soit un taux de sondage de 1 / 86. On y a recueilli 3 034 réponses à HID, soit un taux HID / VQS de 1 / 42 et un taux global de sondage pour HID de 1 / 3612, voisin donc des 23 premières strates.

- L' Hérault (875 000 habitants en domiciles ordinaires), qui a bénéficié d'une extension d'échantillon pour VQS et pour HID, a compté pour sa part 16 150 réponses exploitables à VQS (soit un taux de sondage de 1 / 54) et 1 480 réponses à HID, soit taux HID / VQS de 1 / 10.9 et un taux global de sondage pour HID de 1 / 590, plus de six fois supérieur aux précédents.

Enfin, le nombre de personnes à enquêter par région dépendait en partie de la capacité locale du réseau d'enquête , car une des différences avec l'enquête précédente est que les questionnaires VQS étaient remis aux personnes au moment du recensement de la population de mars 1999 puis récupérer ultérieurement, alors que pour l'enquête HID, des enquêteurs allaient sur place interroger les personnes.
Ce point sera expliqué plus en détail et sera accompagné d'un exemple dans le paragraphe suivant.

Toutes ces raisons ont donc fait que les probabilités de tirage à l'intérieur de chaque strate HID étaient très inégales.
On présentera les moyennes, extrêmes et dispersion des taux de sondage par strate HID dans le chapitre suivant.

Toutefois, dans chaque zone d'enquête (et donc dans l'ensemble du tirage HID dans l'échantillon VQS), les probabilités de tirage dans les 10 strates étaient proportionnelles à une échelle fixe (les coefficients de tirage). C'est-à-dire que d'une zone à l'autre on n'avait pas la même probabilité de tirage pour la même strate HID, mais le vecteur des dix probabilités pour une zone était colinéaire au vecteur ci-dessous :

	Coefficients de tirage
strate 01	0,68
strate 02	18,75
strate 03	5,62
strate 04	19,44
strate 05	20
strate 06	30
strate 07	52
strate 08	32
strate 09	65
strate 10	40

II.2.2.La réalisation concrète du tirage.

On a réalisé ce tirage en partant du découpage des régions en zones de délégués que l'on avait utilisé pour le sondage VQS. Les zones qui ont été enquêtées restent celles qui ont été sélectionnées au cours du tirage VQS ; par contre, comme seul un département a demandé une extension d'échantillon pour l'enquête HID, on n'a pas tenu compte de toutes les zones de délégués qui avait été tirées en surplus dans certains départements dans le but de réaliser l'extension VQS. Par exemple, dans la région PACA, 27 ZD avaient été tirées à cause de l'extension d'échantillon demandée par les Bouches du Rhône ; mais on a proposé de tirer des individus pour l'échantillon HID que dans 22 ZD.

Les répondants VQS ont été regroupés par ***zones d'enquête***. Le plus souvent, une zone d'enquête correspondait aux répondants VQS d'une ZD, mais elle pouvait aussi regrouper deux ZD. Dans la région 54, il n'y avait qu'une seule zone d'enquête.
Dans chacune de ces zones d'enquête, le but était de tirer un nombre déterminé d'individus et, selon l'effectif, un ou plusieurs enquêteurs serait chargé d'aller à leur domicile les interroger.

Le responsable de l'enquête à l'INSEE de Paris a déterminé le nombre d'individus qu'il voudrait dans chaque zone d'enquête en fonction du nombre de répondants VQS obtenus dans la zone, puis a demandé l'avis des directions régionales (DR).

En effet il était nécessaire de négocier avec les DR car le nombre de personnes à enquêter dans une zone dépendait de la charge de travail, de la compétence de l'enquêteur, de sa rapidité ; s'il n'avait pas suffisamment d'expérience, on réduisait son effectif et quand cela était possible, d'autres enquêteurs étaient chargés d'interroger davantage de personnes. De plus, il fallait être suffisamment formé pour pouvoir aborder un sujet aussi délicat que les incapacités physiques.

Il pouvait également arriver que sur une zone d'enquête il n'y ait pas d'enquêteur parce que le précédent était parti à la retraite et que la DR n'a pas pu le remplacer. Dans ce cas la DR a proposé à l'INSEE de Paris une ou deux autres zones d'enquête parmi celles qui avaient été tirées pour l'enquête VQS et que l'on avait mis de côté.
Prenons le cas de la région PACA. L'INSEE a proposé à la DR de tirer un échantillon de 26 personnes dans la zone de PERI et 26 personnes dans la zone de TALLONE. Mais la DR ne disposant pas d'enquêteur dans ces zones pour la collecte HID, elle a alors proposé de tirer les 52 personnes HID attendues plutôt dans l'une des zones d'Aix en Provence.

L'effectif HID par zone convenu après discussion oscillait autour de 50 individus, avec un minimum de 26 et pouvait aller jusqu'à 90 dans des régions comme l'île de France.

Dans des zones rurales très dispersées, il n'était pas possible d'interroger beaucoup de personnes car cela aurait été beaucoup trop coûteux à cause des frais de déplacement notamment. Prenons le cas du Limousin. L'ensemble de ses zones d'enquête n'a couvert qu'un tiers de sa superficie et a représenté 90% de sa population.

En ce qui concerne l'Hérault, 8 zones de délégués avaient été tirées pour le sondage VQS, 7 ont été conservées pour le sondage HID, et chacune présentait un très gros effectif HID (260 individus en moyenne) afin d'obtenir un échantillon de 1800 personnes et réaliser les estimations locales. Dans ce cas, la zone de délégués comprenait plusieurs zones d'enquête et un enquêteur était chargé d'interroger 20 à 70 personnes.

Pour se fixer les idées : il y avait un peu plus de 400 zones d'enquête avec un effectif moyen de 50 individus présents dans l'échantillon HID ; globalement, une zone d'enquête représentait donc 130 000 personnes, avec évidemment nettement moins dans les zones rurales et plus dans les régions à forte densité de population, sauf dans le département de l'Hérault où une zone d'enquête représentait plutôt 15 000 personnes.

Pour la réalisation concrète du tirage, dans une zone d'enquête, on a classé les répondants à VQS selon les 10 strates HID. On a calculé le nombre d'individus à tirer dans chaque strate en appliquant les différents taux de sondage et on a tiré ces individus suivant une méthode de sondage aléatoire simple dans chacune des 10 strates.

Diverses particularités ont provoqué quelques rectifications des taux de sondage dans de rares zones d'enquête :

- d'abord, il pouvait arriver que dans une zone le nombre d'individus du groupe 6 (les plus sévèrement handicapés) soit insuffisant pour fournir le nombre d'individus HID voulu ; on a dû compenser en augmentant arbitrairement le taux de sondage dans les groupes 1 et 2 de la même zone d'enquête ;

- ensuite, après contrôle, lorsque plus de deux individus tirés faisaient partie du même ménage (en fait le nombre maximal s'est établi à trois), le moins handicapé des trois était retiré de l'échantillon HID ;

- enfin, les effectifs tirés par strate HID étant généralement faibles dans chaque zone d'enquête, les effets d'arrondi ont assez sensiblement modifié les taux réels re tirage.

Par exemple dans la Champagne - Ardenne, la zone de Montry ne contenait pas suffisamment de répondants VQS du groupe 6, donc on a été obligé de tirer 15 personnes supplémentaires dans le groupe 1. Après contrôle, on s'est rendu compte qu'il y avait des cas où il y avait trois personnes tirées dans le même ménage et ce cas on a gardé celles de groupe le plus haut et d'âge le plus faible.

L'échantillon de l'enquête était finalement constitué de 21 760 personnes vivant dans 20 116 ménages. On a recueilli 16 945 réponses à HID, soit un taux de déchet de 22.2 % en raison des refus, déménagements à une adresse inconnue, absence pendant toute la période de collecte, décès et entrée en institution sanitaire et sociale depuis la date du recensement (donc hors du champs de l'enquête).

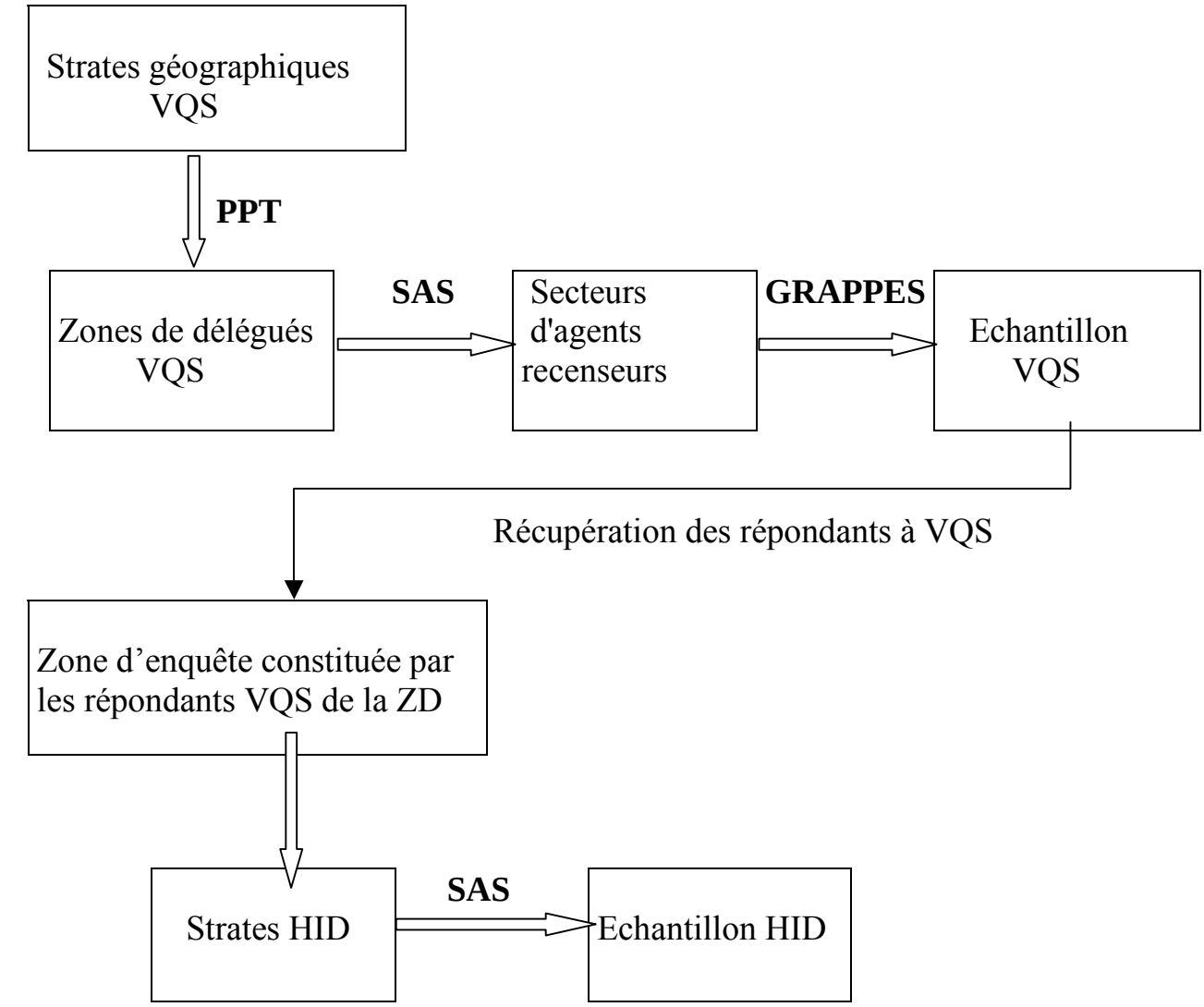
II.3. Résumé.

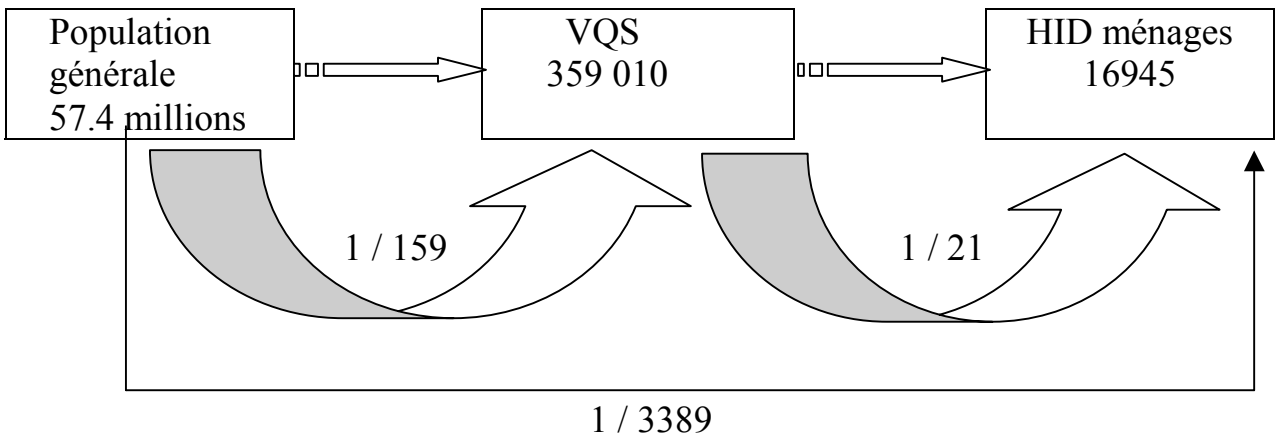
La deuxième phase du plan de sondage est un sondage stratifié par VQS dans chaque zone d'enquête, à probabilités inégales, avec un tirage d'individus par sondage aléatoire simple dans chaque croisement zone d'enquête * strate HID.

Le schéma de la page suivante résume les différentes étapes du plan de sondage de l'enquête HID.

PPT = tirage avec une probabilité proportionnelle à la taille ;
SAS = tirage par sondage aléatoire simple.

III. Synthèse.





Chapitre 2
Etude des pondérations de l'enquête

On rappelle que le plan de sondage mis en œuvre par l'INSEE pour réaliser l'enquête HID est un plan de sondage en deux phases avec post - stratification :
1ère phase : pré enquête VQS, stratifiée, à deux degrés et aréolaire ;
2ème phase : enquête HID, stratifiée par VQS et à probabilités inégales.

L'enquête VQS ayant été effectuée pendant le recensement de la population de mars 1999, on ne savait pas a priori le nombre exact d'individus qui seraient interrogés par VQS. La contrainte était d'enquêter environ 1 800 personnes dans chaque département, et 20 000 personnes en cas d'extension VQS. Faute d'informations au moment du tirage, nous sommes amenés à calculer a posteriori les probabilité de tirage des individus de l'échantillon.

Le but de ce chapitre est de déterminer les probabilités de tirage étape par étape afin d'établir les pondérations des individus ayant répondu à HID. Nous allons dans un premier temps calculer les pondérations des individus de l'échantillon d'enquête VQS, puis dans un deuxième temps calculer les pondérations HID. Enfin, nous établirons les pondérations finales.

I. Les pondérations VQS.

L'enquête VQS a été menée suivant un plan de sondage stratifié et à deux degrés. 36 strates géographiques ont été déterminées. Dans chaque strate, on a effectué le tirage de zones de délégués, puis on a tiré des secteurs d'agents recenseurs à l'intérieur de chaque zone de délégués, enfin toutes les personnes en ménage dans le secteur tiré ont été enquêtées. Toutefois, la strate 13 n'a pas été concernée par l'enquête VQS.

I.1. Probabilité de chaque zone de délégués d'être sélectionnée.

I.1.1. Le choix de la méthode de tirage.

Le premier degré de tirage a été effectué en deux phases:

- Au sein d'un département, on a classé les districts selon une nomenclature socio-économique, puis on a tiré systématiquement un individu selon un pas calculé en fonction de la population 1990 du département et du nombre de zones de délégués (ZD) voulu en cas d'extension VQS. Les districts contenant ces individus ont constitué des points d'entrée, les ZD sélectionnées étant celles incluant ces points d'entrées.

- Ensuite on a regroupé par région les ZD sélectionnées, on les a classées selon une nomenclature de structure familiale, et parmi elles, on a tiré les ZD qu'il fallait dans le cas où il n'y a pas d'extension. Enfin, quand la liste des départements intéressés par

l'extension VQS a été connue, on a conservé toutes les ZD sélectionnées au cours du premier tirage dans le département concerné par l'extension.

En fait, un nombre de ZD à tirer a été défini dans chaque strate alors même que le découpage géographique en zones n'était pas terminé et donc que le nombre total de ZD par strate était encore inconnu. C'est la raison pour laquelle on a tiré des « points d'entrée » à partir du fichier des districts établi à partir du recensement de la population de 1990, et qu'on a décidé que les ZD sélectionnées seraient celles incluant ces districts.

Qu'est -ce- qu'une zone de délégués ?
Pour réaliser le recensement de la population (RP), l'INSEE découpe les départements en petites portions de tailles voisines afin de répartir le travail de collecte entre des délégués qui seront chacun chargé d'une portion. Ces portions de départements regroupent en moyenne 6660 ménages, soit 16200 personnes et sont appelées zones de délégués. Donc le premier travail à effectuer en amont du RP est cette cartographie.
Ensuite, chaque délégué s'occupe du recrutement et de la formation d'agents qui seront chargés de collecter à domicile les informations. Chaque agent aura à sa charge un secteur, donc le deuxième travail consiste à découper la zone de délégués en secteurs d'agents recenseurs (AR). Ces secteurs comprennent en moyenne 215 ménages, soit 530 personnes.

Pour se fixer les idées :
1 ZD → 30 secteurs d'AR → 90 districts RP → 6660 ménages → 16200 personnes.
1 secteur d'AR → 3 districts RP → 215 ménages → 530 personnes.
1 district RP → 72 ménages → 175 personnes.
1 ménage → 2.4 personnes.
RP métropole 1999 → 58 000 000 personnes en ménages
→ 23 770 000 ménages
→ 330 000 districts
→ 110 000 secteurs d'AR
→ 3568 ZD.

En ce qui concerne le RP de mars 1999, le travail de cartographie s'est terminé vers le mois d'octobre 1998, alors qu'il fallait sélectionner les ZD qui participeraient à l'enquête VQS dès le mois de juin 1998 à cause de tout le travail que demande une enquête. Il n'était pas possible de commencer plus tôt le travail de cartographie car ce travail doit prendre en compte toutes les habitations de la commune. En effet, le secteur d'un agent est clairement défini en terme d'habitations, par exemple : l'AR n° 5 a en charge la rue Anatole France du n° 1 au n° 15, uniquement le trottoir de gauche, puis la rue Z etc. Donc si le découpage est effectué trop tôt, on risquerait d'ignorer les nouvelles constructions car pour que les ZD de 1999 fussent définies à temps pour le tirage de juin 1998, il aurait fallu commencer le travail un an à l'avance, ce qui aurait été trop prématuré.

La préparation de l'enquête s'est déroulée de la façon suivante :

- en juin 1998 : tirage des points d'entrée à la direction générale (DG) de l'INSEE à Paris;
- consultation des directions régionales (DR) pour savoir si la zone qui sera définie autour de ces points leur convient, il était question notamment de savoir s'il y a un enquêteur en poste dans la zone pour l'enquête HID ; dans le cas inverse, certaines zones ont été remplacées par d'autres zones de même structure urbaine ;

- en octobre, le travail de cartographie était achevé donc les ZD sélectionnées pour l'enquête VQS étaient connues ;
- au début du mois de novembre, les DR ont envoyé à la DG la liste des secteurs d'AR des ZD sélectionnées ;
- tirage des secteurs d'AR ;
- novembre / décembre : formation des personnes qui encadreront le RP avec une formation supplémentaire pour les délégués VQS ;
- janvier / février 1999 : formation des AR ;
- 08 mars 1999 : début du RP et de l'enquête VQS.

La DG a été obligée de tirer les points d'entrée avant la fin de la cartographie en raison du peu de temps dont elle disposait entre le mois de novembre et le mois de mars pour réaliser tout le travail qu'a demandé cette enquête : formation des délégués VQS, acheminement des bulletins de réponses, délais matériels, etc.

C'est pourquoi la DG a utilisé le fichier des districts 1990 pour le tirage des ZD et n'a pas effectué le tirage avec les ZD 1999. Il est à préciser que les responsables de l'enquête à la DG désiraient une certaine représentativité socio-démographique de l'échantillon, donc le tirage systématique trié était tout à fait approprié. Tirer directement les ZD n'aurait pas offert cette représentativité. Par ailleurs, il n'était pas possible d'effectuer le tirage systématique trié à partir d'une base de 1999 puisque le RP n'étant pas réalisé, aucune population 1999 n'était connue, pas plus que les caractéristiques socio-démographiques.
En somme, réaliser la sélection des ZD VQS par le tirage de points d'entrée à partir du fichier des districts 1990 était la méthode la plus appropriée.

I.1.2. Calcul des probabilités de tirage.

Comme le tirage est stratifié, la probabilité de tirage des ZD dépend a priori de la strate. Dans une strate s_j (j = 1, ….., 36), on tire un nombre fixé n_j de points d'entrée.
La probabilité P_z de tirer la zone de délégués z dans la strate s_j est la somme des probabilités P_i de tirer les districts i qui constituent cette zone moins la somme des probabilités P_{ij} d'intersection (tirer les districts i et j dans la zone z):

$$P_z = \sum_{i \in z} P_i \; - \; \sum_{i \in Z} \sum_{j \neq i} P_{ij}$$

La probabilité de tirer simultanément les districts i et j dans la zone z est presque nulle donc on approxime la probabilité P_z par :

$$P_z \approx \sum_{i \in z} P_i$$

Le tirage des districts a été effectué en deux phases : tirage départemental puis tirage régional.
Cherchons la probabilité de tirage au cours du tirage départemental.
Le tirage systématique a assuré à chaque individu du département la même probabilité de tirage donc le tirage des points d'entrée (districts) s'est effectué proportionnellement à la taille de leur population en 1990. Ainsi, la probabilité de tirer un district i dans le département d est définie par :

$$P_{i1} = ndvqs \times \frac{X_{ip90}}{X_{dp90}}$$

Avec : X_{ip90} = la population du district 1990 au RP de 1990 ;
X_{dp90} = la population du département au RP de 1990 ;
ndvqs = nombre de ZD à tirer dans le cas d'une extension VQS.

Pour réaliser la deuxième phase du tirage, les districts tirés au cours du premier tirage ont été regroupés par région afin de tirer le nombre de ZD qu'il fallait dans le cas où il n'y a pas d'extension. Les ndvqs districts sélectionnés dans un département ont été affectés d'une population de $\frac{X_{dp90}}{ndvqs}$. Ici également, le tirage systématique a assuré à chaque individu de la région la même probabilité de tirage donc le tirage des districts s'est effectué proportionnellement à leur taille. Ainsi, la probabilité de tirer le district i dans la région r est définie par :

$$P_{i2} = ndehf \times \frac{\frac{X_{dp90}}{ndvqs}}{X_{rp90}}$$

Avec : X_{rp90} = la population de la région au RP de 1990 ;
ndvqs = nombre de ZD tirées dans le département d'où provient le district i ;
ndehf = nombre de ZD à tirer dans la région .

Finalement, la probabilité de tirer le district i est le produit des probabilités des deux phases :

$$P_i = P_{i1} \times P_{i2} = \left(ndvqs \times \frac{X_{ip90}}{X_{dp90}} \right) \times \left(ndehf \times \frac{\frac{X_{dp90}}{ndvqs}}{X_{rp90}} \right)$$

$$= ndehf \times \frac{X_{ip90}}{X_{rp90}}$$

Pour revenir à la strate, s'il s'agit d'une strate départementale (extension VQS), on conserve toutes les ZD sélectionnées au cours du premier tirage.
Finalement, la probabilité de tirer le district i dans la strate s_j est définie par :

$$P_i = n_j \times \frac{X_{ip90}}{X_{sp90}}$$

Avec X_{sp90} = la population de la strate au RP de 1990 ;
n_j = le nombre de ZD à tirer dans la région .

Donc : $$P_z \approx n_j \times \frac{\sum_{i \in z} X_{ip90}}{X_{sp90}} \approx n_j \times \frac{Xz_{p90}}{X_{sp90}}$$

Avec X_{zp90} = la population au RP de 1990 de l'ensemble des districts constituant la ZD z .

Ainsi, on considérera que la probabilité de tirage d'une ZD est proportionnelle à la taille de sa population en 1990.

I.1.3. La démarche de recherche de la population en 1990 des ZD.

La difficulté dans le calcul des probabilités de tirage des zones de délégués réside dans le fait que l'on ne connaît pas a priori les populations en 1990 des zones de délégués obtenues par la cartographie pour le RP de 1999. En effet, le RP est effectué en moyenne tous les sept ans donc, en raison des changements qui peuvent apparaître dans la configuration territoriale, il est nécessaire d'effectuer le travail de découpage en ZD à chaque RP. Quand on parle de changements, on fait allusion notamment à l'apparition de nouveaux quartiers résidentiels, à la destruction d'immeubles, à la délocalisation d'habitations (par exemple, un petit village peut avoir été détruit et ses habitants relogés ailleurs en vue de la construction d'une ligne TGV), à l'exode rural, aux nouvelles constructions. Donc il n'est pas étonnant que les ZD 1999 soient très différentes des ZD 1990. Une fois les ZD concernées par l'enquête VQS identifiées, il reste à déterminer leur population de 1990.

Le RP de 1999 a permis de former de nouveaux districts (districts 1999) appartenant aux ZD 1999. Pour trouver la population en 1990 des ZD 1999, il faut reconstituer l'ensemble des districts 1999 de ces ZD et rechercher leur population en 1990.
391 ZD ont été concernées par l'enquête VQS et 35476 districts relèvent de ces ZD. Ces districts sont différents des districts 1990 donc on ne connaît pas leur population en 1990.

Trois configurations sont possibles :

- soit les districts appartiennent à l'une des 844 grosses communes ;
- soit ils appartiennent à une petite commune totalement incluse dans la ZD ;
- soit ils appartiennent à une petite commune à cheval sur plusieurs ZD.

Dans le premier cas, il faut rechercher la population 1990 îlot par îlot car une grosse commune peut déborder sur plusieurs ZD. Un îlot au sens du RP est une très petite portion de commune que ne traverse aucune route, par exemple il peut s'agir d'un pâté de maisons, d'un immeuble. De même, la configuration d'une commune en îlots peut changer d'un RP à l'autre donc on ne connaît pas la population 1990 des îlots 1999.
Monsieur Houssay, qui travaille sur le recensement à la DG, a construit un fichier qui s'appelle « la table de passage de Houssay » dans lequel il convertit les îlots 1999 en proportions d'îlots 1990 relativement à leur surface, c'est à dire que pour chaque îlot 1999 d'une grosse commune, il explique de quel(s) îlot(s) 1990 il provient et dans quelles proportions, et à partir de ces données, on obtient la population 1990 de l'îlot 1999.
159 grosses communes ont ainsi été concernées.

Par exemple : le district n° AAA contient les îlots AAA1, AAA2, etc. On recherche dans la table la correspondance en îlots 1990 et on lit que l'îlot AAA1 contient 100 % de l'îlot 1990 n° X1, 20 % de l'îlot X2 et 10 % de l'îlot X3. Donc la population 1990 de l'îlot 1999 AAA1

vaut : POPX1 + 0.2*POPX2 + 0.1*POPX3 où POPX1 est la population 1990 de l'îlot 1990 X1.
En ce qui concerne les nouveaux quartiers, il suffit de dire que les îlots 1999 proviennent d'îlots 1990 dont la taille de la population vaut zéro.

Quand le district appartient à une commune totalement incluse dans la ZD, on recherche directement la population 1990 de cette commune. Puis il reste à retrouver la population 1990 des autres districts.

Si le district appartient à une commune à cheval sur plusieurs ZD, nous sommes obligés de traiter la partie qui nous intéresse îlot par îlot car on connaît la population 1990 soit par commune, soit par îlot. Malheureusement, on n'a la correspondance d'un îlot 1999 en îlots 1990 uniquement s'il s'agit d'une grosse commune. Donc pour les communes plus petites à cheval sur plusieurs ZD, on est obligé d'estimer la population 1990 des îlots 1999.
Le RP 1999 a permis d'avoir la population 1999 de l'îlot 1999 (notée POP99_îlot99) et de la commune (POP99_com). On dispose également de la population 1990 de la commune obtenue par le RP 1990 (POP90_com).
On estime la population 1990 de l'îlot 1999 par :

$$\text{POP90\_îlot99} = \text{arrondi} \left(\text{POP99\_îlot99} \times \frac{POP90\_com}{POP99\_com} \right)$$

Ainsi on prend en compte le taux d'évolution de la taille de la commune, taux que l'on suppose uniforme sur toute la commune.

La population 1990 de l'ensemble des ZD VQS est de 6 196 852.
Au final :

- 39 % (2 392 006) ont été déterminés à l'aide du fichier de passage de Houssay ;
- 59 % (3 491 851) proviennent d'une évaluation directe des population de petites communes ;
- seulement 5 % (312 995) proviennent d'une estimation.

I.2. Probabilité de chaque secteur d'agent recenseur d'être sélectionné.

I.2.1. Les différentes particularités des zones de délégués.

La probabilité de tirage du secteur d'AR dépend de sa zone de délégué. Cependant, afin d'avoir des taux de sondage uniformes, le nombre de secteurs à tirer dans chaque zone dépend du nombre de secteurs total.
Rappelons que le tirage des zones de délégués pour l'enquête VQS a été associé à celui de l'enquête EHF (Etude de l'Histoire Familiale). En fait, on a tiré des ZD communes aux deux enquêtes, et comme l'échantillon EHF devait être plus grand que l'échantillon VQS, on a tiré plus de secteurs d'AR pour EHF dans les ZD sélectionnées ; un AR ne s'est occupé que d'une seule enquête.

Nous distinguons quatre types de zones de délégués :

- zone sans aucune extension ;
- zone avec extension VQS ;

- zone avec extension « langue » ;
- zone avec extension VQS et extension « langue » (cas du département 62).

Par ailleurs, certaines ZD ont été rétrécies après avoir été sélectionnées. Dans ce cas, le tirage des secteurs d'agents recenseurs ne s'effectue qu'à l'intérieur de la nouvelle zone, donc la probabilité de tirage augmente.

I.2.2. Calcul des différents taux de sondage.

Le tirage des secteurs d'AR dans la ZD s'est effectué par sondage aléatoire simple donc la probabilité de tirage d'un secteur est égale au taux de sondage dans la ZD.

Pour réaliser les deux sondages (EHF et VQS), on a décidé de tirer un secteur d'AR sur quatre. Le nombre total moyen de secteurs d'AR par ZD est de 30.8, donc on décide de tirer 7.7 sur les 30.8 pour les deux sondages.

Cas 1 : zone sans aucune extension.

