Mise en oeuvre du logiciel POULPE pour estimer la précision de l'Enquête HID

Pascal ARDILLY, *INSEE, Unité de Méthodologie Statistique*
Odile JOINVILLE, *Institut de Statistique de l'Université de Paris*
Pierre MORMICHE, *INSEE, Enquêtes et Etudes Démographiques*

PLAN

1. **Poulpe : présentation du logiciel**
2. **Faire entrer HID dans le moule Poulpe**
3. **Poulpe : l'aide à la correction des erreurs**
4. **Du logiciel à un outillage diffusable aux utilisateurs de l'enquête**

1. Présentation du logiciel

Poulpe permet l'estimation

- **de variances,**
- **d'intervalles de confiance**
- **et d'effets de sondage**

dans les plans complexes

Un logiciel puissant…qui traite :

- ***Tirage à probabilités inégales*** **: (formule proposée par JC Deville)**

$$\hat{V} = \frac{n}{n-1} \sum_{k \in s} (1 - \Pi_k) \left(\frac{y_k}{\Pi_k} - \sum_{k \in s} a_k \frac{y_k}{\Pi_k} \right)^2$$

où

$$a_k = \frac{1 - \Pi_k}{\sum_{k \in s} (1 - \Pi_k)}$$

Tirage à plusieurs degrés : (formule de Des Raj)

$$\hat{V} = f\left(\hat{Y}_i \middle| i \in U\right) + \sum_{i \in U} w_{is} \hat{V}_i$$

où

- $\sum_{i \in U} w_{is} \hat{Y}_i$ **estime le vrai *Y* sans biais**
- $f\left(Y_i \middle| i \in U\right) = \hat{V}_1\left(\sum_{i \in U} w_{is} Y_i\right)$
- **- les Y_i , sont inconnus –**
- $\hat{V}_i = \hat{V}_{2/1}\left(\hat{Y}_i\right)$

- ***Tirage en 2 phases : (cas post-stratifié) :***

$$\hat{V} = \sum_{h=1}^{H} \frac{1-\frac{n_{h,2}}{n_{h,1}}}{n_{h,2}} \cdot n_{h,1}^2 \cdot \frac{1}{n_{h,2}-1} \sum_{k \in S_{h,2}} (z_k - \bar{z}_h)^2$$

où $z_k = y_k / \Pi_k$

et $\bar{z}_h = \frac{1}{n_{h,2}} \sum_{k \in S_{h,2}} z_k$

- ***Tirage de Poisson : (modélisation de la non réponse finale)***

$$\hat{V} = \sum_{k \in s} \frac{1 - \theta_k}{\theta_k} \cdot y_k^2$$

Poulpe utilise TROIS fichiers essentiels

- **Description du plan de sondage (unités d'échantillonnage et méthode de tirage)**
- **Liste des unités d'échantillonnage et informations nécessaires au calcul des Π_k**
- **Fichier des données individuelles**

CINQ étapes fondamentales

- **enrichissement et contrôle de l'arbre décrivant le plan de sondage (ARBGEN)**
- **calcul des probabilités d'inclusion « locales » (CALPII)**
- **chargement de la liste des variables d'intérêt (CHARLIS)**
- **estimations de variance des totaux des variables d'intérêt (ESTIVAR)**
- **estimations de variance pour des statistiques complexes (ratios, ρ, ...) (ESTIFON)**

Atouts de Poulpe

- **Universalité (…ou presque !)**
- **Recalcul des pondérations**
- **Traitement d'estimateurs complexes et prise en compte de redressements**
- **Documentation très complète**

