Méthodologie de l'enquête emploi

Formation : Mathématiques et sciences sociales

Eric Lesage

Crest-Ensai

29 septembre 2011

1. Enquête Emploi
2. Échantillonnage
3. Redressement

L'enquête emploi

- L'enquête Emploi est l'une des pièces centrales du dispositif statistique de connaissance de l'emploi et du chômage
- Elle est la seule source permettant de mettre en oeuvre la définition du chômage au sens du Bureau International du Travail

Objectifs de l'enquête emploi

1. Production de **statistiques descriptives**
   - le site de l'Insee
   - besoin d'un échantillon avec les poids d'enquête
   - le nombre de chômeurs, d'actifs, le taux de chômage au sens du BIT
2. Estimation de paramètres de modèles statistiques
   - *Une décomposition du non-emploi en France*, Laroque et Salanié (2000), Eco et Stat
   - besoin d'un échantillon avec les informations explicatives du plan de sondage

L'échantillon de l'enquête emploi

- L'enquête emploi est une enquête trimestrielle
- La collecte est effectuée auprès d'un échantillon de 57 000 ménages répondants, soit 108 000 personnes de 15 ans et plus
- La collecte se fait en partie par interview en vis-à-vis (contraintes de collecte)
- Le taux de réponse est de 85%

Les défis méthodologiques de l'Enquête emploi

- Produire des estimateurs précis mais simples à calculer
- Corriger le biais causé par les non-répondants
- Calculer la précision des estimateurs

1 Enquête Emploi

2 Échantillonnage

3 Redressement

Plan de sondage

L'échantillon doit :

- être sélectionné au hasard (échantillon probabiliste)
- être « représentatif » de la population française (produire des estimateurs faiblement biaisés et les plus précis possibles)
- permettre un calcul de précision des estimations
- coûter le moins cher possible (respect de contraintes de collecte ou financières)

Le plan de sondage est une probabilité

- On étudie une population finie $U = \{1, \ldots, N\}$
- On considère la **probabilité** $p(s)$ sur l'ensemble $\mathcal{S} \subset \mathcal{P}(U)$ des échantillons possibles (plan de sondage)
- Un algorithme de tirage permet de sélection un échantillon s_0 en respectant cette probabilité
- L'important est que $\forall k \in U$, la probabilité de sélection (ou d'inclusion dans s) de l'individu k, noté π_k, soit non nulle (le π_k ne sont pas nécessairement égales)

Le sondage aléatoire simple de taille n

- $\mathcal{S}$ est l'ensemble des échantillons de taille n ($Card(\mathcal{S}) = C_N^n$)
- $p(s) = cte = \frac{1}{C_N^n}$ (équiprobable)
- C'est un plan de base qui requiert peu d'information auxiliaire
- La probabilité d'inclusion $\pi_k = \frac{n}{N}$
- Il existe de nombreux algorithmes pour tirer un échantillon suivant ce plan de sondage

Le paramètre d'intérêt en sondage

Le paramètre d'intérêt est une statistique descriptive :

$$t_y = \sum_{k \in U} y_k$$

ou

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k$$

Exemple : nombre de chômeurs et taux d'activité de la population

Le principe d'estimation original du sondage

Estimateurs dits « linéaires »

$$\hat{t}_y = \sum_{k \in s} w_k(s) y_k$$

où $w_k(s)$ est le **poids d'enquête** associé à l'individu k

L'estimateur Horvitz-Thompson (HT)

$$\hat{t}_{y,\pi} = \sum_{k \in s} \frac{1}{\pi_k} y_k$$

- C'est l'estimateur de base. De nombreux estimateurs en sont dérivés
- Avantage des poids non aléatoires et de l'absence de biais (par construction)
- Utilisé à l'étape de l'échantillonnage pour prévoir la précision et optimiser le plan

L'information auxiliaire

- Pour construire un plan de sondage complexe afin de diminuer les coûts ou améliorer la précision des estimateurs
- Information connue $\mathbf{x}'_k = (x_{k,1}, \ldots, x_{k,p})$, pour tout k de U

k	$\mathbf{x}_1$	$\ldots$	$\mathbf{x}_P$
1	$x_{1,1}$	$\ldots$	$x_{1,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$x_{k,1}$	$\ldots$	$x_{k,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$x_{N,p}$	$\ldots$	$x_{N,p}$