Le but était d'obtenir un échantillon de 1800 personnes pour l'enquête VQS par département. En moyenne, 2 ZD ont été tirées par département dans le cas où il n'y a pas d'extension VQS ; un secteur d'AR comporte en moyenne 560 personnes vivant en ménage ordinaire donc le nombre de secteurs à tirer par ZD vaut : 1800 / (2 * 560) = 1.6

On tire donc 1.6 secteurs d'AR sur les 30.8

Dans ce cas, le taux de sondage pour l'enquête VQS vaut : $\boxed{txv = \frac{1.6}{30.8}}$

Pour déterminer le nombre n de secteurs d'AR à tirer dans une ZD, on applique ce taux de sondage au nombre N de secteurs total de la zone :
n = txv * N, le taux de sondage vaut bien taux = n / N = (txv * N) / N = txv .
Ainsi le nombre de secteurs d'AR tirés par ZD est proportionnel au nombre de secteurs d'AR total de la zone.

Cas 2 : zone avec extension VQS.

S'il s'agit d'un département ayant demandé une extension d'échantillon VQS, on devait tirer un échantillon plus important, donc davantage de secteurs d'AR dans le département. Cet accroissement a été réparti en une augmentation du nombre de ZD tirées dans l'étape précédente, et en une augmentation de la proportion des secteurs d'AR tirés pour l'enquête VQS dans chacune de ces ZD. Corrélativement, la proportion des secteurs EHF dans chaque ZD a diminué, mais de façon à ce que leur nombre total sur le département reste stable par rapport aux autres départements.

On compte le nombre N_1 de ZD que l'on aurait eu sur le département s'il n'y avait pas eu d'extension VQS et le nombre N_2 de ZD que l'on a effectivement sélectionnées compte tenu de l'extension VQS.
En fait, le nombre N_1 prend en compte uniquement les ZD sélectionnées au cours du deuxième tirage et le nombre N_2 représente le nombre N_1 plus le nombre de ZD conservées pour l'extension VQS.

S'il n'y avait pas d'extension VQS, le nombre de secteurs d'AR tirés pour EHF dans le département serait : $N_1 * 6.1$
Dans le cas d'une extension, ce nombre peut s'écrire : $N_2 * 6.1$
Afin de tirer plus d'AR pour VQS tout en conservant le même nombre d'AR EHF que dans les autres départements, on décide donc de tirer dans chaque ZD : $6.1* N_1 / N_2$ secteurs , ainsi le nombre de secteurs EHF tirés dans le département vaut : $N_2 * (6.1* N_1 / N_2) = N_1 * 6.1$

Pour trouver le nombre d'AR VSQ à tirer, on part du principe qu'il faut dans chaque département à extension VQS un échantillon d'au moins 20 000 personnes. Dans ces départements, on a tiré un minimum de 8 ZD ; chaque ZD comporte en moyenne 30.8 secteurs d'AR et un secteur représente en moyenne 560 personnes vivant en ménage ordinaire. Donc il fallait 20000 / 560 = 35.7 secteurs d'AR à tirer dans le département et à répartir dans les 8 ZD. Par conséquent, dans chaque ZD, il fallait tirer 35.7 / 8 = 4.5 secteurs d'AR pour l'enquête VQS.
On calcule la moyenne N_1 / N_2 :
Dans le cas général, on a décidé de tirer en moyenne une ZD sur 10, alors qu'en cas d'extension VQS on a décidé de tirer une ZD sur 4 en moyenne. Donc en moyenne, N_1 / N_2 vaut 0.4 et le nombre moyen de secteurs EHF tirés par ZD vaut 2.3
Finalement, sur les 30.8 secteurs d'AR de la ZD, il fallait tirer 6.8 pour les deux enquêtes.
Ainsi le nombre de secteurs VQS tirés dans le département vaut : $N_2 * 6.8 - N_1 * 6.1$ au lieu de $N_2 * 1.6$, ce qui aurait été insuffisant pour obtenir l'échantillon de 20 000 personnes.

Le taux de sondage VQS dans le cas 2 vaut : $$txv = \frac{1}{30.8} \times \left(6.8 - 6.1 \times \frac{N_1}{N_2} \right)$$

Ainsi, txv augmente avec le nombre de ZD rajoutées pour l'extension VQS, la taille de l'échantillon VQS diffère selon les départements.

Cas 3 : zone avec extension « langue » .

L'extension langue concerne une fraction de département et sous cette particularité, plus de ZD ont été tirées, l'objectif étant d'accroître le nombre de secteurs d'AR tirés pour l'enquête EHF dans cette partie du département, tout en gardant un nombre total de secteurs VQS stable sur le département par rapport aux autres départements.

On détermine N_0 le nombre de ZD qui auraient été tirées dans cette partie du département s'il n'y avait pas eu d'extension langue. On considère que le nombre de ZD tirées dépend de la population et que la répartition des ZD tirées est uniforme sur la région. On dispose du nombre de ZD tirées dans le reste de la région, de la population 1990 de la strate et de la population 1990 de la strate à extension langue.

Par exemple, la strate 25 comprend la partie du département 64 (région 72) concernée par l'extension langue et la strate 24 comprend le reste de la région 72. 13 ZD ont été tirées dans la strate 24 alors que 4 ZD ont été tirées dans la strate 25. La population 1990 de la strate 24 est de 2 482 693 habitants alors que celle de la strate 25 est de 248 554 habitants. Donc théoriquement, le nombre de ZD tirées dans la strate 25 serait : $248554 * 13 / 2482693 \approx 1.5$

Donc pour la strate 25, N_0 vaut 1.5 , c'est le nombre de ZD qu'on aurait constaté dans cette partie du département 64 s'il n'y avait pas eu d'extension langue, et N_2 vaut 4, c'est le nombre de ZD que l'on a effectivement tirées compte tenu de l'extension langue.

Donc dans le cas général, le nombre de secteurs d'AR pour l'enquête VQS vaudrait $N_0 * 1.6$
Mais dans le cas 3, il vaut $N_2 * 1.6$
Afin de tirer plus d'AR pour EHF tout en conservant le même nombre d'AR VQS que dans les autres départements, on décide donc de tirer dans chaque ZD : $1.6 * N_0 / N_2$ secteurs , ainsi le nombre de secteurs VQS tirés dans le département vaut : $N_2 * (1.6 * N_0 / N_2) = N_0 * 1.6$

Le taux de sondage VQS vaut : $\boxed{txv = \frac{1}{30.8} \times 1.6 \times \frac{N_0}{N_2}}$

Cas 4 : zone avec extension VQS et extension « langue » (cas du département 62).

Le département du Pas de Calais est le seul à présenter cette double particularité. 4 des 21 ZD sélectionnées dans ce département sont concernées par l'extension langue.
La partie du département 62 concernée par l'extension langue et l'extension VQS forme la strate 15, le reste du département 62 forme la strate 14, cette strate n'est concernée que par l'extension VQS.

Calculons le nombre de ZD qui auraient été tirées dans la strate 15 s'il n'y avait pas eu d'extension langue, par rapport au nombre de ZD tirées dans la strate 14.
17 ZD ont été tirées dans la strate 14 pour une population 1990 de 1 270 096 habitants ; la population 1990 de la strate 15 est de 146 340. donc, théoriquement, le nombre de ZD tirées dans la strate 15 serait de : $146340 * 17 / 1270096 \approx 2$
Donc N_0 vaut 2 et N_2 vaut 4.
Le taux de sondage dans le reste du département 62 vaut 0.162 donc le taux de sondage dans la strate 15 vaut : $0.162 * N_0 / N_1 = 0.081$

Finalement, le taux de sondage dans ce cas vaut $\boxed{txv = 0.081}$

Il reste à traiter le **cas où la zone a été rétrécie** après avoir été sélectionnée.

C'est le cas de 32 % des ZD. On suppose que le rétrécissement de ces zones a été de 20 % .
Appelons Z la zone effectivement sélectionnée et la Zr la nouvelle zone. Soit N(Z) le nombre de secteurs d'agents recenseurs présents dans la zone Z et N(Zr) le nombre de secteurs d'agents recenseurs présents dans la zone Zr. On peut pour ainsi dire que $N(Zr) = 0.8 * N(Z)$.
Donc $N(Z) = 1.25 * N(Zr)$.
Soit n le nombre de secteurs d'AR voulu dans la zone. Alors $txv (Z) = n / N(Z)$.
Mais on ne tire que dans Zr, $txv = txv (Zr) = n / N(Zr)$ donc $txv = 1.25 * txv (Z)$.

Ainsi, la probabilité de tirer un secteur quand la ZD sélectionnée a été rétrécie vaut 1.25 fois la probabilité de tirage que l'on a effectivement calculée.

Finalement, le taux de sondage dans ce cas vaut $\boxed{txv\ effectif = 1.25 * txv}$

On obtient la dispersion suivante dans les 391 zones de délégués VQS pour txv :

CAS	_FREQ_	MOYENNE	MINIMUM	MAXIMUM	ECART
1	271	0.05199	0.05199	0.06494	0.01294
2	100	0.11427	0.08097	0.18314	0.10216
3	16	0.02354	0.01948	0.02922	0.00974
4	4	0.08100	0.08100	0.08100	0.00000
Total	391	0.06705	0.01948	0.18314	0.16533

Dans le cas 1, le maximum de 0.06494 correspond aux zones de délégués rétrécies.
La dispersion des taux de sondage dans le cas 2 vient du fait que le nombre de ZD rajoutées par département en raison de l'extension VQS diffère entre les départements.

I.2.3. Détermination du nombre de secteurs d'AR à tirer dans une ZD.

Les probabilités de tirage txv sont déterminées pour chaque ZD et pour trouver l'effectif n de secteurs à tirer, on applique ce taux au nombre total N de secteurs de la ZD :
n = txv * N. Comme le n calculé est rarement un nombre entier, il faut résoudre des problèmes d'arrondi. Par exemple, si on trouve n = 1.2 on décide de tirer dans la ZD un seul secteur, mais les 0.2 qui manquent sont reportés sur une autre ZD de la strate. Si on trouve que n = 2.7 on décide de tirer 3 secteurs mais on retira les 0.3 qui sont en trop sur une autre ZD de la strate.
Voici un exemple de résolution du problème d'arrondi sur les 11 premières ZD tirées dans la région parisienne. On en tire 34 au total. C'est une strate sans extension et qui n'a pas été rétrécie donc le taux de sondage txv vaut 0.052

point d'entrée	N	n calculé	n arrondi
751080830012	23	1,2	1
751090910051	24	1,2	2
751090930018	26	1,4	1
751111140042C	20	1,0	1
751131330039C	22	1,1	1
751151510142	23	1,2	1
751161620118	23	1,2	1
751171720113	22	1,1	1
751181820122	21	1,1	1
751191920006	24	1.2	2

On vérifie que le total des n calculés est égal au total des n arrondis. Ici ces totaux valent tous les deux 12.
Le point d'entrée correspond au numéro de district et définit la zone de délégués.
Au total, sur les 391 ZD, 763 secteurs d'AR ont été tirés.

I.3. Probabilité d'un individu d'appartenir à l'échantillon.

L'ensemble des personnes vivant en ménage dans un secteur d'AR tiré est interrogé pour l'enquête VQS.
Par conséquent, pour qu'un individu de la population française fasse partie de l'échantillon, il faut d'une part que sa ZD soit sélectionnée dans sa strate, et d'autre part que son secteur d'AR soit tiré. D'après les formules de probabilité, la probabilité de tirer un individu dans une strate est égale à la probabilité de tirer sa ZD dans la strate multipliée par la probabilité de tirer son secteur sachant que sa ZD a été sélectionnée**.**

La probabilité de sélectionner la ZD z dans la strate s_j vaut P (zone / strate) = $P_z = n_j \mathbf{x} \frac{Xz_{p90}}{X_{sp90}}$

La probabilité de tirer le secteur sachant que z est sélectionnée vaut P (secteur / zone) = txv.
Donc la probabilité de tirer l'individu vaut :

P (ind / strate) = P (zone / strate) **x** P (secteur / zone) = $n_j \mathbf{x} \frac{Xz_{p90}}{X_{sp90}}$ **x** txv .

$$\boxed{P \text{ (ind / strate)} = n_j \mathbf{x} \frac{Xz_{p90}}{X_{sp90}} \mathbf{x} \text{ txv}}$$

Que se passe t -il si la zone de délégués a été rétrécie après avoir été sélectionnée ?

Dans ce cas, comme on n'a pas l'ensemble des districts 1999 de la zone initiale Z, on ne peut pas trouver sa population 1990 par les méthodes expliquées au paragraphe I.1.3 mais uniquement celle de Zr.

Le rétrécissement a été de 20 %, on écrit que Zr = 80 % de Z donc la population 1990 de Zr vaut : POP90_Zr = 0.8 * POP_Z donc POP90_Z = 1.25 * POP90_Zr.

P_z = (n_j **x** POP90_Z) / X_{sp90}
= (n_j **x** POP90_Zr) **x** 1.25 / X_{sp90}
= P_{zr} **x** 1.25

Par ailleurs, pour que l'individu appartienne à l'échantillon, il faut non seulement que son secteur soit tiré, mais il faut en plus que son secteur fasse partie de la zone Zr. Donc la probabilité pour que le secteur soit tiré sachant que la zone Z est sélectionnée est égale à la probabilité que le secteur soit dans Zr fois la probabilité de tirer le secteur dans la zone Zr.

La probabilité que le secteur fasse partie de Zr vaut P (Zr / Z) = 0.8
P (secteur / zone) = P (secteur / Zr) **x** P (Zr / Z)
= txv (Zr) **x** 0.8

Donc la probabilité pour un individu d'appartenir à l'échantillon vaut :
P (ind / strate) = P (zone / strate) **x** P (secteur / zone)
= (P_{zr} **x** 1.25) **x** (txv (Zr) **x** 0.8)
= P_{zr} **x** txv (Zr)

Finalement, dans le cas où la ZD aurait été rétrécie après avoir été sélectionnée, il suffit de faire les calculs sur la nouvelle zone : on détermine la population 1990 de Zr et le taux de sondage txv (Zr) est celui calculé à l'étape précédente, c'est à dire le txv effectif.

On définit la pondération des individus de l'échantillon VQS comme l'inverse de cette probabilité de tirage.

I.4. Tableau des pondérations VQS.

Pour des raisons de longueurs, on observera les pondérations VQS uniquement pour 10 ZD :

- 2 dans la strate 1 car il s'agit d'une grande région ;
- 2 dans la strate 3 car il s'agit d'un département à extension VQS ;
- 2 dans la strate 25 pour l'extension langue ;
- 2 dans la strate 14 et 2 dans la strate 15 car il s'agit d'une extension langue et d'une extension VQS dans le même département.

Le tableau suivant présente pour chacune des 10 zone de délégués VQS :

- STRATE : la strate géographique ;
- CODEDEL : le code de la zone de délégués ;
- TXV : la probabilité de tirage des agents recenseurs ;
- POP90_zone : la population en 1990 de la zone ;
- NBZ : le nombre de ZD tiréés dans la strate ;
- POP90_Strate : la population en 1990 de la strate ;
- PZONE : la probabilité de tirage de la ZD ;
- PROBA : la probabilité de tirage des individus habitant dans la ZD ;
- POIDS : la pondération des individus habitant dans la ZD. (POIDS = 1 / PROBA)

STRATE	CODEDEL	TXV	POP90_ Zone	NBZ	POP90_ Strate	PROBA	POIDS
1	DR523022	0.051948	11207	34	7285186	0.002717	368.047
1	DR523028	0.051948	11065	34	7285186	0.002683	372.770
3	DR603057	0.098901	18111	14	1060479	0.023647	42.289
3	DR603058	0.098901	17341	14	1060479	0.022641	44.167
14	DR203149	0.080978	24259	17	1270096	0.026294	38.032
14	DR203155	0.080978	8459	17	1270096	0.009168	109.069
15	DR203187	0.081000	18703	4	146340	0.041409	24.149
15	DR203189	0.081000	20921	4	146340	0.046320	21.589
25	DR083158	0.019481	15144	5	248554	0.005935	168.504
25	DR083163	0.019481	15972	5	248554	0.006259	159.768

On vérifie bien que les pondérations des individus sont moins importantes dans les strates départementales (extension VQS) et infra - départementales (extension langues).
On remarque que le poids des individus pour un même type de zone dépend fortement de la taille de cette zone, voyons le cas des deux exemples de la strate 14.

Sur l'ensemble des 391 ZD VQS, les poids sont très dispersés avec un écart - type de 111.9019 pour une moyenne de 180.8034 ; l'unique ZD tirée dans la Corse (strate 36) enregistre le poids maximum de 886.2090 et le minimum des poids (20.0118) est observé dans la strate 15 qui représente la fraction du département 62 ayant bénéficié d'une extension VQS et d'une extension langue.

II.5. Traitement de la non-réponse VQS.

Parmi les 416 000 individus interrogés par VQS, 359 010 ont fourni une réponse suffisamment complète pour pouvoir être exploitée, soit un taux d'échec de 14% .Ce taux relativement élevé risque d'introduire un biais des estimateurs, si le refus de répondre n'est pas équitablement distribué dans la population. Afin de limiter ce problème, on décide de repondérer les individus répondants en utilisant le principe du calage sur marges. Comme variables auxiliaires (ou marges de calage), on utilise :

- la strate géographique (31 modalités ne tenant plus compte des subdivisions propres à l'enquête EHF) ;

- la tranche d'unité urbaine en 9 modalités (TUU).

On dispose des effectifs N_k de la population métropolitaine vivant en ménage au RP99 pour chacune des modalités k des deux variables auxiliaires. On rappelle cette enquête HID ne concerne que les personnes vivant en domicile ordinaire, par opposition à la vie en collectivité et en institutions. On voudrait que l'échantillon des répondants présente la même structure que la population d'origine pour les modalités de ces deux variables.

Le redressement par calage sur marge consiste à définir un nouveau jeu de pondérations W_i qui permette d'estimer parfaitement chacun des effectifs connus N_k et qui soit construit de telle sorte qu'il s'éloigne le moins possible de la pondération d'origine d_i :

1) $\sum_{i \in R} W_i X_{ik} = N_k$ où R est l'ensemble des répondants et X_{ik} qui vaut 1 si i vérifie la modalité k, 0 sinon, car on a priori $\sum_{i \in R} d_i X_{ik} \neq N_k$.

2) minimiser $\sum_{i \in R} D(W_i, d_i)$ où D(a, b) est la distance entre les nombre réels a et b.

W_i et d_i sont liés par la relation : $W_i = \lambda_{k,l} * d_i$, où $\lambda_{k,l}$ est le coefficient de redressement qui dépend uniquement du croisement de la modalité k de la strate géographique et de la modalité l de TUU. Donc à chacun des répondants d'une même ZD sera affecté le même coefficient de redressement, et par conséquent le même poids redressé.

Pour traiter ce cas, l'INSEE dispose d'un logiciel de calage, appelé CALMAR et mis au point en SAS par O. Sautory, prenant comme paramètres les effectifs marginaux des variables qualitatives sur lesquelles on veut se caler ainsi que les poids initiaux (ici d_i) . Ce logiciel permet de choisir l'expression de la fonction distance et ici, nous utiliserons l'estimation par le raking ratio. Il donne en sortie les valeurs des poids finaux W_i .

On obtient de nouvelles pondérations pour les 359 010 répondants à VQS, de manière à ce que, quand on calcule l'effectif pondéré des répondants pour chacune des modalités des deux variables auxiliaires, on obtienne les effectifs exacts N_k de la population métropolitaine.

Voici les pondérations redressées que l'on obtient pour les individus des dix zones de délégués précédentes, POIDS2 étant le nouveau poids :

STRATE	CODEDEL	POIDS	POIDS2
1	DR523022	368.047	480.441
1	DR523028	372.770	486.607
3	DR603057	42.289	61.684
3	DR603058	44.167	51.487
14	DR203149	38.032	42.346
14	DR203155	109.069	121.440
15	DR203187	24.149	27.404
15	DR203189	21.589	24.498
25	DR083158	168.504	167.665
25	DR083163	159.768	158.973

POIDS2 est en général supérieur à POIDS et quand on calcule l'effectif repondéré pour l'ensemble des répondants VQS, on obtient la population métropolitaine vivant en ménage au RP99, alors que ce n'était pas le cas avec la pondération initiale, du fait de l'échec de collecte.

II. Les pondérations HID.

La deuxième phase du plan de sondage est un tirage stratifié par le « groupe VQS » et par l'âge (en une ou deux modalités) avec des probabilités fortement inégales de façon à sur représenter fortement les plus certainement et sévèrement handicapés. En effet, les personnes du groupe 6 ont été tirées selon un taux de sondage élevé, alors que celles du groupe 1, à plus grand effectif dans la population des répondants à VQS, ont eu au contraire un taux de sondage minimal.

Les répondants à HID ont été regroupés par zones d'enquête, une zone d'enquête correspond en général aux répondants d'une zone de délégués VQS. Dans une zone d'enquête, les répondants à VQS sont classés selon les dix strates HID et le tirage des individus de l'échantillon HID a été effectué par sondage aléatoire simple.
La probabilité de tirage des individus d'une zone d'enquête est donc égale au taux de sondage de sa strate HID dans cette zone d'enquête.

II.1. Détermination des effectifs à tirer dans chaque strate HID pour l'échantillon HID.

Comme il a été expliqué au paragraphe II.2 du chapitre 1, pour diverses raisons, il n'était pas possible d'appliquer des taux de sondage uniformes sur toutes les zones d'enquête. Mais on a pu établir une échelle fixe de façon à sur représenter les personnes les plus handicapées et les probabilités de tirage devaient être proportionnelles à cette échelle ; c'est-à-dire que d'une zone d'enquête à l'autre, on n'avait pas la même probabilité de tirage pour la même strate HID, mais le vecteur des dix probabilités pour une zone devait être colinéaire au vecteur ci-dessous :

	Coefficients de tirage
strate 01	0,68
strate 02	18,75
strate 03	5,62
strate 04	19,44
strate 05	20
strate 06	30
strate 07	52
strate 08	32
strate 09	65
strate 10	40

Le responsable de l'enquête à la DG a convenu, après discussion avec les DR, du nombre d'individus qu'il souhaitait enquêter dans chacune des 336 zones d'enquête.
Soient N l'effectif des répondants à VQS de la zone d'enquête, n le nombre d'individus que l'on veut tirer parmi ces N répondants pour constituer l'échantillon HID et N (j) l'effectif des répondants dans la strate HID j. Deux contraintes sont à respecter dans la répartition de n dans les dix strates HID :

- les probabilités de tirage doivent respecter l'échelle fixée ;
- le taux de sondage global doit être égal à : $\frac{n}{N}$.

A chaque individu correspond une strate HID et à chaque strate HID correspond un coefficient de pondération qui va de 0.68 (strate 01) à 65 (strate 09) , donc à chaque individu correspond un coefficient de pondération, variable notée 'PONDER'. On calcule la somme des valeurs prises par la variable PONDER sur la zone d'enquête, somme notée PONDERC. La

contribution de l'individu i dans cette somme est $\frac{\text{PONDER (i)}}{\text{PONDERC}}$.

Il s'agit d'un tirage uniforme à l'intérieur de chaque croisement zone d'enquête * strate HID, de plus on désire que la probabilité soit proportionnelle à la variable PONDER. Donc il faut choisir comme probabilité de tirer l'individu i :

$P_i = \frac{\text{PONDER (i)}}{\text{PONDERC}}$ **x** n , étant donné que pour tout i, n **x** PONDER (i) < PONDERC ,

ce qui assure à tous les individus d'une strate j la même probabilité de tirage : P (j).

On a bien : $\sum_{i=1}^{N} P_i = n$, soit encore : $\sum_{J=1}^{10} \sum_{i=1}^{N_J} P_i = \sum_{J=1}^{10} N_j$ **x** P (j) = n .

Soit n (j) le nombre d'individus à tirer dans la strate j.

Le taux de sondage dans la strate j vaut donc: taux (j) = $\frac{n(j)}{N(j)}$.

Il s'agit d'un tirage uniforme donc il faut que taux (j) soit égal à P (j).

Par conséquent, il faut que n (j) vaille : $\boxed{n(j) = \frac{PONDER(j)}{PONDERC} \mathbf{x} N(j) \mathbf{x} n}$.

$\boxed{P(j) = \frac{PONDER(j)}{PONDERC} \mathbf{x} n}$, donc les probabilités de tirage sont bien proportionnelles aux

PONDER (j), et le coefficient de proportionnalité vaut : $\boxed{K = \frac{n}{PONDERC}}$

II.2. Calcul des probabilités de tirage.

Pour une zone d'enquête donnée, le tirage des n (j) individus parmi les N (j) répondants VQS de la strate HID j a été effectué par sondage aléatoire simple donc la probabilité des N (j) individus d'appartenir à l'échantillon HID est constante et elle est définie par le taux de sondage :

$$\boxed{P(j) = \frac{n(j)}{N(j)}}$$

II.2.1. Programme de calcul des probabilités de tirage.

On dispose du fichier des 359 010 répondants à VQS ainsi que du fichier des 21 760 individus de l'échantillon HID.
Ces deux fichiers renseignent, pour chaque individu, le numéro de département, de commune, de district, de logement, le prénom et la date de naissance, âge, le groupe VQS, la strate HID.
Le fichier des répondants VQS est le fichier ″ vqsdef ″, et le fichier des individus HID est le fichier ″ hidredrp ″. Il reste à définir les zones d'enquête pour chacun des individus.
Pour effectuer le tirage de l'échantillon HID, 22 programmes (1 par région) de tirage avaient été créés dans lesquels les zones d'enquête ont été définies. On va donc récupérer les définitions des zones d'enquête dans ces programmes.
On rappelle que l'on n'a pas tenu compte de toutes les ZD VQS dans les départements à extension VQS car seul l'Hérault a demandé une extension d'échantillon pour l'enquête HID. Donc parmi les répondants VQS de certaines ZD, on n'a tiré aucun individu HID. C'est pourquoi certaines zones d'enquête définies pour le fichier des répondants VQS n'apparaîtront pas dans le fichier de l'échantillon HID.

Ensuite, la procédure SUMMARY de SAS nous permettra de compter pour chaque zone d'enquête et pour chaque strate HID l'effectif des répondants VQS et l'effectif des individus HID, afin d'établir les taux de sondage.

Le programme SAS nous permettra également d'observer la dispersion des taux de sondage à l'intérieur des strates HID.

Ce programme (c:\user\odile\tirageHID\tauxsondage.sas) est présenté en annexe 3.

II.2.2. Présentation des résultats.

366 zones d'enquête ont été définies mais le tirage de l'échantillon HID s'est effectué dans seulement 336 zones d'enquêtes pour les raisons que nous avons expliquées au paragraphe II.2 du chapitre 1. Regardons la répartition des individus HID dans les zones d'enquête. La table tirage.parzone nous présente le taux global de tirage par zone d'enquête. Nous ne présentons les résultats que pour les 40 premières zones d'enquête :

Taux globaux de sondage par zone d'enquête

Obs	Zone d'enquête	Nombre de répondants VQS	Effectif des individus HID tirés	Taux globaux de sondage
1		359010	21760	0.06061
2	011BOU	665	58	0.08722
3	012DIV	920	55	0.05978
4	021ESS	326	33	0.10123
5	022VIR	742	70	0.09434
6	031MAG	464	44	0.09483
7	032VIC	367	39	0.10627
8	0401DI	949	29	0.03056
9	0402MA	436	24	0.05505
10	04999H	432	0	0.00000
11	0501CH	760	57	0.07500
12	0601CA	550	56	0.10182
13	0602NI	690	66	0.09565
14	071ANN	571	28	0.04904
15	081ASF	422	44	0.10427
16	082CHA	802	82	0.10224
17	101STA	326	32	0.09816
18	102TRO	821	81	0.09866
19	121MIL	543	60	0.11050
20	122MOY	779	62	0.07959
21	1301AI	2735	122	0.04461
22	1302AU	1176	48	0.04082
23	1303FU	2370	60	0.02532
24	1304GR	1499	59	0.03936
25	1305IS	792	56	0.07071
26	1306TA	1291	58	0.04493
27	1307VE	1430	55	0.03846
28	1308MA	603	35	0.05804
29	1309MA	2029	92	0.04534
30	1310MA	1736	117	0.06740
31	1311MA	1863	103	0.05529
32	1312MA	1859	89	0.04788
33	1313MA	1107	59	0.05330
34	141CAE	436	41	0.09404
35	142MOY	378	38	0.10053

36	143ROT	291	29	0.09966
37	14999H	285	0	0.00000
38	151AUR	275	28	0.10182
39	152RIO	252	26	0.10317
40	161COG	1083	85	0.07849

On remarque bien que les taux sont très dispersés pour les différentes raisons expliquées au paragraphe II.2 du chapitre 1. Sur les 336 zones d'enquête dans lesquelles un échantillon HID a été tiré, la moyenne du taux est de 0.0771 avec un écart-type de 0.0316, un minimum de 0.007 et un maximum de 0.1441 pour l'Hérault.

Le tableau qui présente les taux de sondage réels est la table SAS tirage.tauxsond. On n'en présentera qu'un extrait pour des raisons de longueur. Voici les taux de sondage HID pour les quatre premières zones d'enquête :

Taux de sondage HID

Obs	Zone d'enquête	Strate HID	Nombre de répondants VQS	Effectif des individus HID tirés	Taux de sondage réels
1	011BOU	1	472	3	0.00636
2	011BOU	2	33	6	0.18182
3	011BOU	3	33	2	0.06061
4	011BOU	4	12	2	0.16667
5	011BOU	5	26	5	0.19231
6	011BOU	6	18	5	0.27778
7	011BOU	7	14	8	0.57143
8	011BOU	8	16	5	0.31250
9	011BOU	9	34	20	0.58824
10	011BOU	10	7	2	0.28571
11	012DIV	1	810	9	0.01111
12	012DIV	2	11	4	0.36364
13	012DIV	3	45	4	0.08889
14	012DIV	4	5	2	0.40000
15	012DIV	5	16	6	0.37500
16	012DIV	6	5	3	0.60000
17	012DIV	7	14	14	1.00000
18	012DIV	8	1	1	1.00000
19	012DIV	9	10	10	1.00000
20	012DIV	10	3	2	0.66667
21	021ESS	1	257	3	0.01167
22	021ESS	2	10	4	0.40000
23	021ESS	3	17	2	0.11765
24	021ESS	4	4	2	0.50000
25	021ESS	5	12	3	0.25000
26	021ESS	6	10	5	0.50000
27	021ESS	7	2	2	1.00000
28	021ESS	8	4	2	0.50000
29	021ESS	9	7	7	1.00000
30	021ESS	10	3	3	1.00000
31	022VIR	1	614	7	0.01140
32	022VIR	2	17	6	0.35294
33	022VIR	3	24	2	0.08333
34	022VIR	4	5	2	0.40000
35	022VIR	5	25	9	0.36000

Obs	Zone d'enquête	Strate HID	Nombre de répondants VQS	Effectif des individus HID tirés	Taux de sondage réels
36	022VIR	6	16	9	0.56250
37	022VIR	7	8	8	1.00000
38	022VIR	8	10	6	0.60000
39	022VIR	9	21	20	0.95238
40	022VIR	10	2	1	0.50000

On voit bien que les taux de sondage dans la strate 1 sont très faibles, ils sont de l'ordre de 1 %, alors que les taux sont relativement élevés dans les strates 7 à 10 (groupes VQS 5 et 6). Par exemple, le taux de sondage global de la zone d'enquête 011BOU vaut 0.08722, ceci aurait été la probabilité de tirage de chaque individu de la zone si on ne l'avait pas stratifiée, alors que les individus de 7 et 9 ont une probabilité de tirage supérieure à 50 % , et ceux de la strate 1 ont une probabilité inférieure à 1 % .