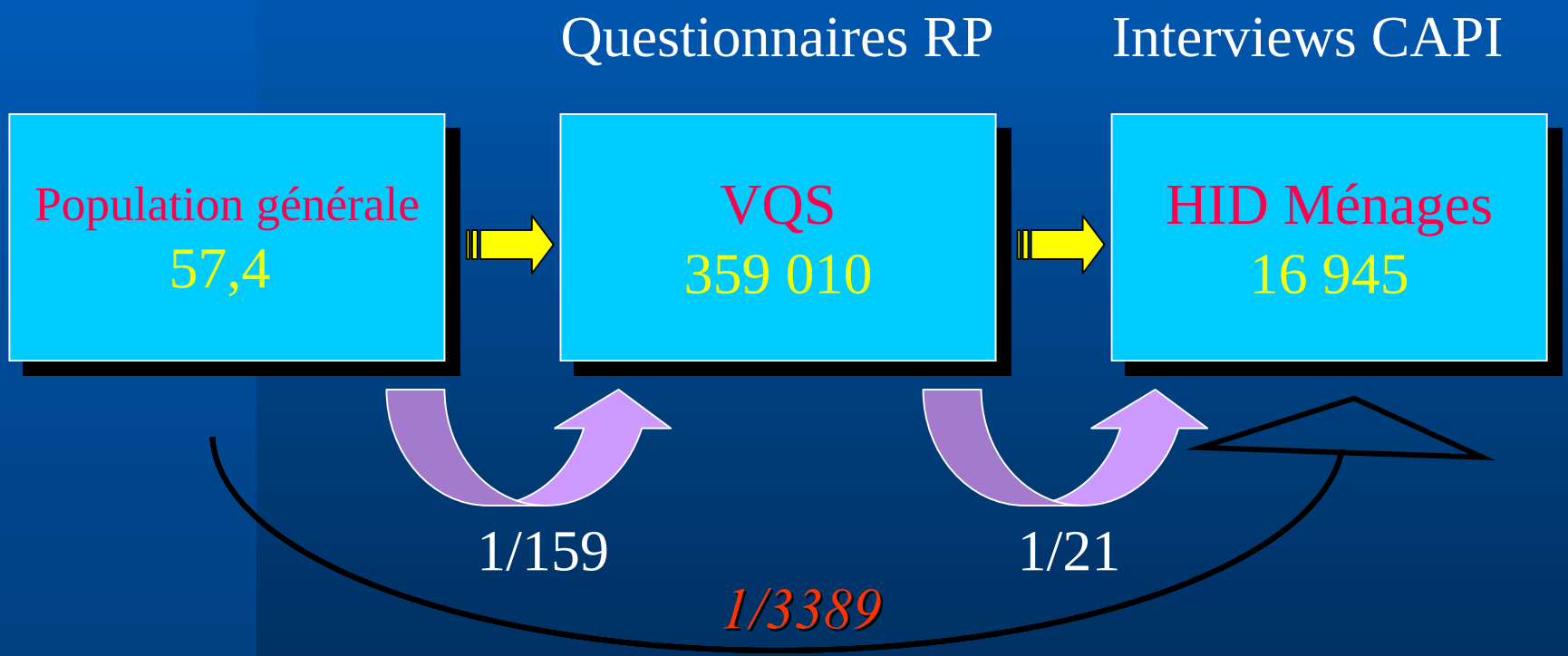
Inconvénients de Poulpe

- **Lourdeur du travail préparatoire**
- **Manque de convivialité**
- **Pas de traitement des cas n=1**
- **Problème avec de très gros échantillons**

2.1 LE PLAN DE SONDAGE (1-Ensemble)

- **Une phase de filtrage : l'enquête VQS**
  - 417 500 personnes interrogées (359 000 réponses)
  - Questionnaire court lié au RP *(dépôt-retrait)*
- **Puis l'extraction de l'échantillon HID**
  - Probabilités de tirages inégales (de 1 à 100) pour :
    - sur représenter les personnes HANDICAPEES
    - et garder un groupe témoin
  - 21 760 individus
- **Enfin les non-réponses (**16 943**)**

ECHANTILLONNAGE



2.1 LE PLAN DE SONDAGE (2-VQS)

Les 4 étapes de VQS :

1. 36 strates géographiques → *pour tenir compte des régions et des extensions locales*
2. 391 zones de délégués tirage PPT
3. 763 secteurs d'agents recenseurs tirage SAS
4. Taux de non-réponse moyen de 14 % → *variable d'un SAR à l'autre*

Donc :

- un sondage aréolaire avec des effets de grappe
- un taux moyen de 1/160 ème

2.1 LE PLAN DE SONDAGE (3-HID)

Une post-stratification complexe :

1. 6 « groupes VQS » → *indiquant la sévérité du handicap sur la base des réponses à VQS*
2. Croisés par l'âge → **10** « strates HID »
3. Dans chacune des 366 « zones d'enquête »

Donc :