Le plan de sondage complexe de l'enquête emploi

C'est un plan :

- à 5 degrés (secteurs, aires géographiques, ménages, individus, répondants)
- stratifié et équilibré au 1er degré (choix des secteurs)

Un secteur regroupe 6 aires géographiques d'environ 20 résidences principales

Les variables d'équilibrage de l'enquête emploi au niveau des secteurs

- La répartition des résidences selon le zonage en aires urbaines
- La répartition des ménages par quintile de revenus
- La répartition des résidences suivant leur ancienneté
- Le nombre de locataires
- Le nombre de Logements sociaux
- Le nombre de Logements collectifs
- Le nombre de ménage dont le chef de ménage à plus de 55 ans
- Le nombre de résidences principales

L'échantillonnage équilibré

La motivation, l'objectif :

- Trouver un plan $p(s)$ « meilleur » que le plan aléatoire simple grâce à l'utilisation de l'information auxiliaire
- Réduire $\mathcal{S}$ en un ensemble $\widetilde{\mathcal{S}}$ d'échantillons qui restituent des caractéristiques identiques à la population totale (estimation exacte des totaux connus)
- Exemple : part des femmes et part des hommes

L'échantillonnage équilibré

Retour à un cas général

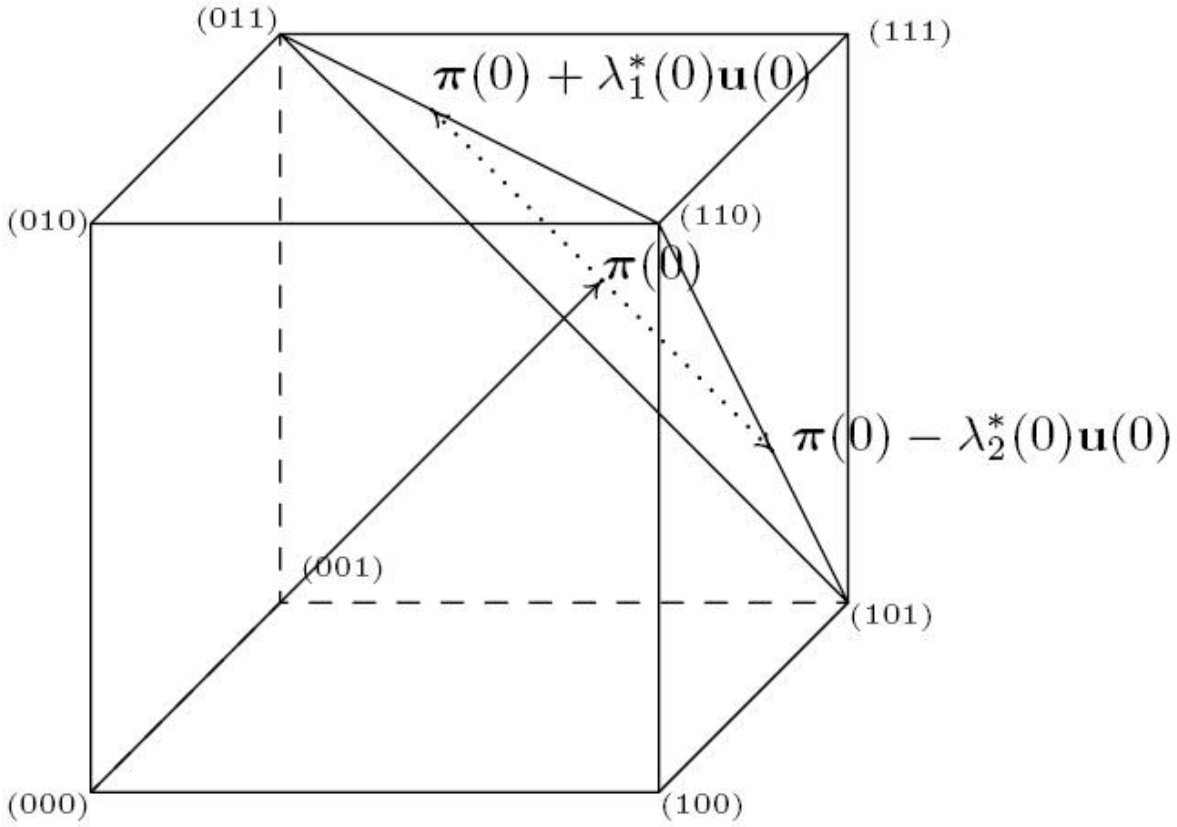
- $\mathbf{x}'_k = (x_{k,1}, \ldots, x_{k,p})$ information auxiliaire connue sur U
- $\pi' = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)$, vecteur des probabilités d'inclusion des individus (fixé)
- On veut un tirage tel que :
  - $\widetilde{\mathcal{S}} = \{s \in \mathcal{S}, \hat{t}_{\mathbf{x},\pi} = t_{\mathbf{x}}\}$ (échantillons équilibrés)
  - $p(s)$ vérifie $\sum_{s \in \mathcal{S}'} p(s) S = \pi$
    où le vecteur aléatoire
    $\mathbf{S}' = (S_1, \ldots, S_N) = \left(I_{[1 \in s]}, \ldots, I_{[N \in s]}\right)$
- il existe plusieurs méthodes pour trouver des plans de sondage équilibrés (cf méthode exhaustive)