Pour mieux constater la répartition de l'échantillon HID entre les 10 strates, la table tirage.parstrate présente les taux globaux de tirage par strate :

Taux globaux de sondage par strate HID

Obs	strate HID	Nombre de répondants VQS	Effectif des individus HID tirés	Taux globaux de sondage
1		359010	21760	0.06061
2	1	269493	1747	0.00648
3	2	12632	2124	0.16814
4	3	19331	929	0.04806
5	4	4489	811	0.18066
6	5	15705	2651	0.16880
7	6	9453	2321	0.24553
8	7	6597	2788	0.42262
9	8	6905	1828	0.26474
10	9	10976	5433	0.49499
11	10	3429	1128	0.32896

Globalement, 0.6 % des individus de la strate 1 ont été tirés contre 49.5 % des individus de la strate 9 dans le but de sur représenter les individus les plus sévèrement handicapés, et aussi parce que l'effectif des répondants VQS des strates 7 à 10 était très faible.

Il est intéressant d'observer la dispersion des taux de sondage réels à l'intérieur d'une même strate HID. On se rappelle que bien que les taux de sondage attendus pour une même strate HID seraient différents d'une zone à l'autre, le calcul des effectifs HID par strate a été fait de manière à avoir des probabilités de tirage proportionnelles à une échelle fixe. Regardons donc la dispersion obtenue par strate HID :

Examen des dispersions des taux de sondage par strate HID

Obs	STRATE	MOYENNE	MINIMUM	MAXIMUM	ECTYPE	ECART
1		0.32565	0.000602	1.00000	0.27671	0.99940
2	1	0.00783	0.000602	0.04335	0.00502	0.04275
3	2	0.22224	0.018519	1.00000	0.13352	0.98148
4	3	0.06519	0.005435	0.25000	0.03589	0.24457
5	4	0.27009	0.023810	1.00000	0.16854	0.97619
6	5	0.22381	0.014286	1.00000	0.11864	0.98571
7	6	0.34267	0.024194	1.00000	0.17537	0.97581
8	7	0.57928	0.017241	1.00000	0.26966	0.98276
9	8	0.37551	0.014286	1.00000	0.20285	0.98571
10	9	0.67352	0.038961	1.00000	0.27870	0.96104
11	10	0.49141	0.026316	1.00000	0.25979	0.97368

On remarque que, sauf en ce qui concerne les strates 1 et 3, dans toutes les autres strates , le taux de sondage peut atteindre 100 %, pour un minimum de l'ordre de 2 %. La dispersion des taux est très élevée dans toutes les strates.

Parallèlement, on peut regarder la dispersion des pondérations dans les strates HID. La pondération des individus est défini par l'inverse de leur probabilité de tirage :

Examen de la dispersion des poids par strate HID

Obs	STRATE	MOYENNE	MINIMUM	MAXIMUM	ECTYPE	ECART
1		27.878	1.0000	1660.00	98.360	1659.00
2	1	212.193	23.0690	1660.00	235.427	1636.93
3	2	7.189	1.0000	54.00	7.582	53.00
4	3	24.019	4.0000	184.00	23.640	180.00
5	4	5.554	1.0000	42.00	4.593	41.00
6	5	7.107	1.0000	70.00	8.126	69.00
7	6	4.619	1.0000	41.33	5.138	40.33
8	7	2.775	1.0000	58.00	4.154	57.00
9	8	4.525	1.0000	70.00	6.352	69.00
10	9	2.221	1.0000	25.67	2.553	24.67
11	10	3.335	1.0000	38.00	4.173	37.00

Comme on pouvait s'y attendre, à l'inverse des probabilités de tirage, les individus de la strate 1 présentent un poids très grand, qui peut atteindre 1660 : un individu tiré représente en moyenne 212 répondants VQS de la strate. Les individus tirés dans la strate 2 représentent en moyenne 24 répondants VQS de la strate. Dans les autres strates, les pondérations sont faibles ; ainsi, la situation des individus de ces strates sera décrite avec beaucoup plus de précision.

Notons que la somme des poids des 21 760 individus HID est inférieur à 359 010 du fait du défaut de couverture (seules 336 des 366 zones d'enquête ont participé au tirage HID).

III. Les pondérations finales.

Il s'agit d'un tirage en deux phases avec post-stratification donc la probabilité pour un individu d'appartenir à l'échantillon final des 21 760 individus HID est égale à la probabilité globale de première phase multipliée par la probabilité de deuxième phase.

Cependant, le taux de déchets de l'enquête HID est encore plus élevé que le taux de non-réponse à l'enquête de filtrage : on enregistre 22.2 % de non-réponse à HID. Afin de corriger le biais et limiter la diminution de précision introduits par le phénomène de non-réponse et par le défaut de couverture, on décide de réaliser un redressement de l'échantillon en repondérant les répondants à HID en plusieurs étapes.

III.1. Traitement des échecs de collecte de HID.

On commence par caler le fichier des répondants HID sur l'échantillon de l'enquête HID. On a pris en compte, parmi les critères susceptibles d'influer sur les taux de refus et d'échec :

- l'âge des individus (décennal, en 9 modalités dont la dernière est 80 ans et plus) ;
- leur sexe ;
- la taille urbaine de leur lieu d'habitation, en 9 modalités ;
- le nombre de personnes du ménage, de 1 à 6 et plus ;
- l'état de handicap « probable », déduit de la variable synthétisant les réponses aux questions de VQS, en 10 modalités ;
- le type de logement (individuel, collectif, autre).

Ces six variables ont été relevées lors du recensement de mars 99 ou de son complément VQS. Leurs totaux pour les 21 760 individus de l'échantillon ont constitué les marges de calage. Il est important de traiter la non-réponse à l'aide de ces variables auxiliaires pour diminuer au maximum le biais. Prenons par exemple le cas de la strate HID : on peut imaginer que les individus ayant beaucoup de difficultés répondent bien car ils se sentent concernés et voudraient améliorer leur situation tandis que les individus n'ayant aucune difficulté à déclarer sont ceux qui répondent le moins bien, ce qui provoquerait l'introduction d'un biais positif, on sur-estimerait le nombre de personnes souffrant d'un handicap donné. On peut aussi imaginer le scénario inverse : si ce sont les individus les plus handicapés qui refusent de répondre (par exemple dans le cas de handicap mental), ceci provoquerait un biais négatif de l'estimateur, on sous-estimerait le nombre de personnes souffrant de ce type de handicap.

On affecte un poids initial qui vaut 1 pour chacun des 16 945 répondants et le logiciel CALMAR permet d'obtenir de nouveaux poids, de sorte que les totaux pour chacune des modalités de ces six variables coïncident avec les marges ce calage.

Voici les totaux observés pour la variable tranche d'âge et on vérifie que le total des coefficients des répondants HID pour les différentes modalités est conforme aux effectifs de l'échantillon :

TRAGED	ECHANTILLON	POIDS DES REPONDANTS
0-9 ans	955	955.00
10-19 ans	981	981.00
20-29 ans	1298	1298.00
30-39 ans	1763	1763.00
40-49 ans	2586	2586.00
50-59 ans	2911	2911.00
60-69 ans	3256	3256.00
70-79 ans	5396	5396.00
+ 80 ans	2614	2614.00
TOTAL	21760	21760.00

Une fois le coefficient de redressement pour la correction de la non-réponse HID déterminé, on multiplie ce coefficient par le poids HID (l'inverse de la probabilité de tirage HID), ce qui donne une nouvelle pondération HID.

III.2. Amélioration de l'échantillon.

On décide d'utiliser des informations auxiliaires afin d'améliorer la précision de l'estimation, en essayant de corriger les déformations de tirage. On prend en compte parmi les données dont on dispose celles qui sont susceptibles d'influer sur les incapacités ou handicaps des personnes et on réalise un redressement par calage sur les marges de ces variables.
Ce redressement est fait en trois temps :

- calage du fichier des répondants HID sur le fichier des répondants VQS ;
- calage du fichier des répondants HID sur les effectifs du RP99 ;
- calage du fichier des répondants HID sur la pyramide des âges du RP projetée à la date de l'enquête (décembre 1999).

III.2.1. Calage du fichier des répondants HID sur le fichier des répondants VQS.

On a tenu à ce que l'échantillon HID présente certaines particularités, par exemple les personnes les plus sévèrement et les plus certainement handicapées ont été sur-représentées, un plus gros échantillon a été tiré dans l'Hérault en vue d'estimations locales. Le redressement aura pour but de repondérer les répondants à HID de sorte que leurs effectifs pondérés leur donnent une structure identique à celle du fichier dans lequel a été tiré l'échantillon HID (fichier des répondants VQS).
On a tenu compte, parmi les critères disponibles dans le fichier des 359 010 individus ayant répondu de façon suffisamment complète à VQS pour participer au tirage de l'échantillon HID, de six variables :

- le plus important, car il déterminait une stratification du tirage selon une échelle de probabilités allant presque de 1 à 100, est la strate HID ou état de handicap probable, construit en synthétisant les réponses aux questions de VQS, en 10 modalités (cf. paragraphe II.2 du chapitre 1) ;
- l'âge des individus, décennal, en 9 modalités ;
- leur sexe ;

- la strate géographique en 31 modalités ;
- le nombre de personnes du ménage ;
- le type de logement en deux modalités (collectif, autre y compris individuel).

Bien qu'on ait tenu à sur-représenter les plus handicapés (groupes VQS 5 et 6), on s'arrange pour que le poids de chaque groupe VQS soit égal à l'effectif des répondants VQS du groupe, car sinon les estimations de prévalences risquent d'être fortement biaisées. Ainsi, la situation des plus handicapés sera bien décrite du fait du fort taux de personnes interrogées, sans pour autant fausser les résultats. De même, on s'intéresse à la possibilité de différences de prévalence entre les hommes et les femmes, entre les différentes tranches d'âge, les types de ménage, il est donc important que les effectifs pondérés des répondants correspondent aux effectifs de la population dans laquelle l'échantillon a été tiré (répondants VQS).

Pour la procédure de calage sur les marges de ces six variables, on utilise comme poids initial le nouveau poids HID obtenu à l'étape précédente.
Ce poids initial est rectifié de sorte que les totaux pondérés pour chacune des modalités des six variables auxiliaires coïncident avec les marges ce calage. La somme des poids des 16 945 répondants HID vaut 359 010.

On observera ci-après la répartition de l'échantillon des 359 010 personnes VQS selon la strate de tirage HID. Naturellement, la répartition pondérée des individus ayant répondu à HID, obtenue à cette étape du redressement, donne strictement les mêmes effectifs :

STRATE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	269493	75.07	269493	75.07
2	12632	3.52	282125	78.58
3	19331	5.38	301456	83.97
4	4489	1.25	305945	85.22
5	15705	4.37	321650	89.59
6	9453	2.63	331103	92.23
7	6597	1.84	337700	94.06
8	6905	1.92	344605	95.99
9	10976	3.06	355581	99.04
10	3429	0.96	359010	100.00
TOTAL	359010			

Une fois qu'on a corrigé les pondérations HID pour traiter la non-réponse HID et le défaut de couverture du tirage HID et pour améliorer les estimations, on obtient un poids global égal au produit du poids HID par le poids VQS calculé et corrigé dans la partie I.

III.2.2. Calage du fichier des répondants HID sur les effectifs globaux du RP.

On décide de réaliser le redressement des pondérations des répondants HID par calage sur les effectifs de la population métropolitaine suivant les deux variables que l'on a utilisées pour le traitement des non-réponses VQS :

- la strate géographique en 31 modalités, qui déterminait largement les probabilités de tirage des différents composants de l'échantillon VQS ;
- la tranche d'unité urbaine en 9 modalités.

En vue d'estimations locales issues de l'enquête VQS, 8 départements et une région avaient demandé une extension d'échantillon pour le tirage de VQS. Le redressement permettra donc de retrouver chez les répondants HID une répartition pondérée par strate géographique identique à la structure de la population métropolitaine.

Pour la procédure de calage sur les marges de ces deux variables, on utilise comme poids initial un poids égal au produit du poids HID par le poids VQS. On obtient alors un nouveau jeu de pondération dont la somme est égale à la population métropolitaine au RP99.

Comme on pouvait s'y attendre compte tenu de la stratification du tirage HID dans VQS et des faibles probabilités affectées aux individus du groupe 1 (personnes déclarant n'avoir aucune difficulté) âgées de moins de 70 ans, les pondérations sont particulièrement élevées dans la strate HID 01. On observe la dispersion des poids dans l'Hérault et dans le reste de l'échantillon, vu que les habitants de l'Hérault ont un poids plus faible.

Dispersion des poids hors Hérault

STRATE	_FREQ_	moyenne	minimum	maximum	ecartype
1	1541	35062.86	7255.60	127232.75	18195.61
2	1899	1518.94	333.81	5122.64	737.16
3	850	4622.90	964.44	18931.46	2399.52
4	734	1326.09	295.09	4620.04	608.85
5	2438	1368.95	257.14	5236.86	688.00
6	2096	939.12	184.71	3471.99	472.60
7	2552	496.76	93.40	2164.93	257.84
8	1648	970.14	199.04	3429.79	469.02
9	5017	426.11	81.17	1787.26	224.29
10	1022	720.85	155.23	2583.74	345.71

Dispersion des poids dans l'Hérault

STRATE	_FREQ_	moyenne	minimum	maximum	ecartype
1	206	4309.76	2399.95	7576.07	1266.19
2	225	181.77	109.98	319.14	50.43
3	79	588.94	304.04	995.29	155.51
4	77	170.43	99.19	287.83	40.55
5	213	171.59	87.94	305.75	50.40
6	225	116.17	67.73	202.88	30.14
7	236	65.04	32.31	105.25	16.70
8	180	116.08	66.90	194.14	28.76
9	416	55.31	28.08	101.97	13.42
10	106	84.75	52.18	142.87	20.61

La moyenne des coefficients dans la strate 01 dépasse légèrement 35 000 en dehors de l'Hérault (4 300 dans ce département), mais en outre la dispersion est assez forte, puisqu'on trouve une dizaine de valeurs supérieures à 100 000 et une dizaine de valeurs inférieures à 9200 (de 2 400 à 7 500 pour l'Hérault).

On a donc procédé à un « écrêtage » des valeurs extrêmes en resserrant l'éventail des pondérations dans la strate 01, séparément dans l'Hérault et dans le reste de l'échantillon. Celui-ci a eu pour principe de conserver la moyenne des poids (et donc le poids global) de la strate, et nous a amenés à prendre de nouveaux poids dont le rapport à cette moyenne soit la puissance 1 / 3.9 hors Hérault et 1 / 1.57 dans l'Hérault.

Dispersion des poids après écrêtage hors Hérault

STRATE	_FREQ_	moyenne	minimum	maximum	ecartype
1	1541	35062.86	23975.61	49971.80	4548.89
2	1899	1518.94	333.81	5122.64	737.16
3	850	4622.90	964.44	18931.46	2399.52
4	734	1326.09	295.09	4620.04	608.85
5	2438	1368.95	257.14	5236.86	688.00
6	2096	939.12	184.71	3471.99	472.60
7	2552	496.76	93.40	2164.93	257.84
8	1648	970.14	199.04	3429.79	469.02
9	5017	426.11	81.17	1787.26	224.29
10	1022	720.85	155.23	2583.74	345.71

Dispersion des poids après écrêtage dans l'Hérault

STRATE	_FREQ_	moyenne	minimum	maximum	ecartype
1	206	4309.76	2998.16	6235.10	813.505
2	225	181.77	109.98	319.14	50.426
3	79	588.94	304.04	995.29	155.514
4	77	170.43	99.19	287.83	40.551
5	213	171.59	87.94	305.75	50.398
6	225	116.17	67.73	202.88	30.137
7	236	65.04	32.31	105.25	16.699
8	180	116.08	66.90	194.14	28.757
9	416	55.31	28.08	101.97	13.418
10	106	84.75	52.18	142.87	20.612

On est ainsi parvenus à des éventails de 24 000 à 50 000 hors Hérault et de 3 000 à 6 000 dans l'Hérault.

Les répartitions pondérées par strate HID obtenues avant et après cette correction sont strictement les mêmes :

STRATE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	42497425	74.27	42497425	74.27
2	2183303	3.82	44680729	78.09
3	3083593	5.39	47764322	83.48
4	748174.6	1.31	48512496	84.78
5	2599368	4.54	51111865	89.33
6	1531750	2.68	52643615	92.00
7	1047127	1.83	53690742	93.83
8	1218534	2.13	54909276	95.96
9	1743743	3.05	56653019	99.01
10	565563.1	0.99	57218582	100.00
TOTAL	57218582			

III.2.3. Calage du fichier des répondants HID sur la pyramide des âges du RP projetée à la date de l'enquête.

Les données du RP se situant à la date du 8 mars 1999 alors que la date centrale de l'enquête intervenait 9 mois plus tard (le 1er décembre 99), on a décidé d'utiliser une projection de la pyramide des âges en décembre 1999 car l'âge est un élément central dans les travaux sur le handicap. Puisqu'on disposait de la date de naissance des répondants HID, on a décidé de réaliser un dernier calage sur la population RP projetée au 1/12/99, par sexe et âge quinquennal, en « âges atteints dans l'année ».

On utilise comme poids initial le poids obtenu à l'étape précédente.
On a noté une conséquence particulière : comme l'échantillon VQS a été tiré dans une population née avant le 8 mars 1999, différente de celle qu'on aurait eu si on avait tiré directement l'échantillon le 1er décembre 1999, les poids des individus de moins de cinq ans ont été un peu tirés vers le haut.
Le fait d'avoir utilisé les marges de la population projetée à la date de l'enquête a eu pour effet d'accroître la population globale d'environ 20 000 personnes, correspondant à l'accroissement de population vivant en domicile ordinaire entre le 8 mars et le 1er décembre 1999. On est ainsi passé à une estimation globale de 57 431 807 au lieu de 57 218 582.

IV. Conclusion.

Dans le but d'obtenir de meilleures estimations et de pallier les phénomènes de non-réponse et de défaut de couverture, nous n'utiliserons pas comme pondération finale le poids classique égal à l'inverse du produit de la probabilité de tirage de première phase par la probabilité de tirage de deuxième phase, mais un poids redressé par plusieurs critères déterminants et établi en plusieurs étapes.

Les objectifs de l'enquête nous ont amené à affecter des poids très différents aux individus HID : un répondant HID représente en moyenne 3 389 métropolitains vivant en ménage ordinaire, mais les pondérations vont de 28 à 50 000.
Ces grands écarts de pondérations finales viennent du fait que le tirage de HID a été stratifié avec des taux de sondage très différents et du fait qu'il y a eu une extension d'échantillon dans le département de l'Hérault.

Chapitre 3
Calcul d'une estimation de la variance en utilisant le logiciel POULPE

La description du plan de sondage et l'établissement des pondérations finales nous permettrons de mettre en œuvre des calculs d'estimation de la précision de l'enquête. Nous utiliserons pour cela le logiciel POULPE, conçu à cet effet, mis au point et utilisé par l'INSEE. Nous pourrons ainsi établir des estimations d'intervalles de confiance pour plusieurs variables d'intérêt.

I. Objectifs et principes du logiciel Poulpe.

Poulpe est un logiciel développé à partir du langage SAS MACRO et qui a pour but d'évaluer la précision des données recueillies sur des échantillons tirés de sondages complexes, à plusieurs degrés ou plusieurs phases. Il apporte une estimation de l'erreur due à l'échantillonnage, en estimant la variance de l'estimateur d'Horvitz-Thompson sur les variables d'intérêt, et fournit l'intervalle de confiance à 95% centré sur le total pondéré.

I.1. Quels plans de sondage le logiciel permet-il de traiter ?

Poulpe permet de traiter la plupart des plans de sondage d'enquêtes statistiques, à plusieurs degrés et en plusieurs phases.

Le logiciel couvre le domaine suivant :

- les sondages à 1 ou n degrés en une phase,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage stratifié,
- les sondages à 1 ou n degrés, en deux phases, lorsque la seconde phase est un sondage Poissonnien,
- les sondages à 1 ou n degrés, en trois phases, lorsque la deuxième phase est un sondage stratifié, et la troisième phase un sondage Poissonnien : ceci permet de traiter les enquêtes en deux phases dont la deuxième phase est stratifiée avec une correction de non réponse, en considérant cette dernière comme un sondage Poissonnien.

Type d'enquête	1ère phase	2ème phase	3ème phase
Enquêtes en une phase	1 ou plusieurs degrés (probabilités d'inclusion calculées par le logiciel ou fournies)		
Enquêtes en 2 phases	1 ou plusieurs degrés (probabilités d'inclusion calculées par le logiciel ou fournies)	stratifiée (probabilités d'inclusion calculées par le logiciel)	

Type d'enquête	1ère phase	2ème phase	3ème phase
Enquêtes en 2 phases	1 ou plusieurs degrés (probabilités d'inclusion calculées par le logiciel ou fournies)	Poissonnien (probabilités d'inclusion fournies)	
Enquêtes en 3 phases	1 ou plusieurs degrés (probabilités d'inclusion calculées par le logiciel ou fournies)	stratifiée (probabilités d'inclusion calculées par le logiciel)	Poissonnien (probabilités d'inclusion fournies)
Enquêtes en 1, 2 ou 3 phases : effet de sondage	Calcul de l'estimateur de variance pour un pseudo sondage aléatoire simple " équivalent " avec des poids égaux à ceux calculés par le logiciel		
Enquêtes en 1, 2 ou 3 phases : méthode Ultimate Clusters	Calcul de l'estimateur de variance sur un arbre particulier, à partir de poids fournis		

Pour les sondages élémentaires, les formules programmées actuellement portent sur les types suivants:

- sondage aléatoire simple sans remise (SAS),
- sondage à probabilités inégales (en particulier à probabilités proportionnelles à la taille (PPT));
- sondage systématique à probabilités égales (SYS); pour ces sondages, il faut indiquer au logiciel quel était l'ordre de tri des données avant le tirage,
- sondage stratifié (EXH),
- sondage équilibré (SASEQ).

Dans toutes ces configurations, le logiciel offre la possibilité de calculer les estimateurs d'Horvitz-Thompson (H-T), leur variance et l'effet de sondage (Design Effect) : c'est le rapport entre la variance obtenue par le plan de sondage et la variance obtenue par un pseudo sondage aléatoire simple avec des poids égaux à ceux calculés par le logiciel.

II.2. Comment Poulpe calcule -t-il les estimateurs ?

Poulpe passe par une décomposition fine du plan de sondage en phases et étapes (degrés de tirage et stratifications) établie par l'utilisateur, ce qui demande une parfaite connaissance du plan de sondage.

Il commence par calculer les probabilités d'inclusion élémentaires de la première phase, c'est-à-dire celles relatives à un sondage élémentaire à partir :

- du nombre d'entités tirées (ex : nombre de zones de délégués tirées);
- de la taille de l'unité dans laquelle on a tiré (ex : taille de la strate géographique) et de la taille de l'unité tirée (ex : taille de la zone de délégué);
- du type de sondage (ex : tirage PPT).

et en appliquant la formule correspondante.
Une fois toutes les probabilités élémentaires calculées, le logiciel fait le produit pas à pas de tous les degrés de tirage pour obtenir la probabilité globale de la première phase.

Pour les enquêtes en plusieurs phases dont la deuxième est stratifiée, le logiciel calcule la probabilité d'inclusion de la deuxième phase comme le rapport nh / NH, nh étant l'effectif

pour la strate h dans l'échantillon 2ème phase et NH l'effectif total pour la strate h de l'échantillon 1ère phase.
La probabilité d'inclusion pour un tirage poissonnien doit être fournie au logiciel.

La probabilité d'inclusion finale est le produit des probabilités d'inclusion relatives à chacune des phases.

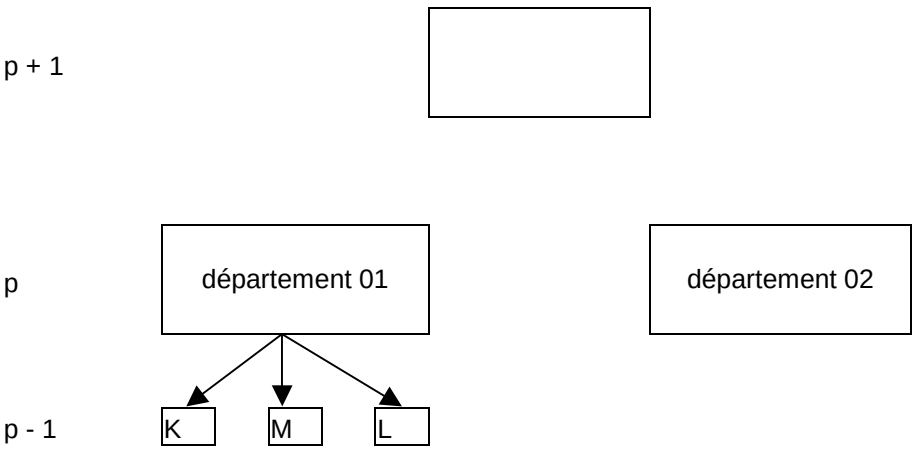
Une fois les probabilités d'inclusion finales calculées, Poulpe passe aux calculs d'estimateurs de statistiques simples (comme le total de la variable) ou complexes (comme les ratios ou encore les fonctions de plusieurs variables). Il fournit deux sortes d'estimations :

- les estimateurs de H-T établis à partir des poids calculés par le logiciel ;
- les estimateurs de H-T calculés à partir des poids fournis par l'utilisateur, comme c'est le cas pour HID.

L'estimateur de H-T du total de la variable d'intérêt Y vaut :

$\hat{Y} = \sum_s \frac{y_k}{\pi_k}$ où π_k est la probabilité d'inclusion finale de l'individu k et s est l'échantillon final.

Pour calculer une estimation de la variance, Poulpe fonctionne pas à pas : il part de la dernière étape de tirage , il calcule la variance en appliquant les formules d'estimation de variance correspondant au type de tirage, puis passe au niveau supérieur qui correspond à l'étape de tirage précédant celui qui vient d'être traité.
Pour calculer la variance au niveau supérieur, il part des estimateurs du niveau inférieur (celui qui vient d'être traité) et les agrège en utilisant **le principe de Raj**.
Représentons les différents de degrés de tirage par le schéma suivant :



Dans la département 01 on tire les communes K, L et M.

Poulpe calcule l'estimateur H-T du total de la variable d'intérêt Y sur chacune des trois communes ($\hat{y}_M$, $\hat{y}_L$ et $\hat{y}_K$) ainsi que l'estimateur de la variance de ces trois estimateurs (U_M , U_L et U_K) en appliquant la formule d'estimation de la variance propre au type de tirage des individus dans chacune des communes.

Pour estimer la variance au niveau supérieur p (donc ici pour calculer la variance au niveau du département 01), le principe de Raj consiste à utiliser les sommes pondérées et les variances calculées au niveau p-1 comme nouvelles 'variables d'intérêt'.

La variance au niveau p est la somme de deux termes :

- la variance de la 'variable d'intérêt' «somme pondérée » obtenue au niveau p-1 ;
- la somme pondérée des variances obtenues au niveau p-1.

Chaque unité du niveau p-1 (donc ici les communes) est pondérée par l'inverse de sa probabilité de tirage w_i.

La formule de Raj dit que l'estimateur de la variance au niveau supérieur p vaut :

$$\hat{V}(\hat{Y}) = f(\hat{y}) + \sum_{S_{p-1}} w_i U_i$$

où S_{p-1} est l'échantillon obtenue au niveau p-1 (donc ici il s'agit des trois communes)

et où f correspond à la formule d'estimation de la variance correspondant au type de tirage des unités du niveau p-1 (donc ici le tirage des communes).

Une fois la variance et l'estimateur H-T calculés pour toutes les unités du niveau p, on passe au niveau p+1 et on applique à nouveau le principe de Raj.

C'est ainsi que Poulpe aboutit à la variance finale. Le calcul est conduit de manière récursive, dans l'ordre chronologique inverse des opérations de tirage des entités : on commence par traiter les niveaux inférieurs (c'est à dire les dernières entités tirées) pour remonter ensuite aux niveaux supérieurs.

Finalement, Poulpe fournit des intervalles de confiance à 95% centrés sur le total pondéré de la statistique demandée. Le total pondéré qu'il utilise à cette étape est calculé à partir des poids fournis par l'utilisateur. Il calcule également sur option l'effet de sondage pour évaluer la pertinence du plan de sondage.

Poulpe permet de traiter les enquêtes qui ont été redressées par un logiciel comme CALMAR (comme c'est le cas pour HID) et l'estimation est faite sur les résidus.

Les formules utilisées par Poulpe pour estimer la variance à chaque niveau et pour calculer l'effet de sondage sont résumées en annexe 4. Une explication détaillée des estimations après un redressement est rédigée en annexe 2.

II. Application à HID.

L'utilisation du logiciel Poulpe demande un travail préparatoire assez conséquent, qui consiste en la mise au point de plusieurs fichiers de travail adaptés. L'application à HID a demandé des remaniements supplémentaires au fur et à mesure du lancement du logiciel.

La mise en œuvre de Poulpe pour estimer la précision de HID a exigé également le calcul des probabilités de réponse à HID. Pour estimer ces probabilités, nous utiliserons une régression logistique.

II.1. Fonctionnement général du logiciel.

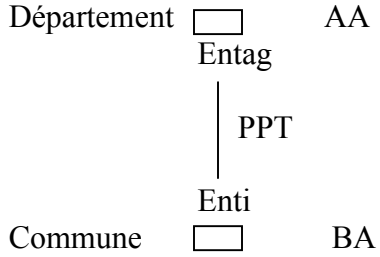
Les programmes sont écrits en langage Macro de Sas . Le logiciel s'appuie sur les informations issues de trois fichiers :

- le fichier arbre décrivant les différentes étapes du plan du sondage, appelé MODELE ;
- le fichier des données résultant de l'enquête, appelé DATA, il comprend l'ensemble des individus tirés à l'issu de la première phase ;
- le fichier géographique , appelé GEO, il contient, pour chacune des unités tirées à chaque étape élémentaire, les effectifs des unités géographiques dans lesquelles ont été tirées ces unités, ces effectifs étant nécessaires au calcul des probabilités d'inclusion.