- un tirage à probabilités très fortement inégales
- un taux moyen de 1/21 ème

TIRAGE DE L'ECHANTILLON HID

- Un choix : le non filtrage

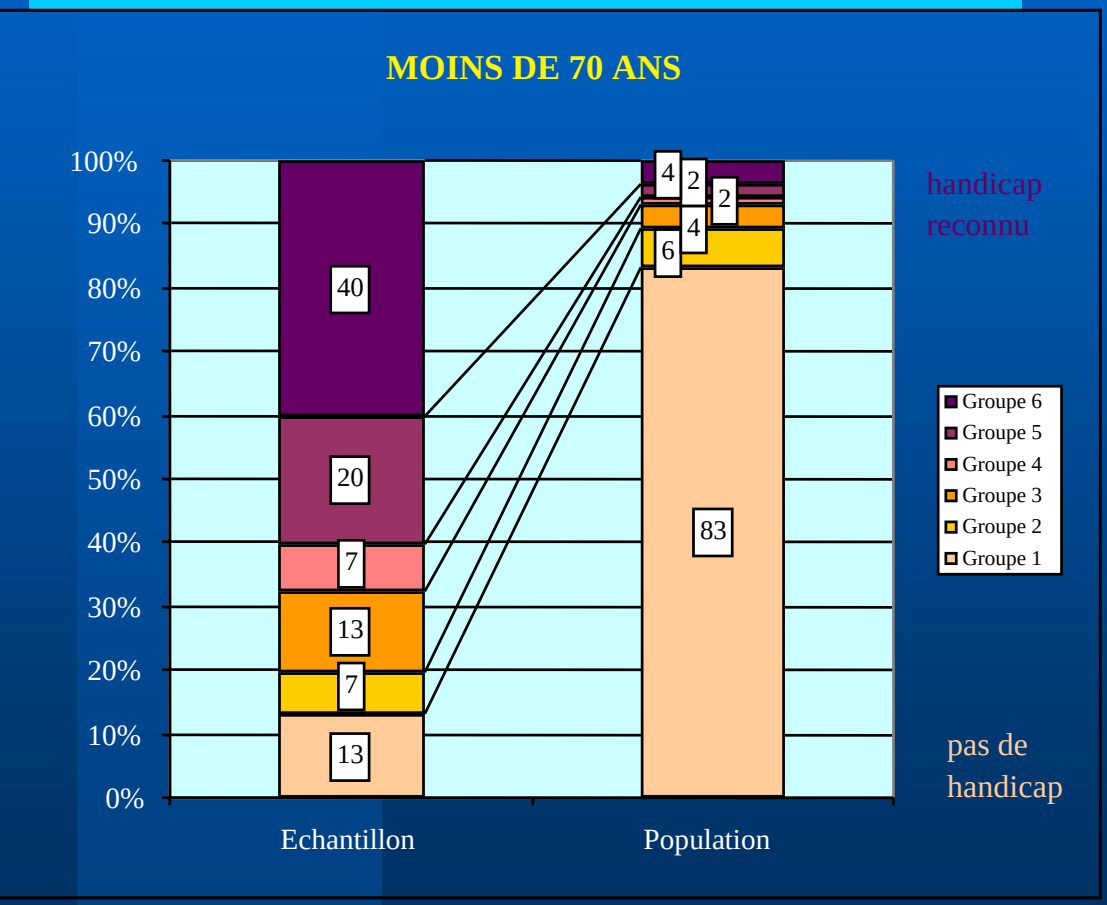
- Base de sondage = échantillon VQS
- Sondage stratifié : 10 strates croisement de 6 groupes VQS et de 2 classes d'âge
- Probabilités de tirage inégales (1à 100)

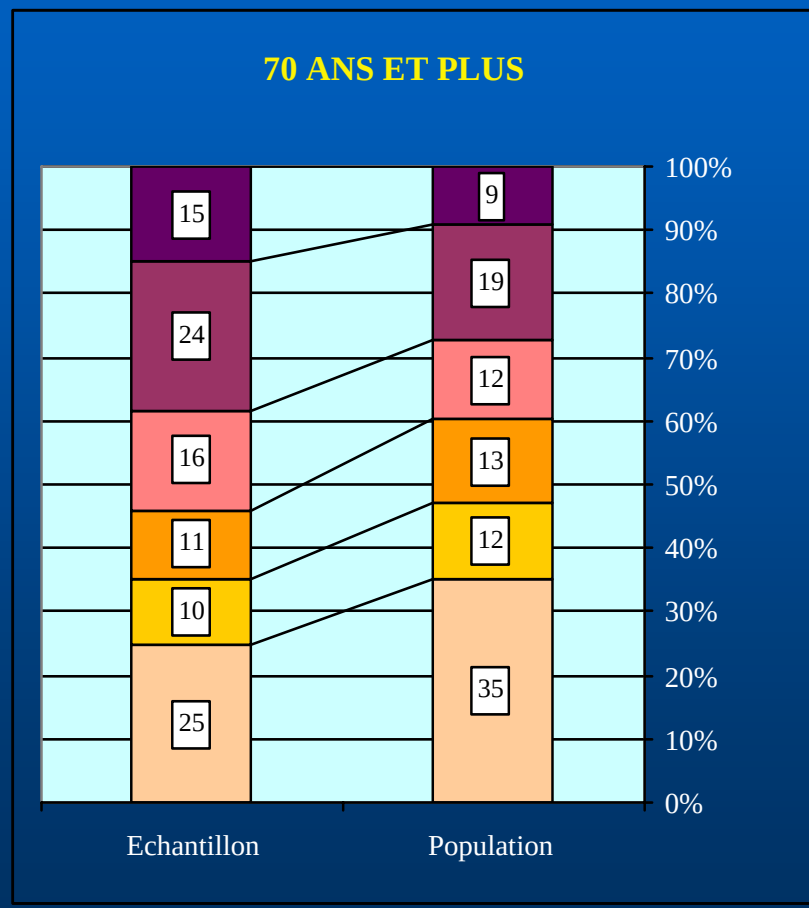
- Sur-représentation croissante avec le groupe VQS
- Parmi les VQS négatifs sur-représentation des plus de 70 ans