Le modèle du sondage doit être construit avec soin, car c'est lui qui permettra à Poulpe d'appliquer les bonnes formules pour le calcul des probabilités d'inclusion et d'estimation de la variance. On modélise les étapes de tirage de la première phase sous la forme d'un arbre qui comprend plusieurs arcs. Chaque étape élémentaire est représentée par un arc, et on identifie les extrémités de l'arc par un code à deux lettres qui permettra de faire le lien entre les étapes élémentaires successives. Pour chaque arc, il faut renseigner :

- l'unité tirée : Enti ;
- l'unité dans laquelle on a tiré (unité d'agrégation) : Entag ;
- le type de tirage élémentaire : Typtir.

Par exemple, si on tire des communes dans des départements proportionnellement à leur taille, le tirage élémentaire est représenté par l'arc suivant :



Le dernier arc de l'arbre représente le tirage des individus, l'identifiant inférieur de cet arc est appelé le code feuille.

Une fois le plan de sondage entièrement modélisé sous la forme d'un arbre, on crée une table SAS pour décrire le modèle, chaque observation correspondant à un arc décrivant un tirage élémentaire, avec au minimum les données suivantes :

- NSUP : identifiant de l'extrémité supérieur de l'arc (ex : AA) ;
- NINF : identifiant de l'extrémité inférieure de l'arc (ex : AB) ;
- ENTI : unité tirée (ex : département commune) ;
- ENTAG : unité d'agrégation (ex : département) ;
- TYPTIR : type de tirage (ex : PPT) ;

D'autres données peuvent être nécessaires, comme le nom de la variable (T_AILLE) qui sert à désigner la taille dans le cas d'un tirage PPT.

Le fichier géographique apporte des informations auxiliaires sur les effectifs des différentes unités d'agrégation (dans le cas d'un sondage SAS, SASEQ, SYS) ou la taille des unités d'agrégation et des unité tirées (dans le cas d'un sondage PPT). Dans cet exemple, le fichier doit renseigner la taille des départements et des communes.
Ce fichier doit permettre un appariement avec les données de l'enquête, donc il doit contenir des identifiants géographiques communs.
On procède comme suit : un code est créé (variable Auxniv), il précise le niveau du tirage auquel correspond chaque unité présente dans le fichier géographique ; ainsi lors de l'appariement, on sait à quelle entité se rapporte le contenu de la variable effectif ou taille.

Dans notre exemple, on aurait :

Département	Commune	Auxniv	NNN	Tailuni
04		1	78	92086
04	125	2	1250	1250

Le département est de niveau 1, la commune est tirée dans le département donc la commune est de niveau 2.
La variable Tailuni est utilisée pour les tirages PPT car elle renseigne sur la taille en unités élémentaires de tirage, c'est-à-dire l'unité que l'on veut sonder, donc ici il s'agit des ménages.
La variable NNN est utilisée pour les autres types de tirage, elle donne le nombre d'unités de niveau inférieur à celui de l'enregistrement en cours.
Ainsi, Poulpe lit que le département 04 contient 78 communes (NNN) et 92086 ménages (Tailuni) et que la commune 04 125 contient 1250 ménages (Tailuni et NNN).

Le fichier des données contient l'ensemble des unités élémentaires (appelés individus) tirés à l'issu de la première phase du plan de sondage. Dans notre exemple il s'agirait de l'ensemble des ménages tirés. Il faut renseigner les identifiants géographiques de tirage pour que Poulpe puisse déterminer la taille de l'échantillon tiré à chaque étape élémentaire de tirage et établir les probabilités d'inclusion.
Par exemple, si on tire des communes dans des départements, il faut renseigner pour chaque ménage le numéro de département et le numéro de commune, afin que Poulpe puisse compter le nombre de communes tirées dans chaque département.
Il faut aussi préciser pour chaque individu le code feuille de l'arbre, qui n'est pas forcément le même d'un individu à l'autre, car il peut exister des plans très complexes (comme c'est le cas

pour la majorité des enquêtes de l'INSEE), où tous les individus ne sont pas tirés suivant la même procédure de tirage.

Pour les enquêtes en plusieurs phases, l'utilisateur doit préciser à quelle phase de l'enquête l'individu appartient en créant une variable PHASE qui vaut 1, 2 ou 3. Si la deuxième phase est stratifiée, il faut préciser les variables de stratification, ainsi, Poulpe calculera les effectifs NH et nh de deuxième phase.

Les variables d'intérêt, les variables de redressement et les variables élémentaires nécessaires au calcul de variables d'intérêt complexes (ex : ratios) sont à entrer dans ce fichier.

Le logiciel fonctionne en quatre étapes :

- un module vérifie la bonne construction du modèle et ordonne les arcs car la structure des formules oblige à traiter les arcs des feuilles à la racine (première étape de tirage).

- un module calcule les probabilités d'inclusion en faisant appel aux trois fichiers;

- un module permet la déclaration des variables d'intérêt ;

- le dernier module calcule les estimations de variance sur des statistiques simples ou sur des fonctions (ex : ratio).

Les deux premières étapes sont lancées une fois pour toutes pour une enquête donnée, seules les deux dernières sont à refaire si on déclare de nouvelles variables d'intérêt.

II.2. Construction des fichiers pour HID.

Un inconvénient majeur a été rencontré dans l'application de Poulpe à HID : le numéro des secteurs d'agents recenseurs, qui ont servi au tirage de l'échantillon VQS (1ère phase du plan de sondage), n'a pas été archivé dans les fichiers de l'enquête issus des différentes collectes. On ne peut donc pas affecter un individu à un secteur d'agent recenseur. Par contre, les fichiers de collecte renseignent sur le district dans lequel l'individu a été enquêté. On rappelle que le secteur d'AR est un regroupement de 3 districts en moyenne. Cependant, les données du recensement ne permettent pas d'établir la correspondance entre un district et son secteur d'AR, le numéro de secteur d'AR est une donnée que les services du recensement n'archivent jamais.

Afin de pallier l'absence d'information sur le secteur d'AR, on propose de « faire comme si » on avait tiré les districts RP99 directement dans les zones de délégués (ZD), ce qui provoquerait une sous-estimation de la variance. Néanmoins, alors que les secteurs d'AR sont de tailles à peu près constantes (530 personnes), les districts interrogés par VQS sont de tailles très variables : pour une moyenne de 184 personnes, la taille des districts varie entre 2 et 1190. Cette forte dispersion des tailles provoquerait une augmentation de la variance, donc finalement le fait de « faire comme si » on avait tiré des districts directement dans les ZD ne sous-estimera pas tant que cela la variance.

II.2.1. La modélisation du plan de sondage.

Le plan de sondage mis en œuvre par l'INSEE pour l'enquête HID est un plan de sondage en deux phases avec post-stratification:

- enquête de filtrage VQS ;
- enquête HID elle-même auprès d'un sous -échantillon des répondants à VQS.

L'arbre sert à la modélisation de la première phase du plan de sondage : il s'agit d'un tirage stratifié, à deux degrés et aréolaire.
Le territoire métropolitain a été découpé en 36 strates et dans chaque strate, des ZD ont été tirées proportionnellement à leur taille en 1990 (premier degré de tirage), puis dans chaque ZD des secteurs d'agents recenseurs ont été tirés par sondage aléatoire simple (deuxième degré de tirage). Enfin, toutes les personnes vivant en domicile ordinaire dans les secteurs d'AR tirés ont été enquêtées.
On va remplacer le secteur d'AR par le district dans le deuxième degré de tirage.

Poulpe considère la stratification géographique comme un premier degré de tirage de type exhaustif (EXH). Donc le tirage des ZD correspond à un deuxième degré de tirage et le tirage des districts correspond à un troisième degré de tirage.

La deuxième phase du plan de sondage (tirage de HID dans VQS) est menée auprès des seuls répondants à VQS. Pour traiter la non-réponse à VQS, on propose de rajouter un quatrième degré de tirage à la première phase, qui consiste à sélectionner les répondants VQS, parmi les individus interrogés dans chaque district, par sondage aléatoire simple. On aboutira donc à un échantillon de première phase de taille égale à 359 010.

Les variables d'intérêt sont renseignées uniquement pour les répondants à HID, comme on l'a déjà expliqué dans le chapitre 2, il est nécessaire de traiter la non-réponse à HID. On va donc rajouter une troisième phase au plan de sondage, qui consistera en un tirage poissonnien des répondants HID parmi les individus interrogés par HID. Il faut fournir au logiciel les probabilité d'inclusion pour un tirage de poisson. Ces probabilités sont estimées par régression logistique, les résultats seront présentés dans la partie II.3.

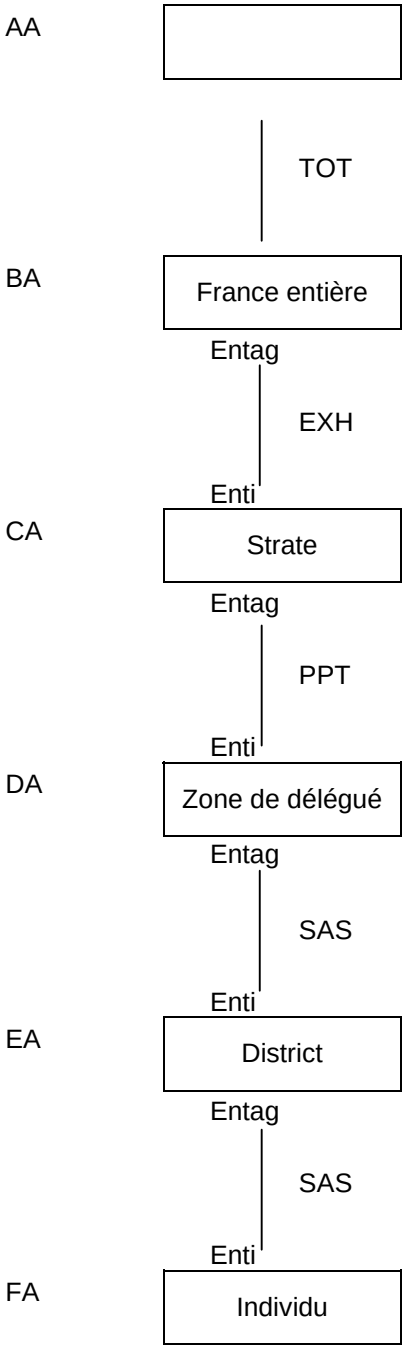
En résumé, on modélise le plan de sondage de l'enquête HID par un tirage en trois phases :

- 1ère phase avec quatre degrés de tirage :
  - premier degré : stratification géographique;
  - deuxième degré : tirage de ZD par tirage PPT ;
  - troisième degré : tirage de districts par SAS ;
  - quatrième degré : tirage de personnes par SAS ;

- 2ème phase stratifiée ;

- 3ème phase de poisson.

On modélise le tirage de la première phase par l'arbre de la page suivante.
Tout arbre est surmonté d'un arc spécial servant à consolider les résultats.



On saisit cet arbre sous la forme d'une table SAS comme le demande Poulpe, ce qui donne le fichier MODELE :

Obs	TYPTIR	ENTI	ENTAG	NSUP	NINF	T_AILLE
1	TOT			AA	BA	
2	EXH	STRATE	STRATE	BA	CA	
3	PPT	STRATE CODEDEL	STRATE	CA	DA	TAILUNI
4	SAS	STRATE CODEDEL DISTRI	STRATE CODEDEL	DA	EA	
5	SAS		STRATE CODEDEL DISTRI	EA	FA	

Codedel désigne la zone de délégué et Distri désigne le district au RP99.

II.2.2. Le fichier géographique.

Dans le fichier géographique, il est nécessaire de rentrer toutes les données auxiliaires dont Poulpe aura besoin pour le calcul des probabilités d'inclusion.
Il s'agit donc de renseigner :

1. la population 1990 de chacune des strates géographiques pour le tirage des ZD ;

2. la population 1990 de chacune des 391 ZD tirées pour VQS ;

3. le nombre total des districts au RP99 que comprend chaque ZD car la modélisation du plan de sondage dit que le tirage des districts dans les ZD a été effectué par sondage aléatoire simple ;

4. la population 1999 de chacun des districts afin de traiter la non-réponse VQS par un sondage aléatoire simple des répondants VQS parmi les individus interrogés (on rappelle que tous les habitants du district ont été interrogés par VQS).

La détermination de la population 1990 des strates et des ZD a déjà été présentée lors du calcul des pondérations VQS (partie I du chapitre 2).
Ces deux populations servent au tirage des ZD, de type PPT donc elles seront renseignées dans la variable TAILUNI.

On dispose d'un fichier du RP99 qui donne pour chaque ZD la liste de ses districts ainsi que la population 1999 de chaque district.
A partir du fichier de l'enquête, on récupère pour chaque ZD la liste des districts interrogés par VQS en utilisant la fonction logique FIRST / LAST de SAS. On en comptabilise 2273 au total .
Et à partir du fichier du RP, on obtient le nombre total de districts pour chaque ZD ainsi que la population 1999 de chaque district interrogé.

Etant donné que l'on ne peut pas estimer la variance à partir d'un échantillon de taille égale à 1, on regarde s'il existe des ZD dans lesquelles un seul district a été interrogé par VQS.
Effectivement, on compte 69 ZD dans lesquelles un seul district a été interrogé par VQS. Ceci n'est pas étonnant, car dans le paragraphe I.2 du chapitre 2, nous avons expliqué que dans les ZD appartenant à une strate géographique sans aucune extension, le but était de tirer en moyenne 1.6 secteurs d'AR sur 30.8 , donc les procédures d'arrondi ont amené les concepteurs de l'enquête à tirer tantôt 1 seul secteur d'AR, tantôt 2 secteurs d'AR. Par ailleurs, les villages et autres très petites communes sont en général mono-district et un secteur d'AR est en général inclus dans une commune. Donc dans ces cas-là, le secteur d'AR contient un seul district et si un seul secteur a été tiré dans la ZD, seul un district a été interrogé par VQS.

On propose donc de regrouper ces ZD par deux ou trois à l'intérieur d'une même strate géographique. On additionne la population 1990 des ZD ainsi que le nombre total de districts.
On aboutit ainsi à 352 zones de délégués au lieu des 391 interrogées par VQS.

Il se trouve également qu' une seule ZD a été enquêtée par VQS dans la strate 36 (région Corse), on propose donc de rattacher cette ZD à la strate 34 (région Provence-Alpes-Côte-d'Azur). Il faut donc additionner la population 1990 des strates 34 et 36.
On aboutit donc à 34 strates au lieu des 35 interrogées par VQS. On rappelle que la strate 13 n'a été interrogée que par l'enquête EHF (Etude de l'Histoire Familiale).

La variable AUXNIV vaudra 1 pour les strates car les strates sont au premier niveau de tirage. Elle vaudra 2 pour les ZD car celles-ci sont au niveau de tirage égal à 2, et les districts sont au niveau 3.

Le programme SAS permettant de construire le fichier géographique est présenté en annexe 3 : c:\user\odile\fichier\geo2.sas .

Le fichier GEO contient 2651 observations :
- 34 strates géographique ;
- 352 zones de délégués ;
- 2273 districts.

Le fichier GEO a la forme suivante : on ne présentera que les deux premières strates, deux ZD par strate et deux districts par ZD.

Obs	STRATE	CODEDEL	DISTRI		AUXNIV	NNN	TAILUNI
1	1				1	.	7285186
2	1	DR523022			2	28	11207
3	1	DR523022	751090910027		3	377	377
4	1	DR523022	751090910040		3	25	25
5	1	DR523035			2	57	22566
6	1	DR523035	751111140042C		3	545	545
7	1	DR523035	751131330039B		3	435	435
8	2				1	.	1320836
9	2	DR603027			2	76	17986
10	2	DR603027	08105	DN03A	3	247	247
11	2	DR603027	08105	DN03B	3	389	389
12	2	DR603039			2	34	13232
13	2	DR603039	51173	0001	3	113	113
14	2	DR603039	51597	0001	3	339	339

Ainsi Poulpe utilisera la variable Tailuni pour le tirage des ZD car il s'agit d'un tirage PPT.
Il lira à Auxniv = 1 que la strate 1 a une taille de 7 285 186 individus et à Auxniv = 2 que la ZD 'DR523022' a une taille de 11 207 individus. Il ira dans le fichier des données compter le nombre m de ZD tirées dans la strate 1 et pourra ainsi calculer la probabilité d'inclusion P_z de la ZD 'DR523022' :

$$P_z = m \times \frac{11207}{7285186}$$

Il n'est pas besoin de renseigner le nombre total de ZD que comprend la strate (variable NNN) car Poulpe n'a pas besoin de cette information.

Pour calculer la probabilité de tirage du district ‘ 751090910027 ‘ dans la ZD ‘DR523022’, comme il s’agit d’un tirage SAS, Poulpe utilise la variable NNN. Il lit à Auxniv = 2 que la ZD ‘DR523022’ comprend au total 28 districts. Il va compter dans le fichier des données le nombre n de districts interrogés par VQS pour la ZD ‘DR523022’.
Et on obtiendra la probabilité P_d de tirage du district ‘ 751090910027 ‘dans la ZD ‘DR523022’ :

$$P_d = \frac{n}{28}$$

Cette probabilité sera la même pour tous les districts de la ZD ‘DR523022’.

Pour calculer la probabilité de tirage d’un individu dans le district ‘ 751090910027 ‘, Poulpe utilise la variable NNN et lit à Auxniv = 3 que le district ‘ 751090910027 ‘ a un effectif de 377 individus. Il compte dans le fichier de l’enquête que k individus ont répondu à VQS dans le district ‘ 751090910027 ‘ donc la probabilité P_i de tirage des individus de ce district vaut :

$$P_i = \frac{k}{377}$$

Poulpe calculera une probabilité globale de première phase qui sera égale au produit des probabilités de tirage à chaque degré de tirage.

II.2.3. Le fichier des données.

Le fichier des données contiendra l’ensemble des individus de la première phase, donc ici il s’agit des 359 010 répondants VQS. Ce fichier doit obligatoirement contenir les répondants VQS et non l’échantillon final car Poulpe a besoin de l’échantillon de première phase pour calculer les effectifs nh (effectif de la strate h dans l’échantillon de 2ème phase) et NH (effectif de la strate h dans l’échantillon de 1ère phase) afin d’établir les probabilités de d’inclusion relatives à la deuxième phase.

Pour chacun des 359 010 répondant VQS, il faut renseigner :

1. les identifiants géographiques pour permettre l’appariement avec le fichier géographique et le calcul des effectifs de première phase:
   - la strate géographique : STRATE ;
   - la zone de délégué : CODEDEL ;
   - le district : DISTRI ;

2. les variables de stratification pour la deuxième phase :
   - la zone d’enquête en 366 modalités: ZONDENQ ;
   - la strate HID en dix modalités : STRATEH ;

3. la phase de l'enquête à laquelle l'individu appartient : PHASE qui vaudra:
   - 1 si l'individu n'appartient pas à l'échantillon HID (il ne participe qu'à la première phase du plan de sondage);
   - 2 si l'individu appartient à l'échantillon HID mais n'a pas répondu à HID (il ne participe qu'aux deux premières phases du plan de sondage) ;
   - 3 si l'individu a répondu à HID (il participe aux trois phases du plan de sondage) ;

4. le code feuille de l'arbre qui vaut 'FA' pour tous les individus : NINFFIC ;

5. la probabilité de réponse à HID pour le tirage poissonnien de la troisième phase: PROBAREP ;

6. le poids final de l'enquête calculé dans le chapitre 2 : POIDSCOR.

Les zones d'enquête sont définies par commune (une commune est incluse dans une unique zone d'enquête). On dispose du fichier des répondants VQS, ce fichier renseigne le code département et le code commune pour chaque répondant VQS, on définit la zone d'enquête comme on l'a fait lors du calcul des pondérations HID au chapitre 2.
Le fichier de l'enquête contient déjà le district, la zone de délégué et la strate géographique. On pense bien-sûr à apporter les modifications dues aux regroupements de ZD (dans le cas où un seul district avait été interrogé par VQS dans la ZD) et à l'introduction de la strate 36 dans la strate 34.

Le fichier dont on dispose contient déjà également la définition de la strate HID.
Pour définir la variable PHASE, on dispose en plus du fichier de l'enquête VQS, un fichier de l'enquête HID qui contient les 21 760 individus tirés pour HID. Ce fichier contient une variable INTERV qui vaut 1 si l'individu a répondu à HID et 0 sinon. De plus, un identifiant (IDVQSDEF) allant de 1 à 359 010 est commun aux deux fichiers.
Nous allons donc utiliser cet identifiant pour faire l'appariement entre les deux fichiers afin de définir les trois échantillons des phases 1, 2 et 3.

Après avoir trié les deux fichiers par la variable IDVQSDEF, on utilise la fonction MERGE BY de SAS avec la fonction logique IN pour isoler les individus n'ayant pas été tirés pour HID et pour ces individus, on affecte la valeur 1 à la variable PHASE.
A partir du fichier de l'enquête HID, on affecte la valeur 2 à la variable PHASE si la variable INTERV vaut 0 et la valeur 3 si INTERV = 1.
On effectue un dernier appariement avec la fonction MERGE BY pour recopier dans le fichier de l'enquête VQS, la valeur de la variable PHASE pour les 21 760 individus ayant été tirés pour HID.

Ainsi, la variable PHASE présente la distribution suivante :

PHASE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	337250	93.94	337250	93.94
2	4815	1.34	342065	95.28
3	16945	4.72	359010	100.00

On a bien: - 16945 répondants HID;
- 16945 + 4815 = 21760 individus de l'échantillon HID,
- 21760 + 337250 = 359 010 répondants VQS.

Poulpe demande que la probabilité d'inclusion pour un tirage poissonien soit fourni par l'utilisateur. On enregistre un taux global de déchet pour HID de 22.2%, soit un taux global de réponse de 77.8 % .
On pourrait utiliser cette probabilité de réponse pour HID et considérer qu'elle est la même pour tous les individus interrogés par HID. Ceci ne serait pas réaliste car on se doute bien que plusieurs facteurs peuvent influer sur la probabilité de réponse de chacun et que cette probabilité ne peut être constante. Par exemple, on peut imaginer que les personnes acceptent de répondre à l'interview selon qu'ils aient ou non des difficultés à déclarer et que la probabilité de réponse différent selon la strate HID. On encore que la probabilité de réponse dépend de l'âge ou du type de commune.

Parmi les éléments dont on dispose dans le fichier de l'enquête HID, on considère ceux qui sont susceptibles d'influer sur la probabilité de réponse à HID :

- la taille urbaine du lieu d'habitation en 9 modalités ;
- l'âge des individus , décennal en 9 modalités ;
- leur sexe ;
- le nombre de personnes du ménage ;
- la strate HID en dix modalités ;
- le type de logement (individuel, collectif ou autre).

Et on suppose que les probabilités de réponse sont homogènes dans les différentes classes, les classes sont obtenues par croisement des modalités de ces différentes variables. Pour chaque classe, on estime donc la probabilité de réponse à HID par le taux de réponse dans la classe.
On réalise pour cela une régression logistique sur les six variables. Les résultats de cette régression seront présentés dans le paragraphe suivant.

Poulpe demande également que figurent dans le fichier des données les variables de calage qui ont été utilisées dans les différentes étapes du redressement.
Il faut rentrer toutes les variables d'intérêt ainsi que les variables élémentaires qui serviront à construire les statistiques complexes.
Les variables d'intérêt brutes sont en général des variables caractères, on crée de nouvelles variables d'intérêt qui seront numériques et dichotomiques 0-1. On décide de recoder les variables selon l'information recherchée.
Prenons le cas de la variable BMOB1 qui demande si la personne est confinée au lit (= 1), dans un fauteuil (= 2) ou à l'intérieur de sa maison (= 3) en raison d'un problème de santé. Comme on s'intéresse au nombre de personnes confinées pour des raisons de santé, on crée une variable d'intérêt CONFIN1 qui vaut 1 si BMOD1 vaut 1, 2 ou 3, et 0 sinon. Ainsi l'estimateur de Horvitz - Thompson du total de la variable CONFIN1 estimera le nombre de métropolitains confinés en raison d'un handicap.
Les différentes variables sur lesquelles portera le calcul d'une estimation de la variance seront présentées dans la partie III de ce même chapitre.

Le programme SAS permettant de construire le fichier des données est présenté en annexe 3 : c:\user\odile\fichier\donnees.sas .

II.3. Calcul des probabilités de réponse à HID par régression logistique.

On dispose de la variable INTERV qu'on utilisera comme variable réponse à HID, elle prend les valeurs 0 et 1.
On réalise la régression logistique sur les variables suivantes :

- la taille urbaine du lieu d'habitation en 9 modalités : TUU;
- l'âge des individus , décennal en 9 modalités : TRAGED;
- leur sexe : SEXE;
- le nombre de personnes du ménage : TAILMEN;
- la strate HID en dix modalités : STRATE2 ;
- le type de logement (individuel, collectif ou autre) : TYPLOG.

Après avoir transformé les variables caractères en variables numériques, on utilise la procédure LOGISTIC de SAS. On commence par réaliser la régression sur le modèle complet.

a) Modèle complet

The LOGISTIC Procedure

Model Information

Data Set	WORK.TABLE
Response Variable	INTERV
Number of Response Levels	2
Number of Observations	21760
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	INTERV	Total Frequency
1	0	4815
2	1	16945

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	5613	6316.9056	1.1254	<.0001
Pearson	5613	5907.9968	1.0526	0.0030

Number of unique profiles: 5620

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	23003.199	22646.087
SC	23011.187	22702.001
-2 Log L	23001.199	22632.087

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	369.1122	6	<.0001
Score	361.5789	6	<.0001
Wald	355.2742	6	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.3027	0.1129	133.1387	<.0001
tuu	1	0.0535	0.00604	78.4402	<.0001
TRAGED	1	0.0178	0.00882	4.0841	0.0433
SEXE	1	-0.0277	0.0335	0.6836	0.4083
tailmen	1	-0.1539	0.0158	94.4709	<.0001
strate2	1	-0.0283	0.00592	22.8150	<.0001
typlog	1	0.1698	0.0332	26.2288	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
tuu	1.055	1.043	1.067
TRAGED	1.018	1.001	1.036
SEXE	0.973	0.911	1.039
tailmen	0.857	0.831	0.884
strate2	0.972	0.961	0.983
typlog	1.185	1.110	1.265

Association of Predicted Probabilities and Observed Responses

Percent Concordant	58.6	Somers' D	0.183
Percent Discordant	40.3	Gamma	0.185
Percent Tied	1.0	Tau-a	0.063
Pairs	81590175	c	0.591

Le modèle est convergent, d'une part les coefficients de déviance montrent que le modèle est adéquat car la p-value correspondante est très nettement inférieure au seuil critique 0.05.
D'autre part, le test global rejette l'hypothèse nulle, selon laquelle tous les coefficients de régression seraient nuls : en effet la statistique du likehood ratio ainsi que la statistique de Wald et celle de Score ont des p-value inférieures à 0.05 .
Par ailleurs, les quatre indices qui mesurent l'association entre la probabilité prédite et la valeur de la variable INTERV sont bons : l'indice C (qui est compris entre 0 et 1) est supérieur à 0.5 , le D de Somers, le Gamma et le Tau-a de Kendall (qui sont compris entre -1 et 1) sont tous les trois positifs. Ce qui signifie que la prévision correspond globalement à la réalité.

Cependant, l'analyse des estimateurs du maximum de vraisemblance des différents coefficients montre que le coefficient relatif à la variable SEXE n'est pas significatif : en

effet, le χ^2 associé a une p-value de 0.4083, ce qui relativement supérieur au seuil critique de 5% .
De plus, le test de l'odd-ratio ne rejette pas l'hypothèque selon laquelle le odd-ratio relatif à la variable SEXE est égale à 1. Le odd-ratio est le rapport entre la probabilité de répondre à HID pour les hommes et la probabilité de répondre à HID pour les femmes, les modalités des autres variables étant constantes. Le test ne rejette donc pas l'hypothèse selon laquelle la probabilité de répondre à HID est identique pour les hommes et les femmes.
Nous allons donc regarder le modèle sans la variable SEXE.

b) Modèle sans la variable SEXE .

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	3761	4309.6875	1.1459	<.0001
Pearson	3761	4060.5756	1.0797	0.0004

Number of unique profiles: 3767

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	23003.199	22644.770
SC	23011.187	22692.697
-2 Log L	23001.199	22632.770

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	368.4287	5	<.0001
Score	361.0480	5	<.0001
Wald	354.7793	5	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.3435	0.1016	174.9252	<.0001
tuu	1	0.0533	0.00603	78.0860	<.0001
TRAGED	1	0.0173	0.00879	3.8597	0.0495
tailmen	1	-0.1530	0.0158	93.8657	<.0001
strate2	1	-0.0280	0.00592	22.4716	<.0001
typlog	1	0.1685	0.0331	25.8859	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
tuu	1.055	1.042	1.067
TRAGED	1.017	1.000	1.035
tailmen	0.858	0.832	0.885
strate2	0.972	0.961	0.984
typlog	1.183	1.109	1.263

Association of Predicted Probabilities and Observed Responses

Percent Concordant	58.6	Somers' D	0.183
Percent Discordant	40.3	Gamma	0.185
Percent Tied	1.1	Tau-a	0.063
Pairs	81590175	c	0.592

Ce deuxième modèle est également adéquat, les coefficients de toutes les variables sont significatifs.
Les critères de AIC et de Schwarz montrent que ce modèle est meilleur que le précédent : ces deux critèes sont plus faibles pour ce modèle que pour le modèle précédent.

On remarque tout de même que le coefficient relatif à la variable TRAGED (la tranche d'âge) est significatif certes, mais à peine : la p-value vaut 0.0495, ce qui est limite à la valeur seuil de 0.05 .De plus, l'intervalle de confiance à 95% de l'odd-ratio relatif à la variable TRAGED vaut : [1.000, 1.035], cet intervalle n'exclue donc pas la valeur 1.

Nous allons donc regarder le modèle sans les variables SEXE et TRAGED.

c) Modèle sans les variables SEXE et TRAGED.

Ce modèle est également adéquat, tous les coefficients sont significatifs. On va comparer les critères de AIC et de Schwarz.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	23003.199	22646.642
SC	23011.187	22686.581
-2 Log L	23001.199	22636.642

Les critères montrent ce modèle est moins bon que les deux premiers.

Le meilleur modèle logistique est donc le modèle sans le sexe. C'est ce modèle que nous utiliserons pour calculer la table des probabilités.

d) Estimation des probabilités de réponse à HID.

La procédure logistique estime la probabilité pour la première modalité de la variable INTERV, donc la procédure estime la probabilité P' de ne pas répondre à HID. On obtient une estimation de la probabilité P de répondre à HID par: P = 1 - P'.

En raison du grand nombre de classes, on ne peut présenter la table des probabilités.
La probabilité P est nommé PROBAREP.
Le taux de réponse à HID vaut 0.778 , on peut regarder la distribution des probabilités autour de cette valeur.

Une proc means sur la variable PROBAREP donne que pour une moyenne de 0.778 et un écart-type de 0.05, la variable varie entre 0.605 et 0.908 .

On utilise la procédure CHART de SAS pour observer la distribution de la probabilité:

```
proba                                                      Cum.              Cum.
                                                    Freq   Freq  Percent  Percent

                ,
0.75 à 0.78     ,******************                 3849   3849    17.69    17.69
                ,
0.78 à 0.85     ,****************************************  9264  13113    42.57    60.26
                ,
0.85 au MAX     ,*********                          1942  15055     8.92    69.19
                ,
MIN à 0.75      ,******************************     6705  21760    30.81   100.00
                ,
                Šffffˆffffˆffffˆffffˆffffˆffffˆffffˆffffˆffffˆfff
                     5    10   15   20   25   30   35   40

                                Percentage
```

On remarque que plus de 50% des individus ont une probabilité estimée supérieure à la probabilité globale de 0.778 dont 9% (soit 1942) avec une probabilité supérieure à 0.85 .
Seuls 17% des individus ont une probabilité proche de 0.778 (entre 0.75 et 0.78).

Le programme SAS permettant le calcul des probabilités de réponse à HID par régression logistique est présenté en annexe 3 : c:\user\odile\fichier\probarepHID.sas .

II.4. Déroulement de l'application de Poulpe à HID.

Il est nécessaire de résumer le déroulement du logiciel afin d'expliquer les remaniements ultérieurs qui ont été apportés aux fichiers construits précédemment.

Le lancement du logiciel s'effectue en quatre étapes, une étape étant lancée en faisant appel à un module, les étapes étant lancées dans l'ordre suivant :

1. vérification de la bonne construction de l'arbre et ordonnancement des arcs : étape ARBGEN ;

2. calcul des probabilités d'inclusion : étape CALPII ;

3. chargement d'une liste de variables d'intérêt : étape CHARLIS ;

4. estimation de totaux et de variances sur mes variables d'intérêt: étape ESTIVAR ou estimation sur statistiques complexes : étape ESTIFON.

On commence par lancer **l'étape ARBGEN**, aucun problème n'est à déclarer, les macros correspondantes sont exécutées sans erreur.

Par contre, le lancement de l'étape suivante, **CALPII**, met en évidence plusieurs difficultés.

Tout d'abord, au niveau des strates de deuxième phase. Le module demande que l'on rentre le nom des variables permettant de construire les strates de deuxième phase : ZONDENQ et STRATEH. Il y a 366 zones d'enquête et 10 strates HID. Le module construit donc les strates de $2^{ème}$ phase : il en compte 366 * 10 = 3660.
Un message d'erreur apparaît, expliquant que le nombre de strates de $2^{ème}$ phase est limité à 99 ! C'est une découverte pour les utilisateurs car le guide d'utilisation du logiciel ne mentionne pas cette contrainte. C'est pourquoi on sera obligé de faire abstraction de la zone d'enquête dans le calcul des probabilités de deuxième phase. On ne peut même pas utiliser la strate géographique en 31 modalités car cela ferait au total 31 * 10 = 310 strates de deuxième phase. Cette contrainte nous obligerait à utiliser juste la strate HID pour la stratification de deuxième phase, alors que le découpage en strates HID a eu lieu à l'intérieur de chaque zone d'enquête et que le tirage s'est effectué dans chaque croisement ZONDENQ*STRATEH.
Et on sera obligé d'affecter une probabilité de tirage identique pour les individus d'une même strate HID.

On propose de distinguer le département de l'Hérault des autres départements car on rappelle que ce département a bénéficié d'une extension d'échantillon pour l'enquête HID afin d'obtenir des estimations locales. Les taux de sondage des différentes strates HID pour ce département sont largement supérieurs à ceux des autres départements. Donc afin de garder cette particularité, on crée une nouvelle variable HERAULT qui vaut 'oui' si l'individu appartient au département de l'Hérault et 'non' sinon. Ainsi, on utilisera les variables HERAULT et STRATEH comme variables de stratification de la deuxième phase, et on obtiendra 2*10 = 20 strates de deuxième phase.

Par contre, comme on l'a expliqué au paragraphe II.2.2 du chapitre 1, on rappelle que toutes les ZD tirées pour VQS n'ont pas participé au tirage de l'échantillon HID. Il ne faut donc pas que les individus relevant de ces ZD soient pris en compte dans le calcul des effectifs NH, sinon les probabilités d'inclusion de la deuxième phase s'en trouveraient faussées. On décide donc d'enlever ces individus du fichier DATA. Il reste 330 191 individus de première phase au lieu des 359 010 de départ.

On relance l'étape CALPII, d'autres erreurs sont mises en évidence. Le module signale la présence d'individus pour lesquels il ne peut calculer la probabilité d'inclusion au

quatrième degré de tirage (tirage des individus dans les district). En allant regarder dans la table des probabilités qu'il a créée, on constate que quatre districts affichent une population 99 nulle, c'est à dire que l'effectif total du district (NNN) est nul ! On vérifie les valeurs dans le fichier du RP99 et cela ne vient pas d'une erreur d'appariement : le fichier de l'enquête VQS contient des individus qui ont été enquêtés dans des districts officiellement non habités. Cela vient sûrement d'une erreur de codification effectuée au moment de la collecte. 167 répondants VQS relèvent de ces quatre districts, parmi lesquels 11 ont été tirés pour HID et un seul a répondu à HID. N'ayant pas les moyens de corriger ces erreurs (de frappe ?), on propose d'ôter les 167 individus du fichier des données.

Par ailleurs, le module signale également des probabilités supérieures à 1 pour la même étape de tirage. En vérifiant, on se rend compte que 83 districts affichent une population 99 (NNN) inférieure au nombre de répondants VQS du district (n) ! Une autre erreur de codification. Pour pallier ce problème, on propose de corriger l'effectif NNN, en le portant égal à n. On supposera donc que tous les habitants de ces districts ont répondu à VQS.

On relance à nouveau l'étape CALPII, cette fois le logiciel a détecté des échantillons de taille = 1 pour cette même étape de tirage : il signale 8 districts pour lesquels il compte un seul répondant VQS. Quant il existe des échantillons de taille = 1, les probabilités d'inclusion sont calculées mais les calculs de variance seront erronés ou incomplets vu que l'on ne peut pas estimer la variance à partir d'un échantillon de taille = 1.
On propose donc de mettre l'individu concerné dans un autre district de sa zone de délégué, et d'additionner les populations 99 des deux districts.

L'étape CALPII est à nouveau lancée et les différentes macros de calcul sont exécutées sans message d'erreur ni avertissement.

L'**étape CHARLIS** ne signale aucun problème et une liste de 12 variables d'intérêt est chargée.

L'**étape ESTIVAR**, la plus longue de toutes, renvoie en sortie les estimations de H-T de totaux pour les 12 variables d'intérêt, mais pas leurs variances. Il signale qu'il ne peut calculer les variances car une macro sas a détecté des erreurs graves qui ont stoppé son exécution. Une observation attentive de la fenêtre LOG de SAS nous conduit à soupçonner un problème interne à la macro elle-même, et non un problème relatif à la construction des fichiers. Nous soupçonnons un problème sur la taille du fichier des données.
Nous décidons donc de réduire le fichier des données aux trois 1ères strates géographiques, on obtient ainsi un fichier de taille 40 318, sur lequel on désire faire un essai. On relance donc les quatre premières étapes et effectivement l'étape estivar est conduite jusqu'au bout et fournit les intervalles de confiance pour les totaux des 12 variables d'intérêt.

Nous avons pu récupérer auprès de l'Unité de Méthodologie Statistique de l'INSEE les sources des différentes macros faisant tourner le logiciel, afin d'étudier l'écriture de la macro chargée de calculer les variances. On constate alors que dans cette macro, la macro variable prenant pour valeur la taille du fichier des données a été définie en numérique sur seulement 5 positions, ce qui limite la taille du fichier des données à 99 999 observations! La macro ne peut donc pas tourner sur un fichier de 330 000 individus. Cette limitation sur la taille du fichier des données n'est pas non plus mentionnée dans le guide d'utilisation du logiciel.

Afin de pallier cette difficulté majeure, nous décidons de sectionner le fichier des données en quatre fichiers de taille à peu près constante, d'environ 80 000 individus, en prenant soin de ne pas couper les strates géographiques. On constitue ainsi quatre groupes.
Comme le tirage de l'échantillon VQS et de l'échantillon HID est indépendant d'une strate géographique à l'autre, le tirage est indépendant d'un groupe à l'autre.
Nous calculerons les estimations de variance séparément. La théorie des sondage nous permet d'additionner les quatre variances obtenues uniquement s'il s'agit d'estimation de totaux :

- l'estimateur H-T du total sur le fichier complet est la somme des estimateurs H-T du total calculés séparément dans les quatre groupes ;
- l'estimateur de la variance de l'estimateur H-T du total sur le fichier complet est la somme des estimateurs de la variance calculée séparément dans les quatre groupes.

Ce problème majeur limite donc notre étude à des estimations de variance sur les totaux des variables d'intérêt.

La procédure Freq de Sas nous permet d'avoir le nombre d'individus par strate géographique, on peut donc sectionner le fichier des données en quatre groupes d'environ 80 000 individus :

Groupe	Strates	Nb de strates	Effectif
1	1 à 8	8	84205
2	9 à 18	9	78147
3	19 à 27	9	79506
4	28 à 35	8	88164
		Total	330022

Le département de l'Hérault constitue la strate 32 donc se trouve dans le groupe 4.

II.5. Résumé des difficultés rencontrées lors de l'application de Poulpe à HID et des solutions qui y ont été apportées.

Le premier inconvénient à rappeler, est l'absence du fichier de l'enquête du numéro de secteur d'agent recenseur. On n'a pas pu affecter aux individus leur numéro de secteur d'agent recenseur. On a donc décidé de modifier le plan de sondage et de faire comme si le district avait été tiré directement dans la zone de délégué.

Plusieurs autres difficultés ont été rencontrées, tantôt liées à la théorie des sondages, tantôt liées à la spécificité du logiciel.

La détection d'échantillons de taille égale à 1 nous a amené à effectuer des regroupements :

- de districts quand on comptait un seul répondant VQS dans le district ;
- de ZD quand un seul district de la ZD avait été interrogé par VQS ;
- de strates : introduction de la strate 36 (Corse) dans la strate 34 (région PACA).

Des erreurs de codification des districts, effectuées sûrement au moment de la collecte VQS, ont conduit à l'élimination de 167 répondants VQS.

En ce qui concerne les problèmes liés à la spécificité du logiciel même, nous listons deux difficultés :

1. La limitation à 99 du nombre de phases de deuxième phase ne permet pas d'utiliser le découpage des répondants VQS en zones d'enquête avant la stratification par la strate HID, car cela fait au total 3660 strates de deuxième phase. On a créé une nouvelle variable HERAULT à deux modalités qui désigne l'appartenance de l'individu au département de l'Hérault. On utilisera les variables HERAULT et STRATEH (strate HID) comme variables de stratification de la deuxième phase. Ainsi, on distinguera le département de l'Hérault du reste de l'échantillon. On a également été amené à ôter du fichier des données les zones de délégué tirées pour VQS, mais n'ayant pas participé au tirage HID.

2. La limitation à 99 999 de la taille du fichier des données nous oblige à sectionner le fichier des données en quatre fichiers d'environ 80 000 individus chacun. Le fait de devoir travailler les fichiers séparément limite notre étude à des estimations de variance de totaux.

Au total, au lieu des 359 010 répondants VQS que nous avions au départ, le fichier des données ne comptera que 330 022 individus. Ce fichier présente la distribution suivante, selon la variable PHASE qui distingue les échantillons de :
- 1ère phase (répondants VQS n'ayant pas été tirés pour HID) ;
- 2ème phase (répondants VQS tirés pour HID mais n'ayant pas répondu à HID) ;
- 3ème phase (répondants HID).

PHASE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	308273	93.41	308273	93.41
2	4806	1.46	313079	94.87
3	16943	5.13	330022	100.00

En conclusion, nous sommes obligés de passer par beaucoup de simplifications et de remaniements afin d'estimer la précision de l'enquête HID. Tous ces remaniements vont conduire à des probabilités de tirage différentes des vrais probabilités de tirage. Cependant, Poulpe offre la possibilité d'avoir deux estimations du total de la variable d'intérêt, une estimation calculée en fonction des pondérations finales établies par le logiciel, et une estimation établie en fonction des pondérations finales imposées par l'utilisateur. Ainsi, en introduisant les pondérations finales calculées au chapitre précédent, nous aurons les véritables estimateurs des différents totaux. De plus, les intervalles de confiance que calcule Poulpe sont centrés sur ces derniers totaux.
La modélisation du plan de sondage permettra juste d'estimer la variance des estimateurs de totaux, nous obtiendrons donc des intervalles de confiance approximatifs.

Il est cependant très regrettable de devoir limiter l'étude à des estimations de totaux car le but de l'enquête HID est d'étudier les prévalences, à partir d'estimation de ratios. Par exemple, regarder s'il existe des différences entre les hommes les femmes, entre les différentes couches sociales. On peut toujours estimer la proportion d'hommes et la proportion de femmes souffrant d'un handicap donné, mais seule la confrontation des intervalles de confiance de ces proportions nous permettra de conclure à une différence significative entre les deux sexes.

Nous avons décidé de prendre contact avec l'informaticien de l'INSEE qui a conçu le logiciel Poulpe dès son retour de congés d'été (fin août), afin qu'il puisse apporter au logiciel les adaptations nécessaires à son utilisation pour HID. On espère au minimum que la prochaine version ne limitera pas la taille du fichier des données. On pourra ainsi poursuivre notre étude par des estimations de variance sur des ratios et des statistiques complexes. En attendant, nous présenterons dans la partie suivante l'analyse des résultats obtenus pour les estimations de totaux.

Il est important de signaler une difficulté d'une autre nature. Le lancement du logiciel est très coûteux en temps et en espace disque. Notre fichier de travail est très gros, il comporte beaucoup de variables, et le logiciel en génère beaucoup d'autres.
La première étape est réalisée une fois pour toute. La seconde étape (Calpii) qui consiste à calculer les probabilités de tirage ne devrait pas avoir besoin d'être réitérée pour traiter de nouvelles variables. En réalité, l'application effectue les calculs d'estimation de variance à partir du fichier de sortie de l'étape Calpii. Et à chaque fois que de nouvelles variables d'intérêt sont ajoutées au fichier des données, l'application exige que l'étape calpii soit relancée.
Or, l'enquête HID a permis de produire plus de 500 variables d'intérêt brutes, de plus, de très nombreuses variables seront créées pour estimer la précision de statistiques complexes, au fur et à mesure des besoins. Doit-on alors lancer Calpii sur un fichier des données qui comprendra toutes les variables ? En plus du temps d'exécution qui risque d'être phénoménal, on risque d'avoir un problème d'espace disque. Ou alors faudra t-il relancer Calpii à chaque nouvelle série de variables ?
L'objectif de cette application à HID est de pouvoir livrer aux équipes de recherche les trois fichiers de travail, plus le fichier de sortie Calpii. Les variables d'intérêt sont fournies sur cdrom sous forme de tables Sas, chaque équipe de recherche rajoutera au fichier des données les variables sur lesquelles porteront ses études.
Nous suggérons de lancer l'étape Calpii sans aucune variable d'intérêt, puis à rajouter au fichier de sortie les variables d'intérêt au fur et à mesure des besoins. Cette solution a l'avantage de permettre de travailler avec un fichier de taille raisonnable et de gagner beaucoup en temps d'exécution.

III. Analyse des résultats d'estimation de la variance.

Comme nous l'avons expliqué dans la partie précédente, notre étude se limitera pour l'instant à des estimations de variance sur des totaux. Cette première étude nous permettra d'avoir une idée sur la précision de l'enquête.

L'analyse de l'effet de sondage nous permettra en outre d'apprécier la pertinence du plan de sondage.

Les dix variables sur lesquelles portera notre étude sont les variables suivantes. Ce sont des variables brutes, elles sont en mode caractère. On les recode en variables numériques dichotomiques, et ce sont les 12 variables dichotomiques que l'on passera au logiciel comme variables d'intérêt. Ce sont les variables soulignées.

- BMOB1 dans MODB_C: indique si la personne est confinée au lit, dans un fauteuil, ou à l'intérieur de son logement en raison d'un handicap ou d'un problème de santé. On recodera cette variable. -> confin1 = 1 si confinée (BMOB1 = 1,2 ou 3).

- DADAPT dans MODD: si elle dispose d'équipement spécialement adaptés à son handicap. -> dadapt1 = 1 si oui (DADAPT = 1).

- C_AIDKI dans MINDIV_C: si la personne bénéficie d'une aide régulière pour les tâches quotidiennes, en raison de problèmes de santé ou d'un handicap. -> aidki1 = 1 si oui.

- R-ALLOC dans MINDIV_C: si elle perçoit une allocation ou pension en raison de son handicap. -> alloc1 = 1 si oui.

- R_INVAL dans MINDIV_C: si reconnaissance officielle d'un taux d'invalidité ou d'incapacité. -> inval1 = 1 si oui.

- AHANDI dans MINDIV_C: si difficultés physiques, sensorielles ou intellectuelles. -> handi1 = 1 si oui.

- BCOLVEZ dans MINDIV_C: indicateur de mobilité.
  -> mob1 = 1 si confinée au lit ou au fauteuil non roulant (BCOLVEZ = 1);
  -> mob2 = 1 si besoin d'aide pour toilette ou pour l'habillage (BCOLVEZ = 2);
  -> mob3 = 1 si besoin d'aide pour sortir (BCOLVEZ = 3).

- NBDEFIC dans MINDIV_C: indicateur du nombre de déficiences.
  -> defi1 = 1 si au moins une déficience (NBDEFIC >= 1).

- RCOTOR dans MODR_C: si elle a déposé un dossier à la COTOREP.
  -> cotor1 = 1 si oui.

- SLIRE, SECRIR, SCOMPT dans MODS_C: indique si la personne agée de plus de six ans sait lire, écrire, compter.
  -> expr1 = 1 si elle ne sait pas lire, ou écrire, ou compter (SLIRE ou SECRIR ou SCOMPT = 3).

Les estimations sont menées séparément sur quatre fichiers, et pour obtenir le résultat final, il suffira d'additionner les résultats obtenus sur les quatre fichiers.

III.1. Examen de la sortie Poulpe pour le groupe 1.

Le groupe 1 contient les répondants VQS des strates 1 à 8, ce fichier contient 84 205 individus.

III.1.1. Résultats.

Poulpe fournit deux estimateurs de Horvitz-Thompson du total de la variable d'intérêt :

- SOMLOG est l'estimateur déterminé à partir des poids finaux calculés par le logiciel sur la base des probabilités d'inclusion;
- SOMPOND est l'estimateur déterminé à partir des poids finaux fournis par l'utilisateur.

On doit s'attendre à ce que SOMPOND soit différent de SOMLOG car les simplifications et les remaniements que nous avons apportés au plan de sondage lors sa modélisation pour Poulpe ont conduit à des probabilités de tirage différentes des véritables probabilités de tirage. Donc les pondérations finales calculées par le logiciel sont différentes des pondérations finales de l'enquête, établies au chapitre précédent.

Poulpe calcule l'estimateur de la variance de l'estimateur du total (SOMLOG) et établit l' intervalle de confiance à 95% centré sur le résultat SOMPOND.
Il arrive que le logiciel ne puisse fournir une estimation approchée de la variance de l'estimateur de H-T lorsque cette variance est faible, c'est-à-dire lorsque la précision est bonne. Le logiciel signale cette situation par un message, et dans le cas HID sur le groupe 1, nous avons reçu le message suivant :

Attention : pour certaines variables très bien
estimées par le plan de sondage, la variance
de l estimateur est proche de zéro, et les
résultats fournis par le logiciel ne sont
plus valables.

PARAMETR VALEUR

EST075 Pour ces variables, ignorez les
variances calculées par le logiciel AIDKI1 DEFI1

Le logiciel fournit les résultats de la page suivante pour les douze variables d'intérêt. Pour les variables AIDKI1 et DEFI1, on ignorera les résultats de variance.

NOM	TOTAUX	VARIANCE	SOMLOG	SOMPOND	ECARTYP	BORNINF	BORNSUP
CONFIN1	256795.30	571619758.86	256795.30	167728.30	23908.57	120867.50	214589.10
DEFI1	10294226.00	-9499866255.77	10294226.00	6315092.00	.	.	.
COTOR1	1225060.00	11126627765.27	1225060.00	488698.60	105482.83	281952.25	695444.95
EXPR1	941775.20	26562195913.09	941775.20	513288.30	162979.13	193849.21	832727.39
DADAPT1	293127.10	1167520505.39	293127.10	212196.90	34169.00	145225.66	279168.14
AIDKI1	2402335.00	-4415952199.46	2402335.00	1229980.00	.	.	.
ALLOC1	1213653.00	8190237408.03	1213653.00	593769.60	90499.93	416389.74	771149.46
INVAL1	1678408.00	16494285606.08	1678408.00	799727.10	128430.08	548004.14	1051450.06
HANDI1	8747144.00	74506934995.53	8747144.00	5270798.00	272959.58	4735797.21	5805798.79
MOB1	33408.28	29301437.97	33408.28	19950.47	5413.08	9340.83	30560.11
MOB2	310396.10	300535790.75	310396.10	187443.00	17335.97	153464.50	221421.50
MOB3	380791.30	1242856681.76	380791.30	206411.60	35254.17	137313.42	275509.78

Nous notons d'amblée la différence entre les estimateurs SOMPOND et SOMLOG : l'estimateur H-T du total établi à patir des poids calculés par le logiciel est en moyenne 1.8 fois supérieur à l'estimateur H-T du total établi à partir des pondérations de l'enquête.
Les différents remaniements apportés au plan de sondage et à l'échantillon a donc globalement augmenté les poids. On sait en outre que l'estimateur H-T SOMPOND est meilleur car non seulement il a été établi à partir des vrais poids, mais en outre les poids ont été redressés de façon à limiter au maximum un éventuel biais introduit par le défaut de couverture.

Nous analyserons les résultats de variance au paragraphe III.3 .
Le logiciel signale que sur le groupe 1, les variables AIDKI1 et DEFI1 sont très bien estimées par le plan de sondage.
Nous devions logiquement nous attendre à un tel résultat, en particulier pour la variable DEFI1. En effet, cette variable indique si la personne interrogée par HID a au moins une difficulté à signaler : elle vaut 1 si oui et 0 si la personne n'a aucune difficulté à signaler. Comme le tirage de l'échantillon HID a été stratifié par la strate HID en dix modalités de handicap croissant (voir le paragraphe II.1 du chapitre 1), on doit s'attendre à ce que la variable DEFI1 vaille :

- 0 sur les strates 1 et 2 (les individus pour lesquels aucune difficulté n'a été déclarée à l'enquête VQS) ;
- 1 sur les 8 autres strates.

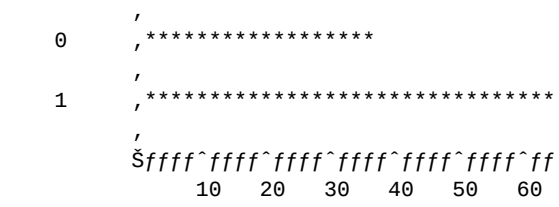
Donc la variance intra-strate de cette variable devrait être faible, ce qui conduit au final à une variance faible.

On peut regarder si la variable DEFI1 est très peu dispersée à l'intérieur des strates HID :

```
---------------------------------------- STRATEH =' 1' ------------------------------------

         defi1                                                Cum.              Cum.
                                                      Freq    Freq   Percent   Percent

                ,
           0    ,**********************************    939     939     70.02     70.02
                ,
           1    ,***************                      402    1341     29.98    100.00
                ,
                Šffff^ffff^ffff^ffff^ffff^ffff^ffff^
                     10   20   30   40   50   60   70

                           Percentage
```

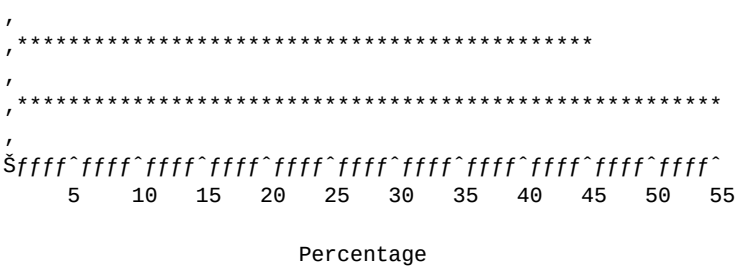
---------- STRATEH=' 2' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	563	563	35.77	35.77
1	1011	1574	64.23	100.00



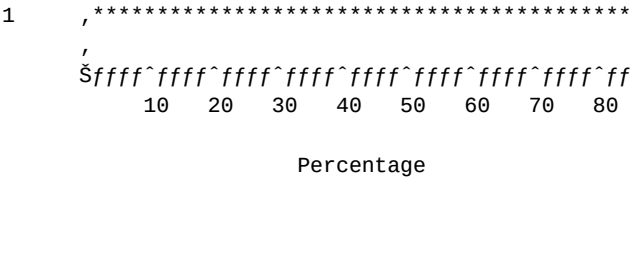
---------- STRATEH=' 3' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	324	324	44.75	44.75
1	400	724	55.25	100.00



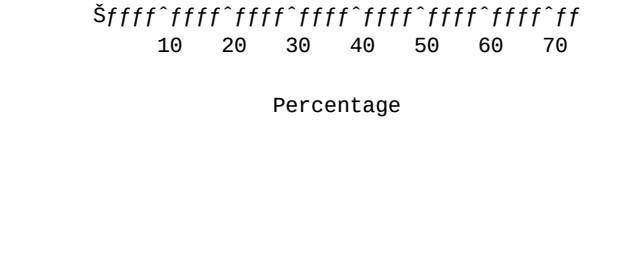
---------- STRATEH=' 4' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	94	94	15.49	15.49
1	513	607	84.51	100.00



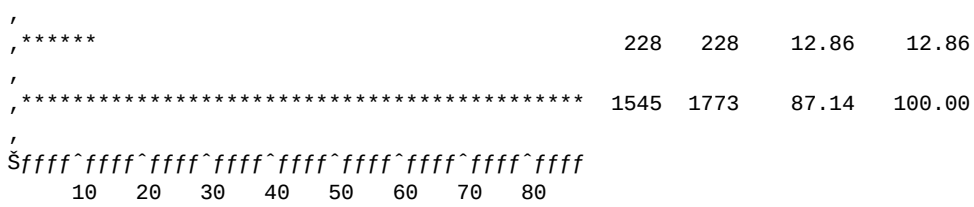
---------- STRATEH=' 5' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	539	539	26.60	26.60
1	1487	2026	73.40	100.00



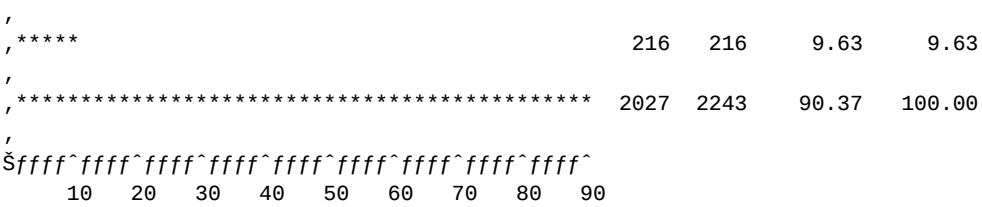
---------- STRATEH=' 6' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	228	228	12.86	12.86
1	1545	1773	87.14	100.00



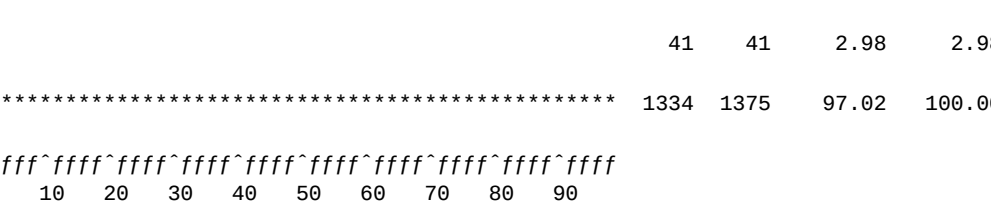
---------- STRATEH=' 7' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	216	216	9.63	9.63
1	2027	2243	90.37	100.00



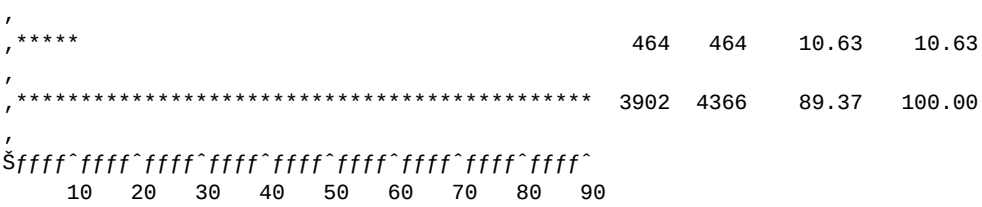
---------- STRATEH=' 8' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	41	41	2.98	2.98
1	1334	1375	97.02	100.00



---------- STRATEH=' 9' ----------

defi1	Freq	Cum. Freq	Percent	Cum. Percent
0	464	464	10.63	10.63
1	3902	4366	89.37	100.00



```
---------------------------------------- STRATEH=10 ----------------------------------------

    defi1                                                                 Cum.              Cum.
                                                                  Freq    Freq   Percent   Percent

          ,
      0   ,**                                                       40      40      4.67      4.67
          ,
      1   ,*********************************************           816     856     95.33    100.00
          ,
          Šffff^ffff^ffff^ffff^ffff^ffff^ffff^ffff^ffff^fff
               10   20   30   40   50   60   70   80   90

                              Percentage
```

L'observation de ces différents histogrammes montre qu'en fait, la variable DEFI1 n'est constante sur aucune des dix strates HID, ce qui est assez curieux. Sur les strates 6 à 10, les réponses sont conformes aux résultats de l'enquête VQS à 90%, puisque 90% des personnes souffrent d'au moins d'une déficience dans chacune des ces strates. Cependant, il n'était pas prévisible de trouver (même à très faible pourcentage) des personnes ayant déclaré des difficultés au moment de l'enquête VQS, qui déclarent au contraire à l'enquête HID n'avoir aucune déficience.
Par contre, il n'est pas étonnant de trouver dans les strates 1 et 2 des individus souffrant d'au moins une déficience, alors qu'ils n'avaient déclaré aucune difficulté à l'enquête VQS. Ce phénomène est particulièrement fort chez les personnes âgées (la strate 2) qui sont 40% à déclarer une déficience.