Coefficients de tirage

	<70 ans	≥70 ans
Groupe 1	0,6875	18,75
Groupe 2	5,625	20,25
Groupe 3	20	20
Groupe 4	30	30
Groupe 5	52	32
Groupe 6	65	40

Caractéristiques de l'échantillon HID





2.2 De la théorie à la pratique

quelques os…

Aucune trace des secteurs d'agents recenseurs n'a été conservée dans les fichiers RP ou VQS

- *On a donc décrit à la place un tirage de districts*
  - Plus nombreux (tirage plus précis)
  - De tailles très inégales (variance plus élevée)

Le logiciel admet au maximum 99 post-strates au lieu des 366 * 10 utilisées dans le tirage HID

- *Il faut donc décrire un tirage dans des zones géographiques regroupant « convenablement » les 366 zones d'enquête*

2.2 De la théorie à la pratique

enfin des résultats !

Variable d'intérêt	Estimateur du total (poids réels)	écart type de l'estimateur HT	coefficient de variation (en %)
DEFI1	22 226 953 -	469 016 -	2,1
AIDKI1	5 017 660 -	152 539 -	3,0
INVAL1	3 483 114 -	290 417 -	8,3
ALLOC1	2 237 528 -	175 058 -	7,8
CONFIN1	581 966 -	62 158 -	10,7
MOB1	119 183 -	15 075 -	12,6

3. Problèmes et solutions (1)

Quelle est la qualité des estimations de variance ?

- *Le logiciel fournit des estimations de totaux ou moyennes et a permis de déceler 3 erreurs :*
  1. L'utilisation des districts à la place des secteurs
  2. Des post-strates HID trop vastes et hétérogènes
  3. Enfin une erreur d'appariement perturbant très fortement les estimations
- *Au terme de ces corrections, on obtient des totaux exacts et des variances 2 à 10 fois moindres*

3. Problèmes et solutions (2)

L'effet des corrections sur les estimations de totaux

Variable d'intérêt	Estimateur du total (poids réels)	Estimation initiale de Poulpe	Estimation Poulpe après correction 1	Estimation Poulpe après correction 2	Estimation Poulpe après correction 3
ALLOC1	2 237 528	5 199 136	3 274 289	3 379 854	2 197 701
DEFI1	22 226 953	37 135 469	23 581 033	23 581 033	22 138 782
HANDI1	17 976 079	31 621 488	19 481 560	19 869 451	18 094 765
INVAL1	3 483 114	7 486 118	4 805 683	4 958 638	3 449 006

3. Problèmes et solutions (3)

L'effet des corrections sur les estimations de variance

Variable d'intérêt	Coefficient de variation initial	CV après correction 1	CV après correction 2	CV après correction 3
ALLOC1	7,1	4,7	3,2	2,7
DEFI1	1,5	1,6	n.d.	1,3
HANDI1	4,0	2,4	1,9	1,8
INVAL1	8,9	5,2	3,7	1,7

4. Du logiciel à un outillage spécifique-1

La première partie de Poulpe *(description du plan de sondage et calcul des PI)* doit être mise au point une fois pour toutes *pour :*

1. *éviter des descriptions variant selon les études*
2. *économiser du temps de calcul*

Poulpe le permet

4. Du logiciel à un outillage spécifique-2

La seconde partie *(calcul des estimations de variance)* d'usage complexe, doit être simplifiée

- *Le logiciel produit des appels de macros qu'on peut récupérer*

En somme, Poulpe est capable d'absorber des plans très divers, et de générer des applications adaptées à chaque enquête

En guise de conclusion

1. Poulpe absorbe et traite des plans très complexes
2. Surtout conserver toute information utilisée dans le tirage et éviter des tirages de taille « 1 »
3. Les estimations de totaux sont un bon outil d'évaluation de la pertinence de la description du plan
4. L'absorption de gros échantillons est coûteuse

Au total, Poulpe permet de comparer des plans divers, de choisir la meilleure description, et de générer une application adaptée à l'enquête traitée