Ceci met déjà en évidence une conséquence de la différence de collecte entre les deux enquêtes.
Pour l'enquête de filtrage (VQS), un court questionnaire a été remis aux personnes au moment du recensement, et l'agent recenseur passait les récupérer ultérieurement. Les questions posées étaient en général peu développées, il fallait répondre par oui ou par non. Les personnes remplissaient elles-mêmes les questionnaires. Cette enquête avait pour but de compter les gens ayant une difficulté physique, sensorielle, intellectuelle ou mentale.
Alors que pour la véritable enquête (HID), un enquêteur se rendait au domicile des personnes afin de leur poser directement les questions ; ces questions étaient très précises, très ciblées, le questionnaire était beaucoup plus long. Les interviewés avaient donc moins de difficultés à répondre aux questions. Le fait de poser des questions très développées ont permis de mettre en évidence des déficiences chez des personnes qui spontanément répondent qu'elles n'ont aucune difficulté à déclarer.

L'observation des résultats du groupe 1 montre déjà que certaines variables sont très bien estimées par le plan de sondage. Il existe néanmoins une variance, qui est très faible et que Poulpe ne fournit pas.

III.1.2. Effets de sondage.

Le logiciel fournit sur option la variance que l'on aurait obtenue si l'échantillon HID avait été tiré directement dans la population totale par un pseudo sondage aléatoire simple,

c'est-à-dire que les individus sont tirés au hasard, mais ils gardent le même poids que celui de l'enquête. On peut ainsi discuter de la pertinence du plan de sondage.
Poulpe calcule donc le design effect, qui est le rapport de la variance obtenue par le plan de sondage, sur la variance que l'on obtiendrait dans le cas d'un sondage aléatoire simple.
On obtient les résultats suivants, on ne tiendra évidemment pas compte des valeurs affichées pour les variables AIDKI1 et DEFI1 :

CONFIN1	DADAPT1	AIDKI1	ALLOC1	INVAL1	HANDI1
0.38361	0.69224	-0.37100	1.18275	1.79111	2.60971
MOB1	MOB2	MOB3	DEFI1	COTOR1	EXPR1
0.14628	0.16569	0.56785	-0.30861	1.64187	5.08565

On remarque que, en ce qui concerne le groupe 1, sept variables sur douze affichent un effet de sondage inférieur à 1, ce qui signifie que le plan de sondage mis au point pour HID permet d'avoir des résultats plus précis que le sondage aléatoire simple. Les variables concernées sont des variables qui rentrent directement dans la définition du handicap des personnes : confin1, dadapt1, aidki1, defi1, mob1, mob2, mob3.
La variable EXPR1 (qui indique si la personne ne sait pas lire, écrire ou compter) présente par contre un effet de sondage égal à 5, ce qui signifie que cette variable serait beaucoup mieux estimée par un sondage aléatoire simple. Ce n'est pas étonnant dans la mesure où l'illettrisme et le handicap sont en général deux concepts séparés.

Cette première analyse sur le groupe 1 nous a permis de noter certains aspects du plan de sondage, mais c'est l'analyse globale que nous privilégierons.

III.2. Résultats des autres groupes.

L'analyse des résultats du groupe 1 a suggéré que certaines variables étaient bien estimées par le plan de sondage. Regardons ce qu'il en ait pour les trois autres groupes.

Pour chacun des trois autres groupes, Poulpe a signalé des variables très bien estimées par le plan de sondage.

Groupe 2 (strates 9 à 18) : CONFIN1, DADAPT1, COTOR1, EXPR1 et MOB3.
Autres variables ayant un effet de sondage inférieur à 1 : Mob1 et Mob2.

Groupe 3 (strates 19 à 27) : AIDKI1, INVAL1, MOB2, MOB3 et DEFI1.
Autres variables ayant un effet de sondage inférieur à 1 : Confin1, Dadapt1, Alloc1, Mob1 et Mob2.

Groupe 4 (strates 28 à 35) : DADAPT1, AIDKI1, ALLOC1 et MOB2.
Autres variables ayant un effet de sondage inférieur à 1 : Mob1 et Cotor1.

On constate que d'un groupe à l'autre, ce ne sont pas toujours les mêmes variables qui ont une variance très faible, cependant on retrouve globalement les sept variables du groupe 1 qui sont mieux estimées par le plan de sondage que par un sondage aléatoire simple.

Mais ce sont les résultats sur le fichier tout entier qui nous permettra de conclure qu'en a la précision de l'enquête HID.

III.3. Estimation de la précision de l'enquête HID.

Comme il s'agit d'estimation de totaux, l'estimateur de Horvitz-Thompson du total de la variable d'intérêt est la somme des estimateurs H-T du total calculés sur les quatre fichiers. Et l'estimateur de la variance de l'estimateur du total est la somme des quatre estimateurs de variance.
Pour les variables pour lesquelles Poulpe a jugé que la précision était très bonne sur un groupe, nous considérons que la variance estimée est nulle dans ce groupe.

Pour chaque variable d'intérêt, nous analyserons trois résultats de précision :

1 : la précision obtenue si on n'améliore pas l'échantillon par un redressement ; donc les calculs sont menés sans tenir compte des variables de redressement.

2 : la précision obtenue sur les données corrigées des fluctuations d'échantillonnage ; on va donc préciser au logiciel les variables de calage.

3 : la précision obtenue sur les données corrigées, en considérant que le plan de sondage est celui d'un sondage aléatoire simple, afin d'examiner la pertinence du plan de sondage HID.

III.3.1. Analyse des résultats.

L'échantillon final HID sur lequel nous estimons le total étant de taille 16 943, on est dans les conditions d'application du théorème central limite. Par conséquent, on peut supposer que l'estimateur du total suit une loi de Gauss, ce qui nous permet d'établir un intervalle de confiance à 95% du total. Soit $\hat{Y}$ l'estimateur H-T du total et $\sigma(\hat{Y})$ son écart-type. Alors il y a 95% de chance pour que le vrai total Y appartienne à l'intervalle :

$$IC = [\ \hat{Y} - 1.96 * \sigma(\hat{Y})\ ,\ \hat{Y} + 1.96 * \sigma(\hat{Y})\ ]$$

On estime la variance de $\hat{Y}$ par $\hat{V}(\hat{Y})$, donc l'intervalle de confiance approché à 95% de Y est :

$$\hat{I} = [\ \hat{Y} - 1.96 * \sqrt{\hat{V}(\hat{Y})}\ ,\ \hat{Y} + 1.96 * \sqrt{\hat{V}(\hat{Y})}\ ]$$

Pour chacune des douze variables d'intérêt, on calcule l'estimateur H-T du total obtenu à partir des pondérations finales de l'enquête, l'estimateur de la variance, ainsi que l'intervalle de confiance à 95% estimé.

Un moyen commode pour comparer le degré de précision des différentes variables consiste à comparer leurs coefficients de variation, qui est le rapport de l'écart type de l'estimateur à la valeur de l'estimateur, sans unité. Le résultat est exprimé en % .

Nous obtenons les résultats suivants :

Variable d'intéret	Estimateur du total (poids réels)	Variance de l'estimateur de HT du total	écart type de l'estimateur HT du total	Borne inférieure (IC à 95 %)	Borne supérieure (IC à 95 %)	coefficient de variation (en %)
AIDKI1	5017660.00	23268269755.18	152539.40	4718682.77	5316637.23	3.0401
ALLOC1	2237528.40	30645469026.39	175058.47	1894413.79	2580643.01	7.8237
CONFIN1	581965.80	3863599766.99	62157.86	460136.39	703795.21	10.6807
COTOR1	2105969.10	19990923768.76	141389.26	1828846.14	2383092.06	6.7137
DADAPT1	851930.00	1837940274.48	42871.21	767902.44	935957.56	5.0322
DEFI1	22226953.00	219976186129.84	469016.19	21307681.27	23146224.73	2.1101
EXPR1	1370555.70	36366676550.13	190700.49	996782.74	1744328.66	13.9141
HANDI1	17976079.00	318425419990.21	564291.96	16870066.75	19082091.25	3.1391
INVAL1	3483114.10	84341974284.11	290416.90	2913896.98	4052331.22	8.3379
MOB1	119182.66	227261953.58	15075.21	89635.25	148730.07	12.6488
MOB2	982243.20	1771495405.55	42089.14	899748.49	1064737.91	4.2850
MOB3	741583.20	1242856681.76	35254.17	672485.02	810681.38	4.7539

Ces estimations de totaux nous permettent de chiffrer les handicaps, l'incapacité et la dépendance.
Nous estimons à 22 millions le nombre de métropolitains vivant en domiciles ordinaires souffrant d'au moins une déficience physique, sensorielle ou mentale ou intellectuelle, soit un 1 métropolitain sur 2.5 .
Environ 3.5 millions de handicapés sont officiellement reconnus, c'est à dire qu'un taux d'invalidité a été attribué à leur état, mais seuls 2 millions de personnes perçoivent une allocation de handicap.
En ce qui concerne la dépendance, nous estimons à 5 millions le nombre de métropolitains vivant en domicile ordinaire et ayant besoin de l'aide régulière ou permanente d'une tierce personne pour accomplir les tâches de la vie quotidienne en raison d'un problème de santé ou d'un handicap : faire les courses, faire le ménage, s'alimenter, se doucher, se déplacer par exemple. Un peu moins d'un million a besoin d'aide pour la toilette et l'habillage mais parmi

les personnes qui n'ont aucune difficulté pour faire leur toilette ou pour s'habiller, 740 000 a besoin d'aide pour sortir.

Les variances estimées sont particulièrement élevées. L'enquête n'est pas très précise, puisque l'échantillon final comporte 16945 individus pour une population de 58 millions de personnes, soit un taux de sondage d'environ 1 / 3500.

En comparant les coefficients de variation, on constate que les variables les plus précises sont : Defi1, Aidki1, Handi1, Mob2, Mob3 et Dadapt1. Ces variables présentent un coefficient de variation compris entre 2 et 5 %. Les variables Defi1 et Handi1 entrent directement dans la définition des dix strates HID, c'est pourquoi il paraît tout à fait normal qu'elles soient plus précises que les autres variables.
Les variables les moins précises sont Confin1, Expr1 et Mob1 qui ont un coefficient de variation supérieur à 10% .

Afin de mieux juger de la précision de l'ensemble des variables, il est intéressant de les comparer aux variances que l'on aurait obtenues en l'absence du redressement, puis dans le cas d'un pseudo sondage aléatoire simple.

III.3.2. Effet du redressement.

Nous avons réalisé un redressement de l'échantillon des répondants VQS afin de corriger la non-réponse, le défaut de couverture et les fluctuations d'échantillonnage VQS et HID. Nous allons donc regarder si le redressement a amélioré la précision des estimateurs.

L'effet de calage est le rapport de la variance avant calage par la variance après calage.

variable d'intéret	variance avant calage	variance après calage	effet de calage
AIDKI1	130525285112.0	23268269755.18	5.6096
ALLOC1	57374007831.8	30645469026.39	1.8722
CONFIN1	9449388243.3	3863599766.99	2.4457
COTOR1	85860653422.2	19990923768.76	4.2950
DADAPT1	4694645599.3	1837940274.48	2.5543
DEFI1	3585995486352	219976186129.84	16.3017
EXPR1	133914155203.4	36366676550.13	3.6823
HANDI1	3480367537473	318425419990.21	10.9299
INVAL1	154292478930.4	84341974284.11	1.8294
MOB1	226465624.0	227261953.58	0.9965
MOB2	8302489193.6	1771495405.55	4.6867
MOB3	8814480498.6	1242856681.76	7.0921

Nous constatons que l'effet de calage est supérieur à 1, sauf en ce qui concerne la variable Mob1 où il est égal à 1. Cela signifie que le fait d'avoir redressé l'échantillon a augmenté la précision de l'ensemble des variables d'intérêt. La moyenne harmonique sur l'ensemble des douze variables de l'effet de calage vaut 3.
Les variables qui enregistrent le plus fort effet de calage sont les variables Defi1 (16) et Handi1 (11). Le redressement a donc permis d'estimer le nombre de métropolitains souffrant d'au moins une déficience avec une précision égale à celle que l'on aurait eu sur un échantillon de taille 16 fois plus grande.

III.3.3. Examen de la pertinence du plan de sondage.

Afin de juger de la qualité du plan de sondage HID, plan très complexe, Poulpe calcule la variance obtenue si l'échantillon final était tiré par sondage aléatoire simple mais à poids égaux à ceux obtenus par le plan de sondage HID. On pourra ainsi savoir dans quelle mesure l'échantillonnage complexe a agi sur la précision de l'enquête.
On calcule pour cela l'effet de sondage qui est le rapport de la variance sur arbre complet par la variance dans le cas d'un pseudo sondage aléatoire simple. Les variances calculées sont les variances après calage.

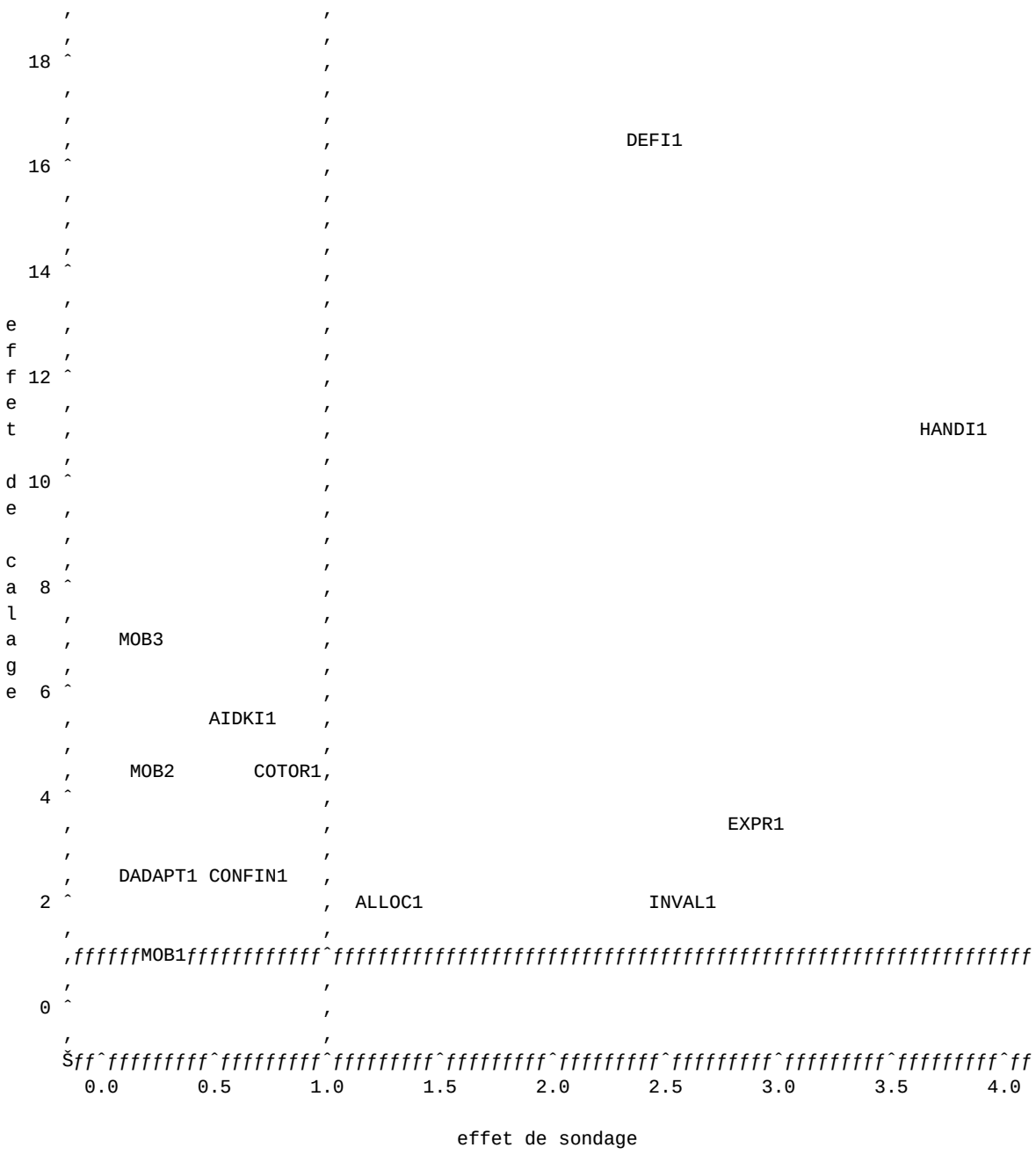
Variable d'intéret	variance sur arbre complet	variance dans le cas d'1 SAS	effet de sondage
AIDKI1	23268269755.18	39558693209.10	0.58820
ALLOC1	30645469026.39	24494192833.49	1.25113
CONFIN1	3863599766.99	5933565998.96	0.65114
COTOR1	19990923768.76	24383442072.54	0.81986
DADAPT1	1837940274.48	7160555351.86	0.25668
DEFI1	219976186129.84	90675215754.75	2.42598
EXPR1	36366676550.13	12437346536.94	2.92399
HANDI1	318425419990.21	84774006810.31	3.75617
INVAL1	84341974284.11	32965229144.21	2.55851
MOB1	227261953.58	983354613.17	0.23111
MOB2	1771495405.55	8877428728.75	0.19955
MOB3	1242856681.76	8372590563.30	0.14844

Nous constatons que sept des douze variables d'intérêt présentent un effet de sondage inférieur à 1 : Mob1,2,3, Dadapt1, Confin1, Aidki1 et Cotor1.
Ce sont justement les variables qui définissent la dépendance des personnes.

Sur l'ensemble des variables étudiées, l'effet de sondage a une moyenne harmonique de 0.77, donc l'échantillon de l'enquête HID a été correctement tiré et la précision est globalement meilleure que celle que l'on aurait obtenue par sondage aléatoire simple.

Les variables Mob1,2,3 et Dadapt1 ont un effet de sondage inférieure à 0.25, ce qui signifie que le plan de sondage a permis d'obtenir une précision quatre fois plus grande que le sondage aléatoire simple ; si les individus étaient tirés au hasard dans la population métropolitaine, il aurait fallu un échantillon quatre fois plus grand pour obtenir la même précision.

On utilise la procédure PLOT de SAS afin d'observer la position des variables d'intérêt dans le plan défini par l'effet de calage et l'effet de sondage.



Les variables Handi1 et Defi1 qui ont un très bob effet de calage présentent au contraire des effets de sondage élevés.
Nous nous attendions à un effet de sondage inférieur à 1 pour la variable Defi1 qui indique si l'individu souffre d'au moins une déficience. Au contraire l'effet de sondage est de 2.42, ce qui est élevé pour une variable comme Defi1. De même, la variable Handi1 présente un effet de sondage de 3.71 . Cette variable correspond à la première question posée à l'enquête HID: « Rencontrez-vous dans la vie de tous les jours des difficultés, qu'elles soient physiques, sensorielles, intellectuelles ou mentales ? ». L'interviewé répond par oui ou non et c'est l'ensemble des réponses au questionnaire qui permettra de synthétiser le nombre de déficiences dans la variable Defi1. Et on a constaté des déficiences chez plusieurs personnes qui avaient spontanément déclaré n'avoir aucune difficulté. Ce qui explique que les variables Defi1 et Handi1 sont moins bien estimées par le plan de sondage que ce que l'on aurait pu croire. Ces variances sont d'autant plus élevées que la dispersion des réponses à ces variables vient principalement des strates HID 1 et 2, dans lesquelles les individus ont un très gros poids, d'où la forte variance.

Ces différents résultats nous permettent d'émettre des suppositions sur la façon dont les personnes ont répondu aux deux questionnaires VQS / HID, sur les handicaps et incapacités dont elles ont pleinement conscience, voire sur les incapacités qu'elles sont disposées à déclarer.
En effet, les variables qui ont bon effet de sondage sont celles qui entrent directement dans la définition des dépendances : le fait d'être confiné, d'avoir besoin d'aide pour les tâches de la vie quotidienne, d'avoir des équipements spécialement adaptés à son handicap, d'avoir besoin d'aide pour se déplacer ou encore le fait d'avoir fait une demande de reclassement professionnel auprès de la Cotorep. Les personnes concernées ont pleinement conscience de leur état, elles ont certainement correctement déclaré les problèmes de cette catégorie à l'enquête VQS. Par conséquent la stratification par la strate HID a permis d'obtenir des variances plus faibles que si on avait tiré les individus au hasard dans la population métropolitaine.
A l'inverse, les variables comme Defi1 ne gagnent pas en précision, au contraire. Seules des questions précises permettent de mettre en évidence les déficiences que les personnes concernées n'identifient pas comme telles.
Il y a aussi une autre catégorie de variables, comme Expr1 (le fait de ne pas savoir lire, écrire ou compter) qui ne gagne pas en précision par le plan de sondage. Est-ce parce que cette déficience intellectuelle ne rentre pas en compte dans la définition du handicap ? Où est-ce parce les personnes déclarent mal (volontairement ou involontairement) ce genre de difficulté ?

III.4. Etude de quelques prévalences.

L'informaticien responsable de la réalisation du logiciel Poulpe a pu apporter dans des délais assez brefs une première amélioration au logiciel qui consiste à augmenter la taille limite du fichier des données, ce qui permet de travailler avec le fichier complet et donc de calculer des estimations de variance sur des statistiques complexes.
Compte tenu du temps d'exécution des différentes macros et des délais dont nous disposons, nous pouvons déjà présenter les premiers résultats d'estimation de ratios.

Une des premières questions que l'on se pose en réalisant des études sur le handicap est de savoir s'il existe une différence significative entre les hommes et les femmes.

Regardons ce qu'il en est en ce qui concerne les déficiences, le fait d'avoir besoin de l'aide régulière d'une tierce personne et le fait de percevoir une allocation de handicap.

Poulpe estime le pourcentage d'hommes et le pourcentage de femmes qui :

- souffrent d'au moins une déficience : TDEFH (pour homme) et TDEFF (pour femme) ;
- ont besoin d'aide aide régulière : TAIDH et TAIDF ;
- perçoivent une allocation de handicap : TALLH et TALLF.

Et il fournit les intervalles de confiance estimés à 95% de ces pourcentages.

variable d'intéret	ratio estimé (poids réels)	ecart-type du ratio estimé	borne inférieure (IC à 95%)	borne supérieure (IC à 95%)
TAIDH	0.06609	.0049482	0.05639	0.07579
TAIDF	0.10741	.0015798	0.10431	0.11050
TALLH	0.05336	.0026426	0.04818	0.05854
TALLF	0.02542	.0029921	0.01955	0.03128
TDEFH	0.36921	.0008458	0.36755	0.37086
TDEFF	0.40383	.0048141	0.39440	0.41327

Ainsi on estime que 36% des hommes souffrent d'au moins une déficience, physique, sensorielle, intellectuelle ou mentale contre 40% des femmes. Et 6% des hommes ont besoin de l'aide régulière d'une personne pour les tâches de la vie quotidienne, contre 6% des hommes.
Par contre, seuls 2.5% des femmes perçoivent une allocation pour un handicap, contre 5% des hommes, soit deux fois moins.

La confrontation des intervalles de confiance de ces différents pourcentages nous permet de conclure à une différence significative entre les hommes et les femmes.

- on trouve plus de déficiences chez les femmes que chez les hommes;
- celles-ci sont plus dépendantes que les hommes ;
- elles sont toutefois deux fois moins nombreuses à bénéficier d'une allocation de handicap.

On peut peut-être expliquer ce dernier constat par le fait que les déficiences ne sont pas de la même nature chez les hommes que chez les femmes. Une grande part des allocations dont bénéficient les hommes correspond à des allocations pour accidents du travail ou maladies professionnelles. Jusqu'à très récemment, c'est le rapport au travail qui structurait la reconnaissance sociale du handicap.
Des études plus approfondies nous permettront de savoir quelles sont les tranches d'âge qui bénéficient le plus des allocations, de quelles déficiences les hommes souffrent le plus, pareillement pour les femmes, ainsi que leurs origines.

Nous pouvons également étudier la dépendance des personnes selon le groupe VQS auquel elles appartiennent, à l'aide de la variable Aidki1 (présence d'une aide régulière pour accomplir les tâches de la vie quotidienne en raison d'un problème de santé ou d'un handicap).
Rappelons les critères des différents groupes.

- Groupe 1: personnes déclarant ne souffrir d'aucune difficulté ;
- Groupe 2: personnes déclarant une seule difficulté ;
- Groupe 3: personnes déclarant « avoir un handicap » ou souffrir d'une « limitation d'activité » ou dépendre d'une aide humaine ;
- Groupe 4: personnes déclarant « avoir un handicap » et personnes déclarant souffrir d'une « limitation d'activité », déclaration appuyée par des items d'aide humaine ou technique.
- Groupe 5: personnes déclarant « avoir un handicap », déclaration fortement appuyée par d'autres items ;
- Groupe 6: personnes déclarant avoir obtenu une reconnaissance de leur handicap (plus, pour les moins de 16 ans : enfants et adolescents inscrits dans une classe ou un établissement spécialisé).

Nous estimons le pourcentage de personnes de groupe ayant besoin d'une aide régulière. En fait nous avons fait des classes de deux groupes VQS.
Poulpe nous renvoie les résultats suivants :

Groupes VQS	ratio estimé (poids réels)	ecart-type du ratio estimé	borne inférieure (IC à 95%)	borne supérieure (IC à 95%)
Groupes VQS 1 et 2	0.03127	0.000561	0.03017	0.03237
Groupes VQS 3 et 4	0.23370	0.011075	0.21199	0.25541
Groupes VQS 5 5et 6	0.36644	0.021952	0.32342	0.40947

Nous estimons que 36% des personnes le plus sévèrement et le plus certainement handicapées (groupes VQS 5 et 6) bénéficient de l'aide permanente ou régulière d'une personne pour les tâches de la vie quotidienne, contre 23% pour les personnes des groupes 3 et 4.
L'estimation de l'IC dit qu'il y a 95% de chance pour que la vraie valeur du pourcentage soit comprise entre 32 et 41 %. Il y a donc un fort pourcentage de personnes sévèrement handicapées qui n'a pas besoin d'aide humaine. Ces personnes arrivent à être indépendantes, à vivre comme les autres chez elles, grâce à des équipements spécialement adaptés à leur état.

D'autres études de prévalences vont être menées, notamment selon l'âge, l'origine de l'incapacité et la classe sociale.

IV. Conclusion.

Deux aspects sont mis en évidence par l'analyse de la précision de l'enquête HID.
D'une part, le redressement de l'échantillon final a largement augmenté la précision des estimateurs. Il a même permis d'estimer le nombre de métropolitains vivant en domicile ordinaire qui souffrent d'une déficience physique, sensorielle, intellectuelle ou mentale avec une précision égale à celle qu'on aurait obtenu sur un échantillon de taille seize fois plus grande.
D'autre part, le plan de sondage HID, très complexe, a permis d'obtenir des estimateurs plus précis que ceux que l'on aurait eu à partir d'un sondage aléatoire simple, sur les variables qui définissent la dépendance des personnes.
Cependant, il faut mentionner un fait très surprenant : les résultats ont montré que le nombre de personnes souffrant d'une déficience n'est pas estimé avec autant de précision qu'il aurait dû l'être. On obtiendrait même un estimateur deux fois plus précis par le sondage aléatoire simple.
Le plan de sondage HID n'est pas optimal pour estimer le nombre de personnes souffrant d'une déficience ou encore le nombre de personnes ne sachant pas lire, ou écrire ou compter.
Mais le calage sur marges permet d'augmenter très fortement la précision.

L'analyse des premiers résultats d'estimation de ratios met en évidence une différence significative entre les hommes et les femmes en ce qui concerne les déficiences, l'aide humaine et la reconnaissance du handicap. Les femmes sont plus nombreuses à souffrir d'au moins une déficience que les hommes (40% contre 36% pour les hommes), elles sont aussi plus nombreuses à bénéficier de l'aide régulière d'une personne pour accomplir les tâches de la vie quotidienne. Par contre, les hommes sont deux fois plus nombreux à percevoir une allocation en raison d'un handicap (5% contre 2.5% pour les femmes).

La mise en œuvre du logiciel Poulpe pour estimer la précision de l'enquête HID a le mérite de souligner les apports et les limites du logiciel d'une part, et les soucis que devraient avoir les concepteurs d'enquête d'autre part.

Poulpe permet de mettre en œuvre des calculs de précision sur la plupart des plans de sondages complexes et de prendre en compte des enquêtes redressées. Il permet de chiffrer le gain de précision apporté par le calage sur marge. Par le calcul des effets de sondage, il permet également de juger de la pertinence de l'échantillonnage d'une enquête.

Cependant, la version actuelle du logiciel ne permet pas de prendre en compte plus de 99 strates de deuxième phase, les macros sont rédigées en version 6.12 de Sas, version qui tend à disparaître des micros au profit de la version 8. De plus, le lancement du logiciel est très coûteux en temps et en espace disque pour des enquêtes à gros échantillons.

Il faudrait que les concepteurs d'enquête prennent en compte les calculs de précision au moment où ils réalisent l'échantillonnage, en évitant notamment de tirer des échantillons de taille égale à 1. Il est en outre très regrettable que certaines données du tirage n'aient pas été archivées pour HID ou par le recensement.

L'application du logiciel à une grosse enquête comme HID a mis en évidence des difficultés non négligeables en raison de cas particuliers qui n'avaient pas été envisagés. Il a été suggéré de modifier le logiciel afin de le rendre plus performant, plus souple et plus rapide.

Citons notamment :
1- l'augmentation de la taille limite du fichier de l'enquête,
2- la possibilité de ne rentrer les variables d'intérêt que dans l'étape d'estimation de la variance, et donc sans avoir à relancer les étapes précédentes (notamment l'étape de calcul des probabilités d'inclusion). Ceci éviterait de rentrer toutes les variables dès la première étape et on pourrait ainsi gagner en temps et en espace disque.

CONCLUSION

Ce stage, effectué dans l'organisme pilote de la statistique publique française, m'a permis d'étudier la théorie des sondages et de voir les différents problèmes et réajustements que demande un sondage. Grâce à ce projet, j'ai pu avoir une première approche des difficultés que l'on peut rencontrer pour effectuer des estimations de variance à partir d'un plan de sondage complexe, et surtout savoir comment les aborder et les résoudre sans s'éloigner de la réalité.
Cette expérience m'a permis de pratiquer intensément le logiciel SAS, et j'ai reçu une formation de deux journées à SAS Macros.

Ce travail m'a offert la possibilité de développer des qualités d'organisation, de rigueur, de travail en équipe. J'ai apprécié le fait qu'une grande autonomie m'ait été laissée dès le début, et particulièrement le fait que des responsabilités me soient confiées.

La plus grande satisfaction de mon stage est d'avoir été considérée comme un membre à part entière de l'équipe de projet HID, et d'avoir contribué à l'avancement du travail d'estimation de la variance.
Des difficultés non négligeables ont été rencontrées dans l'application du logiciel Poulpe à HID, liées aux spécificités mêmes du logiciel. Et ce travail a permis de les mettre en lumière afin que l'INSEE puisse améliorer le logiciel, le rendre plus performant et plus souple d'utilisation.

ANNEXE 1 : Structure du questionnaire HID

I. Le déroulement de l'enquête.

La collecte s'est déroulée en quatre phases :

1. Entre **octobre et décembre 1998**, le questionnaire détaillé de l'enquête HID a été administré à un échantillon d'environ 15 000 personnes vivant en institutions pour personnes handicapées, en institutions pour personnes âgées, dans les services hospitaliers de long séjour et dans les établissements psychiatriques.

2. En **mars 1999**, un court questionnaire de filtrage VQS (Vie Quotidienne et Santé) a été collecté pendant le recensement de la population auprès d'un échantillon d'environ 400 000 personnes vivant en domicile ordinaire. Environ 360 000 personnes ont répondu à l'enquête nationale VQS.

2bis. Entre **octobre 1999 et janvier 2000**, le questionnaire détaillé de l'enquête HID a été administré à un échantillon d'environ 20 000 personnes tirées parmi les répondants à l'enquête de filtrage. Environ 17 000 personnes ont répondu à l'enquête.

3. **Fin 2000**, les personnes vivant en institutions interrogées en octobre 1998 ont été à nouveau enquêtées afin d'analyser l'évolution des situations individuelles.

4. **Fin 2001**, un second passage a également eu lieu auprès des personnes vivant en domicile ordinaire.

II. La structure du questionnaire HID auprès des personnes vivant à domicile.

Le questionnaire débute par un tableau de composition du ménage. Ensuite, les questions sont regroupées par modules, au nombre de 11.

- **Module A : causes et origines des incapacités**. Il s'agit d'un tableau comprenant la liste des déficiences dont souffre la personne, et de leur origine (maladies, accidents, problème de naissance, vieillissement,...).

- **Module B : description des incapacités et de leur ancienneté**. Cette partie du questionnaire passe en revue les différents actes de la vie courante, pour relever les éventuelles difficultés d'accomplissement auxquelles se heurte la personne. En même temps, elle interroge sur l'ancienneté de chacun des problèmes rencontrés ainsi que sur leur origine.

- **Module C : environnement socio-familial** de la personne, recensement des aidants,…

- **Module D : accessibilité du logement**, aménagement de celui-ci pour des raisons de santé et **aides techniques** dont dispose ou que souhaiterait la personne.

- **Module L : conditions de logement** . Cette partie interroge la personne sur sa condition de locataire ou de propriétaire et sur la proximité des équipements (bureau de poste, arrêt d'autobus, supermarché, etc.).

- **Module T : déplacement et transport.**

- **Module S : scolarité et diplômes.**

- **Module E : emploi**, présent ou passé, recherche d'emploi, **origine sociale**.

- **Module R : revenus, allocations, statut juridique, reconnaissance officielle.**

- **Module G :** questionnement général sur les **loisirs, vacances, sports, culture…**

- **Module W : interview à l'aidant principal.** Cette partie n'est réalisée que s'il y a un aidant principal et si l'aide n'est pas fournie par des professionnels. L'aidant est interrogé sur les conséquences de son rôle sur sa vie professionnelle, sur sa vie privée, les conséquences sur sa santé physique et morale.

III. Durée du questionnement HID.

L'interview individuelle HID a duré en moyenne 40 minutes.
La durée de l'interview dépend fortement du degré de handicap probable des personnes. La moyenne est de 33 minutes pour le groupe VQS 1 (aucune difficulté déclarée à l'enquête VQS) et elle est de 41 minutes pour le groupe 6 (handicap lourd).
L'âge influe également sur la durée de l'interview. Elle est de 28 minutes pour les moins de 16 (sûrement pas par qu'un grand nombre ne les concernait pas) et de 43 minutes pour les plus de 60 ans.

Le questionnaire de l'aidant comportait environ 20 questions et a duré 3 minutes en moyenne.

ANNEXE 2 : Eléments de la théorie des sondages

On parle de **sondage** pour tout type d'enquête effectuée auprès d'un échantillon de la population seulement, par opposition à la population tout entière.
L'individu est l'unité que l'on veut sonder, il est aussi appelé **unité d'observation**.
Le **plan de sondage** est l'ensemble des étapes de tirage de l'échantillon.

Soit N la taille de la population et n la taille de l'échantillon tiré. On appelle **taux de sondage** le rapport $\frac{n}{N}$.

On parle de **défaut de couverture** lorsque la population dans laquelle on effectue le tirage n'est pas la population tout entière, certains individus étant volontairement exclus du tirage.
On parle de **non-réponse** lorsque l'on ne dispose pas des réponses pour un ou plusieurs individus de l'échantillon tiré, soit par refus de répondre, soit par absence.

I. Définition des sondages élémentaires.

1. Le sondage aléatoire simple (SAS) consiste à tirer dans une population de taille N un échantillon de taille n sans remise, de façon à ce que chaque individu ait la même probabilité de tirage $\frac{n}{N}$, et cela sans aucune manipulation préalable dans la population.

2. L'algorithme du tirage systématique (SYS)est utilisé pour tirer des individus au hasard, chaque individu ayant la même probabilité de tirage. Si on désire que l'échantillon soit représentatif de la population selon un ou plusieurs critères, on ordonne les individus selon ces critères.

Il faut calculer le pas en nombre d'individus qui vaut Pas = int ($\frac{n}{N}$), int (a) étant la partie entière du réel positif a. Le premier individu est tiré au hasard en début de fichier de la manière suivante : on détermine un nombre au hasard entre 0 et 1 appelé aléa, le premier individu sera l'individu de rang int (aléa* Pas).
Partant de cet individu, le principe consiste à descendre ensuite le long du fichier en retenant un individu tous les Pas individus.

3. Le sondage stratifié (EXH) consiste à découper la population en plusieurs groupes appelés strates, puis à effectuer un tirage dans chacune des strates, les tirages étant indépendants les uns des autres.
Soient N_h la taille de la population de la strate H et n_h la taille de l'échantillon tiré dans la strate h. Un sondage stratifié est dit **à allocation proportionnelle** si $\frac{n_h}{n} = \frac{N_h}{N}$.

4. Le sondage à probabilité proportionnelle à la taille (PPT) est un cas particulier du sondage à probabilités inégales. Chaque unité de tirage n'a pas la même probabilité P_i

d'être sélectionnée, cette probabilité va dépendre de la valeur T_i prise par la variable de taille. Cette variable de taille peut-être la population (par exemple si on tire des communes, ainsi plus la commune est grande, plus elle a de chance d'être sélectionnée), le revenu de l'individu, etc.

La probabilité de section vaut : $P_i = n \mathbf{x} \frac{T_i}{T}$, T étant la somme des T_i sur la population entière.

II. Les sondages en plusieurs étapes.

Le sondage à plusieurs degrés consiste à découper la population en plusieurs groupes, puis à découper chaque groupe en plusieurs sous-groupes, ainsi de suite. La première étape de tirage consiste à tirer des groupes, appelés **unités primaires**. Puis à l'intérieur de chaque groupe tiré, on tire des sous-groupes, appelés **unités secondaires**. Ainsi de suite. On appelle **degré de tirage** chaque étape élémentaire de tirage.
Par exemple, si on désire effectuer une enquête sur les logements français, on peut tirer des départements (unités primaires), puis dans chaque département tirer des logements. Dans ce cas, on parle de sondage à deux degrés car le tirage des logements constitue la deuxième étape de tirage. On peut aussi tirer des départements, puis des communes, puis des logements. On parlera dans ce cas d'un sondage à trois degrés.

Un sondage est dit aréolaire si le dernier degré de tirage consiste à tirer une grappe d'individus. Par exemple, si au lieu de tirer des logements dans les communes on tire des immeubles et qu'on décide de sonder tous les logements des immeubles tirés.

Le sondage en deux phases consiste à tirer dans un premier temps un échantillon de n' individus dans une population de taille N, éventuellement suivant un plan complexe (stratification, tirage à plusieurs degrés). Puis dans un deuxième temps, on tire parmi les n' individus un échantillon de n individus. On appelle **phase de tirage** l'ensemble des étapes de tirage de chaque échantillon.

Le sondage en deux phases avec post-stratification est le cas où la deuxième phase est stratifiée par une variable auxiliaire. Le cas fréquent est de tirer un échantillon de n' individus que l'on va interroger afin de récolter une information Z. La post-stratification consiste à constituer des catégories selon les valeurs de Z, à ranger chacun des n' individus dans sa catégorie, puis à effectuer un tirage dans chaque catégorie.

III. Estimation de la variance après un redressement par calage sur marges.

Soit un échantillon S de taille n tiré dans une population de taille N, suivant un plan de sondage éventuellement très complexe. Soit θ un paramètre à estimer.
Le redressement a pour objectif d'améliorer la précision de l'estimateur de θ en utilisant une ou plusieurs informations auxiliaires.
Le calage sur marges est une méthode largement employée pour redresser les enquêtes de l'INSEE.

Supposons que l'on souhaite estimer la dépense annuelle de l'ensemble des ménages français en nourritures. On a tiré un échantillon S de ménages, chaque ménage tiré a un poids D_i défini par l'inverse de sa probabilité de tirage. Soit Y_i la dépense annuelle en nourriture du ménage i. θ est la somme des Y_i sur la France.
Supposons également que l'on dispose de la structure des ménages de la France par catégories socioprofessionnelles (CSP) ainsi que la structure par taille du ménage, par exemple grâce aux données du recensement de la population.

Soit $N_{j.}$ le nombre de ménages français dont le chef de famille appartienne à la CSP j et soit $N_{.k}$ le nombre de ménages français de taille k.
Il peut se trouver que la somme des poids des ménages de l'échantillon S dont le chef de famille appartienne à la CSP j soit différente de $N_{j.}$ et que la somme des poids des ménages de taille k de l'échantillon S soit différente de $N_{.k}$.

Le redressement par calage sur marge consiste à définir un nouveau jeu de pondérations W_i qui permette d'estimer parfaitement chacun des effectifs connus N_k et qui soit construit de telle sorte qu'il s'éloigne le moins possible de la pondération d'origine D_i :

1) $\sum_{i \in S} W_i X_{ij} = N_{j.}$ où X_{ij} vaut 1 si le ménage i appartient à la CSP j et 0 sinon.

$\sum_{i \in S} W_i Z_{ik} = N_{.k}$ où Z_{ik} vaut 1 si le ménage est de taille k et 0 sinon.

2) minimiser $\sum_{i \in S} D(W_i, D_i)$ où D(a, b) est la distance entre les nombre réels a et b.

W_i et d_i sont liés par la relation : $W_i = \lambda_{j,k}$ **x** D_i , où $\lambda_{j,k}$ est le coefficient de redressement qui dépend uniquement du croisement de la modalité j de la CSP et de la modalité k de la taille du ménage.

Des logiciels comme Calmar permettent de trouver les nouveaux poids W_i .
Le total de la variable Y est estimée sans biais par l'estimateur de Horvitz-Thomsom défini par :

$$\hat{y} = \sum_{i \in S} W_i Y_i$$

Du point de vue de la variance, on montre que lorsque la taille de l'échantillon S est grande, la variance de $\hat{y}$ est approximativement égale à celle d'**un estimateur par la régression** :

$Y_{i,j,k} = e + a_j + c_k + \varepsilon_{i,j,k}$

e est un paramètre réel, les a_j sont les paramètres traduisant les effets spécifiques des diverses modalités de la variable CSP sur Y, les c_k sont les paramètres traduisant les effets spécifiques des modalités de la variable taille du ménage sur Y, et les $\varepsilon_{i,j,k}$ sont les résidus traditionnels dont la somme étendue à tous les ménages de la France vaut 0.

Soient $\hat{e}$, $\hat{a}_j$ et $\hat{c}_k$ les estimateurs des moindres carrés de e, a_j et c_k .
Si l'échantillon S a été tiré par sondage aléatoire simple, la théorie des modèles linéaires nous dit qu'un estimateur sans biais de la variance V de Y est :

$$\hat{V} = \frac{1}{n-1} \sum_{i \in S} \sum_{(j,k)} \left(Y_{i,j,k} - \hat{e} - \hat{a}_j - \hat{e}_k \right)^2$$

Dans le cas d'un sondage aléatoire simple, la variance de l'estimateur du total vaut :

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

Par conséquent, l'estimateur de la variance de $\hat{y}$ vaut :

$$\hat{V}(\hat{y}) = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \frac{1}{n-1} \sum_{i \in S} \sum_{(j,k)} \left(Y_{i,j,k} - \hat{e} - \hat{a}_j - \hat{e}_k \right)^2$$

ANNEXE 3 : Programmes SAS

Programme c:\user\odile\tirageHID\tauxsondage.sas

```
/*******************************************************/
/*         Création du tableau des taux de sondage     */
/*         par zone d'enquête et par strate HID        */
/*                                                     */
/************* Mardi 4 Juin 2002 ***********************/
/*         C:\User\odile\tirageHID\tauxsondage.sas     */
/*                                                     */
/*******************************************************/

options obs = max;

/*  1.  On récupère l'ensemble des répondants VQS      */
/*      ainsi que l'échantillon HID tiré.              */
/*******************************************************/

libname tirage 'c:\user\odile\tiragehid\tables';

data repvqs1; set tirage.vqsdef (keep = dep com strate idvqsdef);
              nb  = 1;
run;

data echhid1; set tirage.hidredrp (keep = dep com strate);
             nb = 1;
run;

/*  2.  On définit les zones d'enquete.               */
/*      (récupérées dans les programmes régionaux).   */
/*******************************************************/

data toto; set repvqs1(in=a) echhid1(in=b);
           c = a; d = b;
           depcom = dep!!com;
           y='Z';

* région 21( Champagne, Picardie exclue mais avec Seine et Marne);
if depcom in ('08126','08476') then zondeg='081ASF';
      else if depcom='08105' then zondeg='082CHA';
      else if depcom='10333' then zondeg='101STA';

[...]

* région 94(La Corse);
     else if dep='2A' then zondeg ='201AJA';
run;

/*  3.  On définit les zones d'enqute pour les       */
/*      communes n'ayant pas participé au tirage HID.*/
/*******************************************************/

data toto2; set toto;
            w ='999H';
            if zondeg=' ' then zondeg =dep!!w;
run;


/*  4.  Finalament on récupère les deux tableaux.    */
/*******************************************************/

data repvqs2; set toto2;
              if c;
run;

data repvqs; set repvqs2(keep = depcom zondeg strate nb idvqsdef);
proc sort; by zondeg strate;
run;

data echhid2; set toto2;
              if d;
```

```
run;

data echhid; set echhid2(keep = depcom zondeg strate nb);
proc sort; by zondeg strate;
run;

* petite vérification;
data verif; set echhid;if substr(zondeg,3,4)='999H';run;

/*  5. On va compter le nombre de répondants VQS       */
/*     par zone d'enquête et par strate HID.           */
/*****************************************************/

proc summary data = repvqs;
     class zondeg;
     var nb;
     output out = tab11 sum = total1;
proc print data = tab11;
run;

proc summary data = repvqs;
class strate;
     var nb;
     output out = tab12 sum = total1;
proc print data = tab12;
run;

proc summary data = repvqs nway;
     class zondeg strate;
     var nb;
     output out = tab13 sum = total1;
proc print data = tab13;
run;

/*  6. On compte le nombre d'individus tirés par        */
/*     zone d'enquete et par strate pour HID.           */
/*****************************************************/

proc summary data = echhid;
     class zondeg;
     var nb;
     output out = tab21 sum = total2;
proc print data = tab21;
run;

proc summary data = echhid;
     class strate;
     var nb;
     output out = tab22 sum = total2;
proc print data = tab22;
run;

proc summary data = echhid nway;
     class zondeg strate;
     var nb;
     output out = tab23 sum = total2;
proc print data = tab23;
run;

/*  7. On construit une table avec le nombre            */
/*     répondants VQS et le nombre d'individus          */
/*     tirés par zone d'enquete et par strate HID       */
/*     afin de calculer les taux de sondage.            */
/*****************************************************/

* taux globaux par zone d'enquete;

data tab11; set tab11(keep = zondeg total1);
proc sort; by zondeg;
run;

data tab21; set tab21(keep = zondeg total2);
proc sort; by zondeg;
run;

data tirage.parzone; merge tab11 tab21;
             by zondeg;
```

```
                        if total2 = . then total2 = 0;
                        taux = total2 / total1;
label total1 = "Nombre de répondants VQS"
      total2 = "Effectif des individus HID tirés"
         taux = "Taux globaux de sondage"
         zondeg = "Zone d'enquete";
run;
proc print data = _last_ label;
title 'Taux globaux par zone';
run;

* taux globaux par strate HID;
data tab12; set tab12(keep = strate total1);
proc sort; by strate;
run;

data tab22; set tab22(keep = strate total2);
proc sort; by strate;
run;

data tirage.parstrat; merge tab12 tab22;
             by strate;
                        taux = total2 / total1;
label total1 = "Nombre de répondants VQS"
      total2 = "Effectif des individus HID tirés"
         taux = "Taux globaux de sondage"
         strate = "Strate HID";
run;
proc print data = _last_ label;
title 'Taux globaux par strate HID';
run;

* taux de sondage;
data tab13; set tab13(keep = zondeg strate total1);
proc sort; by zondeg strate;
run;

data tab23; set tab23(keep = zondeg strate total2);
proc sort; by zondeg strate;
run;

data tirage.tauxsond; merge tab13 tab23 ;
            by zondeg strate;
            if total2 = . then total2 = 0;
            taux = total2 / total1;
label total1 = "Nombre de répondants VQS"
      total2 = "Effectif des individus HID tirés"
         taux = "Taux de sondage réels"
         zondeg = "Zone d'enquete"
         strate = "Strate HID";
run;
proc print data = _last_ label;
title 'Taux de sondage HID';
run;

/*  7.  Examen des moyennes,extrèmes et dispersions       */
/*      des taux de sondage par strate HID.               */
/*      On ne tiendra pas compte des zones d'enquete      */
/*      dans lesquelles on n'a pas tiré d'individu HID.   */

data taux; set tirage.tauxsond;
           if taux NE 0;
run;

proc means data = taux;
           class strate;
           var taux;
                 output out = tirage.disptaux(drop = _type_ _freq_)
                 mean = Moyenne
                 min = Minimum
                 max = Maximum
                 std = Ectype
                 range = Ecart;
run;
proc print;
title 'Examen des dispersions des taux de sondage par strate HID';
run;
```

```
/*  8. Examen de la dipersion des poids.                */
/*     Ici également on ne tiendra pas compte           */
/*     des zones sans individus HID.                    */
/*******************************************************/

data poids; set tirage.tauxsond;
            if taux NE 0;
            poids = 1/taux;
run;

proc means data = poids;
           class strate;
           var poids;
                 output out = tirage.dispoids(drop = _type_ _freq_)
                 mean = Moyenne
                 min = Minimum
                 max = Maximum
                 std = Ectype
                 range = Ecart;
run;
proc print;
title 'Examen de la dispersion des poids par strate HID';
run;

/*  Toutes les tables sont enregistrées dans:           */
/*  C:\User\Odile\tirageHID\tables.                     */
/*******************************************************/

data tirage.repvqs; set repvqs;
run;

data tirage.echhid; set echhid;
run;
```

Programme c:\user\odile\fichier\geo2.sas

```
/*********************************************************************/
/*       Programme de création du deuxième fichier géographique.     */
/*********************************************************************/
/* On considère ici que le second degré de tirage consiste en un ti- */
/* de districts au RP 99 à l'intérieur des ZD VQS.                   */
/* Donc il faut renseigner le nombre de districts par ZD VQS et la   */
/* pop 99 de chaque district interrogé, afin de traiter la non - ré- */
/* ponse VQS par un troisième de gré de tirage SAS. Il faut égale-   */
/* ment renseigner la pop 90 pour chaque strate et pour ZD.          */
/*                                                                   */
/***************** Vendredi 19 Juillet 2002 ************************/
/*                 c:\user\odile\fichier\geo2.sas                    */
/*********************************************************************/

libname fichier 'c:\user\odile\fichier';
libname vqs 'c:\user\odile\tirageVQS';

/* On récupère la liste des districts interrogés avec leur pop 99    */
/* dans le fichier FICHIER \ VQSZD .                                 */
/*********************************************************************/

data zd; set fichier.vqszd ( drop = ct ef dble dble1 );
               if codedel = 'DR283114' and dist = '0004' then do;
                  il0 = '0004'; pop99 = 669; strate = 28;
               end;
               length distri $13;
               il = right(il0);
               distri = dep!!com!!cil!!il!!frdist;
        proc sort; by codedel distri;
run;

data zd2; set zd;
             by codedel distri;
                if first.codedel or first.distri then output;
run;

data distri(drop = pop99); set zd2 (keep = codedel distri pop99 strate);
```

```
                                        NNN = pop99;
                                tailuni = pop99;
                                auxniv = 3;
                                proc sort; by strate codedel distri;
run;

NOTE: The data set WORK.DISTRI has 2273 observations and 6 variables;

* Il faut noter que quatre districts ont une pop 99 nulle !!! ;
* Il s'agit des districts n°21, 22, 23 et 160 ;


/* Le fichier VQS\rp99cs contient la liste des districts de   */
/* chaque ZD au RP 99. On va les compter.                     */
*/
/************************************************************/

data rp99cs; set vqs.rp99cs;
                    proc sort; by codedel;
run;

NOTE: The data set WORK.RP99CS has 303011 observations and 10 variables;

data rp99cs; merge zd2 ( keep = codedel in = x) rp99cs;
                    by codedel;
                    if x;
                  length distri $13;
                  distri = d!!c!!cil!!il!!fil;
            proc sort; by codedel distri;
run;

NOTE: The data set WORK.RP99CS has 35476 observations and 11 variables;

data zondel; set rp99cs;
                    by codedel;
                    retain nb;
                  if first.codedel then do;
                        nb = 0;
                        end;
               nb = nb + 1;
                  if last.codedel then output;
run;

/* On récupère la population 90 des strates et des ZD dans le */
/* fichier VQS\POIDS1 pour le tirage PPT.                     */
/************************************************************/

data t1; set vqs.poids1 ( keep = strate pop90 );
               rename pop90 = tailuni;
               auxniv = 1;
        proc sort; by strate;
run;

data strate ; set t1;
                     by strate;
                     if first.strate then output;
run;

NOTE: The data set WORK.STRATE has 35 observations and 3 variables;

data poids1; set vqs.poids1 ( keep = strate codedel tot );
                    proc sort; by codedel;
run;

data codedel; merge zondel ( keep = codedel nb ) poids1;
                     by codedel;
                     rename tot = tailuni;
                     rename nb = NNN;
                     auxniv = 2;
                     proc sort; by strate codedel;
run;

NOTE: The data set WORK.CODEDEL has 391 observations and 5 variables;

/* Regardons s'il existe des échantillons de taille 1.        */
/************************************************************/
```

```
data ech  ; set zd2;
                      by codedel;
                      retain taille;
                      if first.codedel then do;
                          taille = 0;
                             end;
                      taille = taille + 1;
                      if last.codedel then output;
run;

data verif; set ech;
                      if taille = 1;
                      keep strate dep codedel distri;
                      proc sort; by strate dep codedel;
run;

NOTE: The data set WORK.VERIF has 69 observations and 4 variables;

proc print data = verif;
var codedel strate dep distri;
title 'ZD avec un échantilon de taille = 1 ';
title3 'pour tirage de districts 99 en US ';
run;

/* On ne pourra pas estimer la variance des ZD ayant un       */
/* échantillon de taille = 1. On propose d'effectuer des      */
/* regroupements par deux ou trois, à l'intérieur des strates */
/* quand cela est possible. On a quatre cas:                  */
/* - cas1: un nombre pair de ZD dans la strate                */
/*         -> regroupement par deux ;                         */
/* - cas2: 3 ZD dans la strate -> un seul paquet ;            */
/* - cas3: dep 75 dans la strate 1 a cinq ZD                  */
/*         -> faire deux groupes ;                            */
/* - cas4: strate avec une seule ZD à pb;                     */
/*         -> rattacher la ZD à une autre ZD de sa strate;    */
/* On va additionner la pop90 des ZD,  et je propose          */
/* d'additionner également le nombre de distri de la ZD afin  */
/* d'équilibrer la probabilité résultante des deux premiers   */
/* degrés de tirage.                                          */
/*************************************************************/

proc sort data = verif; by codedel; run;
proc sort data = codedel; by codedel; run;
proc sort data = distri; by distri; run;

data table; merge verif(in = x) codedel;
                      by codedel;
                      if x;
                      rename NNN = nb;
                      rename tailuni = pop90;
                      drop auxniv;
                      proc sort; by distri;
run;

data table2; merge table(in = x) distri( keep = distri NNN );
                       by distri;
                       if x;
                       rename NNN = pop99;
                       proc sort; by strate codedel;run;
run;

/*                 Traitement du cas 1.                     */
/*************************************************************/

data groupe11; set table2( where = (strate not in
                         (1,17,20,23,25,28,31,34,35)) );
                         by strate codedel;
                         code + 1;
run;

NOTE: The data set WORK.GROUPE11 has 42 observations and 8 variables;

data groupe12;set groupe11;
                      by strate codedel;
                      retain num p90 nb2;
                      if first.strate then do;
                            num = 0; p90 = 0; nb2 = 0;
```

```
                        end;
                        num = num + 1;
                    p90 = p90 + pop90;
                        nb2 = nb2 + nb;
                        if num = 2 then do;
                              output;
                              num = 0; p90 = 0; nb2 = 0;
                        end;
run;

NOTE: The data set WORK.GROUPE12 has 21 observations and 11 variables;

data groupe13; merge groupe11 groupe12(in = x);
                         by strate codedel;
                         if x = 0 then code2 = code + 1;
                         else code2 = code;
                         proc sort; by code2 descending code;
run;

data groupe14; set groupe13;
                          by code2 descending code;
                          retain y;
                          if first.code2 then do;
                             y = codedel;
                          end;
                          codedel2 = y;
run;

/*                    Traitement du cas 2.                        */
/****************************************************************/

data groupe21; set table2( where = (strate in (17,20,28)) );
                         by strate codedel;
                         retain code 42;
                         code + 1;
run;

NOTE: The data set WORK.GROUPE21 has 9 observations and 8 variables;

data groupe22; set groupe21;
                        by strate codedel;
                        retain p90 nb2;
                        if first.strate then do;
                             p90 = 0; nb2 = 0;
                        end;
                        p90 = p90 + pop90;
                        nb2 = nb2 + nb;
                        if last.strate then output;
run;

NOTE: The data set WORK.GROUPE22 has 3 observations and 10 variables;

data groupe23; merge groupe21 groupe22;
                         by strate codedel;
                         if code LE 45 then code2 = 45;
                         else if code in (46,47,48) then code2 = 48;
                         else code2 = 51;
                         proc sort; by code2 descending code ;
run;

data groupe24; set groupe23;
                         by code2 descending code;
                         retain y;
                         if first.code2 then do;
                            y = codedel;
                         end;
                         codedel2 = y;
run;

/*                    Traitement du cas 3.                        */
/****************************************************************/

proc sort data = table2; by dep codedel; run;

data groupe31; set table2( where = (strate = 1) );
                         by dep codedel;
                         retain code 51;
```

```
                        code + 1;
run;

NOTE: The data set WORK.GROUPE31 has 13 observations and 8 variables;

data groupe32; set groupe31;
                        by dep codedel;
                        retain p90 nb2;
                        if dep = 75 then do;
                             num + 1;
                             p90 + pop90;
                             nb2 + nb;
                             if num = 2 then do;
                                   output;
                                   p90 = 0; nb2 = 0;
                             end;
                             else if num = 5 then output;
                        end;
                        else do;
                             if first.dep then do;
                                   num = 0; p90 = 0; nb2 = 0;
                             end;
                             num = num + 1;
                             p90 = p90 + pop90;
                             nb2 = nb2 + nb;
                             if num = 2 then do;
                                   output;
                                   num = 0; p90 = 0; nb2 = 0;
                             end;
                        end;
run;

NOTE: The data set WORK.GROUPE32 has 6 observations and 11 variables;

data groupe33; merge groupe31 groupe32(in = x);
                        by dep codedel;
                        if x = 1 then code2 = code;
                        else do;
                           if code = 52 then code2 = 53;
                           else if code in (54,55) then code2 = 56;
                           else code2 = code + 1;
                        end;
                        proc sort; by code2 descending code ;
run;

data groupe34; set groupe33;
                        by code2 descending code;
                        retain y;
                        if first.code2 then do;
                           y = codedel;
                        end;
                        codedel2 = y;
run;

/*               Traitement du cas 4.                          */
/* On prend la première ZD du dep, on lui rajoute le district.*/
/*************************************************************/

proc sort data = vqs.poids1
                out = toto;
                by strate d;
run;

data tata; set toto;
                  by strate d;
                  if d in (86,64,30,83,13) and
                     strate in (23,25,31,34,35);
                  if first.strate or first.d;
run;

data titi; set tata toto( where = (codedel = 'DR563156'));
                  if codedel = 'DR563052' then delete;
                  keep codedel strate;
                  proc sort; by codedel;
run;

NOTE: The data set WORK.TITI has 5 observations and 2 variables;
```

```
data titi2; merge titi(in = x) codedel;
                     by codedel;
                     if x;
                     rename NNN = nb;
                     rename tailuni = pop90;
                     drop auxniv;
run;

proc sort data = distri; by codedel distri;run;

data titi3; merge titi2(in = x) distri( keep = codedel distri NNN);
                     by codedel;
                     if x;
                     rename NNN = pop99;
                     proc sort; by strate codedel;
run;

NOTE: The data set WORK.TITI3 has 26 observations and 6 variables;

data groupe41; set table2( where = (strate in (23,25,31,34,35)) in = x)
                              titi3;
                        drop dep;
                        if x = 1 then p = 1;
                        proc sort; by strate descending p;
run;

data groupe42; set groupe41;
                     by strate descending p;
                        if first.strate or first.p then output;
run;

data groupe43; set groupe42;
                        by strate descending p;
                        retain nb2 p90;
                        if first.strate then do;
                           nb2 = 0; p90 = 0;
                        end;
                        nb2 = nb2 + nb;
                        p90 = p90 + pop90;
                        if last.strate then output;
run;

data groupe44; merge groupe41 groupe43;
                     by strate descending p;
                        proc sort; by strate p;
run;

data groupe45; set groupe44;
                     by strate p;
                  retain y;
                        if first.strate then do;
                           y = codedel;
                        end;
                        codedel2 = y;
run;

NOTE: The data set WORK.GROUPE45 has 31 observations and 11 variables;

/* On constate également que 8 districts interrogés compte   */
/* un seul répondant à VQS. On propose de les rattacher à un */
/* autre district du secteur d'agent recenseur.              */
/*************************************************************/

data groupe51; set distri( where = (codedel in ('DR683030','DR683059'
                                                'DR683099','DR083054','DR083057','DR323041'))
                                                rename = (NNN=pop99)
                                                drop = tailuni auxniv );
                        if codedel = 'DR683030' and substr(distri,9,4)= 'AC11'
                              then substr(distri,9,4)= 'AC09';
                        if codedel = 'DR683059' and substr(distri,9,4)= 'AE13'
                              then substr(distri,9,4)= 'AE14';
                        if codedel = 'DR683099' and substr(distri,9,4)= '0082'
                              then substr(distri,9,4)= '0081';
                        if codedel = 'DR083054' then do;
                              if substr(distri,9,4)= 'AN04' then substr(distri,9,4)= 'AN03';
                              if substr(distri,9,4)= 'AN05' then substr(distri,9,4)= 'AN03';
```

```
                              if substr(distri,9,4)= 'AM15' then substr(distri,9,4)= 'AM14';
                              end;
                        if codedel = 'DR083057' and substr(distri,9,4)= 'BL33'
                              then substr(distri,9,4)= 'BL30';
                        if codedel = 'DR323041' and substr(distri,9,4)= 'AE09'
                              then substr(distri,9,4)= 'AE10';

                        if distri = '76217   AC09 ' then pop99 = 294;
                        if distri = '95585   AE14 ' then pop99 = 65;
                        if distri = '95637   0081 ' then pop99 = 403;
                        if distri = '47001   AM14 ' then pop99 = 75;
                        if distri = '47001   AN03 ' then pop99 = 372;
                        if distri = '47001   BL30 ' then pop99 = 73;
                        if distri = '04070    AE10 ' then pop99 = 20;

run;

data groupe52; set codedel (where = (codedel in ('DR683030','DR683059'
                                                 'DR683099','DR083054','DR083057','DR323041')));
                        NNN = NNN - 1;
run;

proc sort data = groupe51; by distri; run;

data groupe53; set groupe51;
                        by distri;
                        if first.distri then output;
run;

/* Reconstitution du fichier géographique des districts.      */
/************************************************************/

data groupe(drop = pop99); set groupe14(keep = strate codedel2 distri pop99)
                                groupe24(keep = strate codedel2 distri pop99)
                                                groupe34(keep = strate codedel2 distri
pop99)
                                                    groupe45(keep = strate codedel2 distri
pop99)
                                                    groupe53(rename = (codedel =
codedel2));
                                        tailuni = pop99;
                                        NNN = pop99;
                                        auxniv = 3;
                                                  rename codedel2 = codedel;
run;

proc sort data = groupe43; by codedel; run;
proc sort data = groupe51; by codedel; run;

data distri2; merge distri verif(in = x drop = dep)
                              groupe43(in = w keep = codedel distri)
                                    groupe51(in = z keep = codedel distri);
                    by codedel;
                       if x = 0 and w = 0 and z = 0;
run;

data distri3; set distri2 groupe;
              proc sort; by strate codedel distri;
run;

NOTE: The data set WORK.DISTRI3 has 2265 observations and 6 variables;

/* Il se trouve aussi qu'une seule ZD a été enquetée dans la  */
/* strate 36. Donc on rattache cette ZD à la strate 34(PACA). */

data codedel2; set codedel;
                         if strate = 36 then strate = 34;
run;

data strate2; set strate;
                        if strate = 36 then delete;
                        if strate = 34 then tailuni = tailuni + 243719;
run;

data distri4; set distri3;
                        if strate = 36 then strate = 34;
run;
```

```
/* Reconstitution du fichier des zones de délégués.           */
/*************************************************************/

data groupe2; set groupe12(keep = strate codedel nb2 p90)
                  groupe22(keep = strate codedel nb2 p90)
                                groupe32(keep = strate codedel nb2 p90)
                                groupe43(keep = strate codedel nb2 p90);
                        rename nb2 = NNN;
                        rename p90 = tailuni;
                        auxniv = 2;
run;

data codedel3; merge codedel2 verif(in = x) groupe43(in = w keep = codedel) groupe52(in=z);
                      by codedel;
                         if x = 0 and w = 0 and z = 0;
                         drop distri dep;
run;

data codedel4; set codedel3 groupe2 groupe52;
                         proc sort; by strate codedel;
run;

NOTE: The data set WORK.CODEDEL3 has 352 observations and 5 variables;

/* Le fichier géographique est la concaténation des trois     */
/* fichiers créés : strate, codedel3 et distri3.              */
/*************************************************************/

data fichier.geo2; set strate2 codedel4 distri4;
run;

NOTE: There were 34 observations read from the data set WORK.STRATE;
NOTE: There were 352 observations read from the data set WORK.CODEDEL3;
NOTE: There were 2265 observations read from the data set WORK.DISTRI3;
NOTE: The data set FICHIER.GEO2 has 2651 observations and 6 variables;

proc sort data = fichier.geo2
                out = geo2;
                by strate codedel distri;
run;

proc print data = geo2;
where strate LE 2;
var strate codedel distri auxniv NNN tailuni;
title 'Fichier géographique avec tirage de districts 99 en US ';
title3 'Les strates 1 et 2 seulement';
run;

/* On isole les fichiers pour lesquels il faudra changer la ZD. */
/*************************************************************/

data fichier.changeZD; set groupe14(keep = strate codedel codedel2 distri)
                           groupe24(keep = strate codedel codedel2 distri)
                                                groupe34(keep = strate codedel codedel2 distri)
                                        groupe45(where =(p=1) keep =p strate codedel codedel2
distri );
                                        drop p;
                                        proc sort; by codedel;
run;

NOTE: The data set FICHIER.CHANGEZD has 69 observations and 4 variables;
```

Programme C:\user\odile\fichier\donnees.sas

```
/*********************************************************************/
/*       Programme de préparation du fichier des données.            */
/*********************************************************************/
/* Pour chacune des 359 010 répondants à VQS, il faut renseigner:    */
/* - la strate géographique: STRATE;                                 */
/* - la zone de délégué: CODEDEL;                                    */
/* - l'unité secondaire de tirage,ici c'est le district: DISTRI;     */
/* - la zone d'enquete pour les individus HID: ZONDENQ;              */
/* - la strate HID: STRATEH;                                         */
```

```
/* - la phase de l'enquete à laquelle il appartient: PHASE;          */
/* - la probabilité de réponse à HID: PROBAREP;                      */
/* - le taux de réponse global qui est de 0.778: REPGLOB;            */
/* - le code feuille de l'arbre qui vaut 'FA' partout: NINFFIC.      */
/*                                                                   */
/***************** Mardi 25 Juillet 2002 ****************************/
/*              c:\user\odile\fichier\données.sas                    */
/*******************************************************************/

libname fichier 'c:\user\odile\fichier';
libname joinvill 'm:\f101demo\f170eed\joinville';

options compress = yes;

/* Etape 1: injection du code feuille de l'arbre et définition des   */
/*          zones d'enquete. On prend la définition dans le program- */
/*          me tirageVQS \ tauxsondage.sas .                         */
/*******************************************************************/

data repvqs; set fichier.vqsdef2( keep = dep com frdist strate idvqsdef );
                      NINFFIC = 'FA';
             depcom = dep!!com;
             y='Z';

* région 21( Champagne, Picardie exclue mais avec Seine et Marne);
           if depcom in ('08126','08476') then zondeg='081ASF';
      else if depcom='08105' then zondeg='082CHA';

[...]

* région 94(La Corse);
     else if dep='2A' then zondeg ='201AJA';
run;

* On définit les zones d'enquete pour les communes;
* n'ayant pas participé au tirage HID;

data repvqs2; set repvqs;
              w ='999H';
              if zondeg=' ' then zondeg =dep!!w;
run;

data repvqs3; set repvqs2;
                        drop y depcom w ;
                        rename zondeg = zondenq;
                        rename strate = strateh;
run;

/* Etape 2: On récupère la zone de délégué et la strate dans le      */
/*          fichier vqszd. On définit le district en 13 caractères.  */
/*******************************************************************/

data repvqs4; merge repvqs3 fichier.vqszd(keep = idvqsdef codedel cil il0 strate);
                    by idvqsdef;
run;

data repvqs5; set repvqs4;
                        length distri $13;
                        distri = dep!!com!!cil!!il0!!frdist;
run;

/* Etape 3: On définit la variable d'appartenance à une phase de l'  */
/*          enquete en 3 modalités:                                  */
/*          - PHASE = 1 si c'est un répondant à VQS;                 */
/*          - PHASE = 2 s'il appartient à l'échantillon HID;         */
/*          - PHASE = 3 si c'est un répondant à HID.                 */
/*******************************************************************/

proc sort data = repvqs5; by idvqsdef;run;

proc sort data = fichier.hidreddf( keep = idvqsdef ident tuu9 traged sexe tailmen6 typlog3
typlog2 stratev pondrhid pondhivq poids3 pondinit pondrp pondrpa pondrpr poidsfin pondrpr3
poidscor interv numind)
           out = hidreddf;
           by idvqsdef;
run;
```

```
data rephid; set hidreddf;
                      by idvqsdef;
                      if interv = 1;
run;

data repvqs6; merge repvqs5 hidreddf(in =x) rephid(in=y);
                       by idvqsdef;
                       if x = 1 then do;
                       if y = 1 then PHASE = 3;
                              else if y = 0 then PHASE = 2;
                       end;
                       else if x = 0 then PHASE = 1;
run;

* petite vérification;
proc freq data= repvqs6; tables PHASE; run;

                           /*
                                        Cumulative    Cumulative
  PHASE    Frequency     Percent     Frequency      Percent
     1      337250       93.94        337250        93.94
     2        4815        1.34        342065        95.28
     3       16945        4.72        359010       100.00

On a bien : - 16945 répondants à HID,
                     - 16945 + 4815 = 21760 individus HID,
                     - 21760 + 337250 = 359010 répondants à VQS.

                           */

/* Etape 4: On introduit les deux variables probabilité de réponses  */
/*          à HID. La probabilité PROBAREP a été calculée par sous   */
/*          classes homogènes en réalisant une régression logistique */
/*          sur cinq variables. Le programme correspondant est enre- */
/*          gistré sous: FICHIER \ probarepHID.sas .                 */
/*********************************************************************/

proc sort data = fichier.probarep(keep = idvqsdef probarep)
                out = probarep;
                by idvqsdef;
run;

data repvqs7; merge repvqs6 probarep(in=x);
                       by idvqsdef;
                       if x = 1 then repglob = 0.778;
run;

data repvqs8; set repvqs7;
                     by idvqsdef;
                       if probarep = . then probarep = 0;
                       if repglob = . then repglob = 0;
run;

/* Etape 5: Injection des variables d'intéret ( questions HID ).     */
/*          Le calcul d'intervalles de confiance portera sur les dix */
/*          variables suivantes:                                     */
/*                                                                   */
/* - BMOB1 dans MODB_C: indique si la personne est confinée au lit,  */
/*         dans un fauteuil, ou à l'intérieur de son logement en     */
/*         d'un handicap ou d'un problème de santé. On recodera cette*/
/*         variable. -> confin1 = 1 si confinée ( BMOB1 = 1,2 ou 3 ).*/
/*                                                                   */
/* - DADAPT dans MODD: si elle dispose d'équipement spéciaux.        */
/*         -> dadapt1 = 1 si oui ( DADAPT = 1 ).                     */
/*                                                                   */
/* - C_AIDKI dans MINDIV_C: si présence d'une aide régulière.        */
/*         -> aidki1 = 1 si oui.                                     */
/*                                                                   */
/* - R-ALLOC dans MINDIV_C: si elle perçoit une allocation ou pension*/
/*         en raison de son handicap. -> alloc1 = 1 si oui.          */
/*                                                                   */
/* - R_INVAL dans MINDIV_C: si reconnaissance officielle d'un taux   */
/*         d'invalidité ou d'incapacité. -> inval1 = 1 si oui.       */
/*                                                                   */
/* - AHANDI dans MINDIV_C: si difficultés physiques, sensorielles ou */
/*         intellectuelles. -> handi1 = 1 si oui.                    */
/*                                                                   */
```

```
/* - BCOLVEZ dans MINDIV_C: indicateur de mobilité.                  */
/*         -> mob1 = 1 si confinée ( BCOLVEZ = 1 );                  */
/*         -> mod2 = 1 si besoin d'aide pour toilette/habillage;     */
/*         -> mod3 = 1 si besoin d'aide pour sortir.                 */
/*                                                                   */
/* - NBDEFIC dans MINDIV_C: indicateur du nombre de déficiences.     */
/*         -> defi1 = 1 si au moins une déficience ( NBDEFIC >= 1 ). */
/*                                                                   */
/* - RCOTOR dans MODR_C: si elle a déposé un dossier à la COTOREP.   */
/*         -> cotor1 = 1 si oui.                                     */
/*                                                                   */
/* - SLIRE, SECRIR, SCOMPT dans MODS_C: indique si la personne agée  */
/*         de plus de six ans sait lire, écrire, compter.            */
/*         -> expr1 = 1 si elle ne sait pas lire, ou écrire, ou      */
/*                       compter ( SLIRE ou SECRIR ou SCOMPT = 3 ).  */
/*                                                                   */
/* On récupère les réponses à ces questions sur le cdrom " Premier   */
/* passage auprès des personnes vivant à domicile".                  */
/*********************************************************************/

libname cdrom 'd:\hid99\fichiers\sas';

proc sort data = cdrom.modb_c ( keep = BMOB1 ident numind )
                out = modb_c;
                by ident numind;
run;

proc sort data = cdrom.modd ( keep = DADAPT ident numind )
                out = modd;
                by ident numind;
run;

proc sort data = cdrom.mindiv_c ( keep = C_AIDKI R_ALLOC R_INVAL AHANDI BCOLVEZ NBDEFIC ident
                                         numind )
                out = mindiv_c;
                by ident numind;
run;

proc sort data = cdrom.modr_c( keep = RCOTOR ident numind )
                out = modr_c;
                by ident numind;
run;

proc sort data = cdrom.mods_c ( keep = SLIRE SECRIR SCOMPT ident numind )
                out = modS_c;
                by ident numind;
run;

data reponse1; merge modb_c modd mindiv_c modr_c mods_c;
                        by ident numind;
run;

NOTE: The data set WORK.REPONSE1 has 16945 observations and 14 variables;

data reponse2; set reponse1;
                        by ident numind;
                        confin1 = ( BMOB1 in ('1','2','3') );
                        dadapt1 = ( DADAPT = '1' );
                        aidki1 = ( C_AIDKI = '1' );
                        alloc1 = ( R_ALLOC = '1' );
                        inval1 = ( R_INVAL = '1' );
                        handi1 = ( AHANDI = '1' );
                        mob1 = ( BCOLVEZ = '1' );
                        mob2 = ( BCOLVEZ = '2' );
                        mob3 = ( BCOLVEZ = '3' );
                        defi1 = ( NBDEFIC GE '01' );
                        cotor1 = ( RCOTOR = '1' );
                        expr1 = ( SLIRE = '3' or SECRIR = '3' or SCOMPT = '3' );
run;

proc sort data = reponse2; by ident numind; run;

proc sort data = repvqs8; by ident numind; run;

data repvqs9; merge repvqs8 reponse2;
                        by ident numind;
run;
```

```
NOTE: The data set WORK.REPVQS9 has 359068 observations and 60 variables;

/* Problème ! Il se trouve que pour 58 répondants à HID, les identi- */
/* fiants du ménage ou du numéro d'individu différent entre les deux */
/* fichiers. On va donc les rectifier dans le fichier REPONSE2.      */
/*********************************************************************/

data modif1; set repvqs9;
                        if idvqsdef = .;
run;

proc print data = repvqs9;
where phase = 3 and aidki1 = . ;
var idvqsdef ident numind;
run;

proc print data = repvqs9;
where idvqsdef = . ;
var idvqsdef ident numind;
run;

data modif2; set modif1;
                        if substr(ident,11,1) = '1' then substr(ident,11,1) = '0';
                        if numind = 2 then numind = 1;
                        if ident = '26012399000' then do;
                               ident = '26012499000'; numind = 1;
                               end;
                        if ident in ( '31095699000','52007699000','53021899000','53093299000',

'54015699000','54070599000','74008799000','74025299000')
                        then numind = 2;
run;

data reponse3; merge reponse2 modif1(in=x);
                          by ident numind;
                          if x = 0;
run;

data reponse4; set reponse3 modif2;
                          proc sort; by ident numind;
run;

data repvqs10; set repvqs9;
                          if idvqsdef NE . ;
                          proc sort; by ident numind;
run;

data repvqs11; merge reponse4 repvqs10;
                          by ident numind;
run;

/* Il se trouve que 167 répondants à VQS sont collectés dans quatre  */
/* districts officiellement non habités en 1999. Parmi eux, on compte*/
/* 11 individus HID dont un seul répondant. On propose de les retirer*/
/* du fichier des données pour Poulpe.                               */
/*********************************************************************/

data repvqs12; set repvqs11;
                          where idvqsdef not between 271727 and 271875
                          and idvqsdef not between 290243 and 290260;
                          proc sort; by dep com cil il0 frdist;
run;

/* Des regroupements de zones de délégués et de districts se sont    */
/* avérés neccessaires ( voir FICHIER\GEO2.sas ). Nous allons donc   */
/* apporter ces modifications dans le fichier des donnees.           */
/*********************************************************************/

data changezd; set fichier.changezd ( keep = codedel codedel2 );
                          proc sort; by codedel;
run;

proc sort data = repvqs12; by codedel; run;

data repvqs13(rename = (codedel2 = codedel) ); merge repvqs12 changezd(in=x);
                          by codedel;
```

```
                        if x = 0 then codedel2 = codedel;
                        drop codedel;
run;

data repvqs14; set repvqs13;

                                    if codedel = 'DR283114' and strate = . then do;
                                       il0 = '0004'; strate = 28; distri =
dep!!com!!cil!!il0!!frdist;
                                       end;

                                    if codedel = 'DR683030' and substr(distri,9,4)= 'AC11'
                                   then substr(distri,9,4)= 'AC09';
                                 if codedel = 'DR683059' and substr(distri,9,4)= 'AE13'
                                       then substr(distri,9,4)= 'AE14';
                                 if codedel = 'DR683099' and substr(distri,9,4)= '0082'
                                       then substr(distri,9,4)= '0081';
                                 if codedel = 'DR083054' then do;
                                       if substr(distri,9,4)= 'AN04' then
substr(distri,9,4)= 'AN03';
                                       if substr(distri,9,4)= 'AN05' then
substr(distri,9,4)= 'AN03';
                                       if substr(distri,9,4)= 'AM15' then
substr(distri,9,4)= 'AM14';
                                       end;
                                 if codedel = 'DR083057' and substr(distri,9,4)= 'BL33'
                                       then substr(distri,9,4)= 'BL30';
                                 if codedel = 'DR323041' and substr(distri,9,4)= 'AE09'
                                       then substr(distri,9,4)= 'AE10';

                                    if strate = 36 then strate = 34;
run;

/* Pour terminer, afin de ne pas fausser les calculs de propabilité  */
/* de deuxième phase, il convient d'oter du fichier des données les  */
/* ZD qui n'ont pas participé au tirage de l'échantillon HID.        */
/*********************************************************************/

proc sort data = repvqs14; by codedel; run;

data repvqs15; merge fichier.Zdhid(keep=codedel in=x) repvqs14 ;
                        by codedel;
                        if x = 1;
run;

data fichier.donnees; set repvqs15;
                                    if idvqsdef NE .;
                                    proc sort; by dep com cil il0 frdist;
run;

proc freq data = fichier.donnees;
tables phase;
run;
                                                    /*

                    The FREQ Procedure

                                          Cumulative    Cumulative
     PHASE    Frequency     Percent     Frequency      Percent
         1       308273       93.41        308273        93.41
         2         4806        1.46        313079        94.87
         3        16943        5.13        330022       100.00

                                                  */

/* Comme Poulpe n'admet pas plus de 100 strates de deuxième phase,   */
/* on ne pourra pas utiliser l'information ZONDENQ et il faudrait se */
/* limiter aux 10 strates HID. On décide de différencier l'Hérault   */
/* du reste de l'échantillon, donc on crée une variable HERAULT qui  */
/* vaut OUI pour le dep 34 et NON pour les autres.                   */
/*********************************************************************/

data fichier.donnees; set fichier.donnees;
                                    by dep com cil il0 frdist;
                                    if dep = '34' then HERAULT = 'OUI';
                                    else HERAULT = 'NON';
run;
```

```
data fichier.donnees; set fichier.donnees;
by dep com cil il0 frdist;
if PROBAREP = 0 then PROBAREP = .;
run;

data fichier.donnees; set fichier.donnees;
by dep com cil il0 frdist;
REPGLOB = 0.778;
run;

/* Suite à des problèmes d'identification d'effectifs de population, */
/* on est ammené à introduire la variable TAILUNI dans le fichier des*/
/* qui renseigne la pop90 de la zone de délégué pour le tirage PPT.  */
/* le fichier ESSAI est le nouveau fichier géographique.             */
/*********************************************************************/

proc sort data = fichier.essai(keep= auxniv codedel tailuni)
          out = toto;
                by codedel;
run;

proc sort data = fichier.donnees
                out = donnees;
                by codedel;
run;

data fichier.donnees2; merge toto(where =(auxniv = 2)) donnees(in=x);
                                by codedel;
                                      if x;
run;
```

Programme c:\user\odile\fichier\probarepHID.sas

```
/*********************************************************************/
/*          Calcul des probabilités de réponses à HID.               */
/*********************************************************************/
/* On réalise la régression logistique sur les variables suivantes:  */
/* - la taille urbaine du lieu d'habitation en 6 modalités: TUU9;    */
/* - l'age des individus, décennal, en 9 modalités:TRAGED;           */
/* - leur sexe: SEXE;                                                */
/* - le nombre de personnes du ménage en 6 modalités: TAILMEN6;      */
/* - la strate HID en 10 modalités: STRATE;                          */
/* - le type de logement(individuel, collectif, autre): TYPLOG3;     */
/* La variable de réponse à HID est INTERV qui vaut 0 ou 1.          */
/*                                                                   */
/***************** Jeudi 25 Juillet 2002 *************************/
/*          c:\user\odile\fichier\probarepHID.sas                    */
/*********************************************************************/

libname fichier 'c:\user\odile\fichier';

/* On travaille avec le fichier HIDREDFF de l'échantillon HID tiré.  */
/* On ne garde que les variables catégorielles et l'identifiant.     */
/*********************************************************************/

data table; set fichier.hidreddf(keep =
                      idvqsdef interv tuu9 traged sexe tailmen6 strate typlog3);
                      x = input(tuu9,1.);tuu = x;
                      y = input(tailmen6,1.);tailmen = y;
                      z = input(strate,2.);strate2 = z;
                      w = input(typlog3,1.);typlog = w;
run;

/* On regarde le modèle complet.                                     */
/*********************************************************************/

proc logistic data = table;
                        model interv = tuu traged sexe tailmen strate2 typlog /
                        scale = none aggregate;
                        output out = proba pred = prob;
run;

/* Le modèle est adéquat mais le sexe est non-significatif.          */
```

```
/* On regarde le modèle sans la variable sexe.                        */
/*********************************************************************/

proc logistic data = table;
                        model interv = tuu traged tailmen strate2 typlog /
                        scale = none aggregate;
run;

/* Ce deuxième modèle est meilleur que le premier.                    */
/* La tranche d'age n'étant pas très significative, on regarde le     */
/* modèle sans le sexe et la tranche d'age.                           */
/*********************************************************************/

proc logistic data = table;
                        model interv = tuu tailmen strate2 typlog /
                        scale = none aggregate;
run;

/* Le troisième modèle est moins bon que les deux deux premiers.      */
/* Finalement on calculera les probabilités de réponse à HID          */
/* partir du deuxième modèle.                                         */
/*********************************************************************/

proc logistic data = table;
                        model interv = tuu traged tailmen strate2 typlog /
                        scale = none aggregate;
                        output out = proba pred = prob;
run;

data fichier.probarep; set proba(keep = idvqsdef interv tuu9 traged
                                                              tailmen6 strate typlog3
prob);
                                      probarep = 1 - prob;
                                      drop prob;
                                      label probarep = "probabilité de réponse à HID";
run;

/* On regarde la distribution de la variable.                         */
/*********************************************************************/

proc means data=fichier.probarep;run;

data chose; set fichier.probarep;
length proba $11;
if probarep ge 0.6056567 and probarep le 0.75 then proba = 'MIN à 0.75';
else if probarep gt 0.75 and probarep lt 0.78 then proba = '0.75 à 0.78';
else if probarep ge 0.78 and probarep le 0.85 then proba = '0.78 à 0.85' ;
else if probarep gt 0.85 and probarep le 0.9092148 then proba = '0.85 au MAX';
run;

proc chart data = chose;
HBAR proba / type = percent;
run;
```

ANNEXE 4 : Références des formules de calcul des estimateurs de variance utilisées par le logiciel Poulpe.

Pour calculer les estimateurs de variance, Poulpe part de la dernière étape de tirage et remonte successivement jusqu'à la racine en appliquant le principe de Raj. Il applique une formule d'grégation spécifique à chaque type de tirage.

Les formules relatives aux étapes élémentaires de la première phase sont les suivantes. Dans le cas de HID, il y avait trois types de tirage : le sondage aléatoire simple, le tirage à probabilité proportionnelle à la taille et la stratification.
π_k représente la probabilité d'inclusion, y_k représente la valeur de la variable Y pour l'individu k . S'il ne s'agit pas d'un individu mais d'une unité de tirage (exemple le district), pour appliquer le principe de Raj, la formule utilise $y_k = \hat{t}_k$ le total de la variable Y sur l'unité de tirage k.

Sondage aléatoire simple sans remise (référencé : SAS)

Estimateur de la somme :

$$\hat{y} = \frac{N}{n}\sum_s y_k$$

Estimateur de la variance :

$$\hat{V} = \frac{N^2}{n}(1-\frac{n}{N})\frac{\sum_s (y_k - \bar{y})^2}{n-1} + \frac{N}{n}\sum_s V_k$$

avec : $y_k = \hat{t}_k$ et $\pi_i = n/N$

et : $V_k = 0$ pour les feuilles.

Tirage à probabilités inégales (dont les sondages à probabilités proportionnelles à la taille) (référencé : PPT)

Estimateur de la somme :

$$\hat{t} = \sum_s \frac{y_k}{\pi_k}$$

Estimateur de la variance :

$$\hat{V} = \frac{n}{n-1}\sum_k (1-\pi_k)(\frac{y_k}{\pi_k} - \sum_s a_k \frac{y_k}{\pi_k})^2 + \sum_k \frac{\hat{V}_k}{\pi_k}$$

avec : $a_k = \frac{(1-\pi_k)}{\sum_s (1-\pi_k)}$

π_i = *n*Σ(Taille des entités tirables)/(Taille de l'entité tirée)*

et : $V_k = 0$ pour les feuilles.

Stratification (référencé : EXH)

Estimateur de la somme : *Estimateur de la variance :*

$\hat{t} = \sum_s y_k$ $\hat{V} = \sum_s \hat{V}_k$

avec : $y_k = \hat{t}_k$

Pour calculer la **variance sur les trois phases de l'enquête**, le logiciel utilise une formule globale. π_i est la probabilité d'inclusion de la première phase, p_i celle de la deuxième phase (stratifiée) et q_i celle de la troisième phase (tirage poissonnien). La variance est la somme de cinq termes.

Nature	*Référence fichier RES*	*Formules*	*Observations*
Somme	H	$\hat{Y} = \sum_{i \in s_3} \frac{y_i}{\pi_i p_i q_i}$	
Variance, 1er terme	R	$\hat{R} = \sum_{i,j \in s_1} A_{ij} \frac{y_i}{p_i q_i} \frac{y_j}{p_j q_j} + \sum_{h=1}^{h=H} \sum_{i \in s_1} \sum_{j \in s_1} A_{ij} \frac{y_i}{p_i q_i} \frac{y_j}{p_j q_j} \frac{1-f_h}{n_h - 1}$	$A_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j}$ π_{ij} est la probabilité d'inclusion double.
Variance, 2ème terme	J	$\hat{J} = \sum_{h=1}^{h=H} \sum_{i \in s_3 \cap h} A_{ii} y_i^2 \frac{1}{f_h} \left[\frac{1}{q_i} - \frac{1}{f_h q_i^2} \left(\frac{n_h - f_h}{n_h - 1} \right) \right]$	$A_{ii} = \frac{1-\pi_i}{\pi_i^2}$
Variance, 3ème terme	K	$\hat{K} = \sum_{h=1}^{h=H} N_h^2 \frac{1-f_h}{n_h} \left[\frac{1}{n_h - 1} \sum_{i \in s_{2h}} (t_i - \bar{t}_h)^2 \right]$	$t_i = \frac{y_i}{q_i \pi_i}$ si i ∈ s sinon 0 $\bar{t}_h = \frac{1}{n_h} \sum_{i \in s_{2h}} \frac{y_i}{\pi_i q_i}$
Variance, 4ème terme	L	$\hat{L} = \sum_{i \in s3} B_{ii} \frac{y_i^2}{\pi_i^2} \left[\frac{q_i - 1}{q_i^2} \right]$	$B_{ii} = \frac{1-p_i}{p_i^2}$
Variance, 5ème terme	M	$\hat{M} = \sum_{i \in s3} \frac{y_i^2}{\pi_i^2 p_i^2} \left[\frac{1-q_i}{q_i^2} \right]$	

Pour calculer l'effet de sondage, Poulpe utilise la formule suivante, π_k* étant la probabilité d'inclusion finale du répondant HID k. Cette formule permet d'avoir une estimation de la variance que l'on aurait obtenue si l'échantillon avait été tiré par sondage aléatoire simple, le poids final étant le même que celui de l'enquête.

$$\hat{V}_{sas}(\hat{Y}) = \hat{N}\frac{(1-\frac{n}{\hat{N}})}{n}\sum_k \frac{1}{\pi_k^*}(y_k - \bar{y})^2$$

$$\hat{N} = \sum_k \frac{1}{\pi_k^*} \text{ et } \quad \bar{y} = \frac{\sum_k \frac{y_k}{\pi_k^*}}{\sum_k \frac{1}{\pi_k^*}}$$