La méthode du cube (Deville et Tillé, 2004)

- Interprétation géométrique du problème
- $\mathbf{S} \in \{0,1\}^N$ est un sommet du N-cube $C = [0,1]^N$ de R^N
- les équations d'équilibrage $\sum_{k \in U} \frac{\mathbf{x}_k}{\pi_k} S_k = \sum_{k \in U} \mathbf{x}_k$ définissent un sous espace affine de R^N, noté $\mathcal{Q}$, tel que $\mathcal{Q} = \pi + Ker\mathbf{A}$ où $\mathbf{A} = [x_{p,k}]_{p \in [1,P], k \in [1,N]}$
- La méthode du cube consiste en un algorithme qui mène à un sommet de l'intersection $\mathcal{Q} \cap C$ (martingale)

Une illustration de la méthode du cube

A chaque tour on touche une face



La méthode du cube

Processus $\pi(t)$

$\pi(0) = \pi$, ... $\pi(N) = \mathbf{S}$

1 Enquête Emploi

2 Échantillonnage

3 Redressement

Redressement de l'erreur d'échantillonnage

- Du fait de l'aléa, un échantillon est plus ou moins équilibré
- $\sum_{k \in s} \mathbf{x}_k \frac{1}{\pi_k} \neq \sum_{k \in U} \mathbf{x}_k$
- Une information auxiliaire (non présente au moment de la collecte) permet de détecter ce déséquilibre et de le « corriger »
- Exemple : disproportion homme-femme

Estimateur avec la propriété de calage (Estimateur calé)

$\mathbf{x}_k' = (x_{k,1}, \ldots, x_{k,p})$ information auxiliaire connue sur la population U

Un estimateur linéaire $\hat{t}_y = \sum_{k \in s} w_k(s) y_k$ a la propriété de calage si

$$\hat{t}_{\mathbf{x}} = \sum w_k(s) \mathbf{x}_k = t_{\mathbf{x}}$$

Les estimateurs par calage (Deville et Särndal, 1992)

- Une méthode simple mais générale pour produire des estimateurs calés
- Remplacer les poids d'enquête par un nouveau jeu de poids de calage $w_k(s)$ obtenu par la résolution du programme d'optimisation :
  - $\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} d(w_k, d_k)$
    sous contrainte : $\sum_{k\in s} w_k \mathbf{x}_k = t_{\mathbf{x}}$
  - $d(\cdot,\cdot)$ est une pseudo-distance

Les estimateurs par calage

- Le programme se résout à l'aide d'un Lagrangien
- Si la distance utilisée est la distance du χ^2 (i.e. $d(w_k, d_k) = \frac{1}{2}\frac{(w_k - d_k)^2}{d_k}$), on trouve comme solution : $w_k = d_k(1 + \mathbf{x}_k^{'}\lambda)$ (où λ est le vecteur de taille P des multiplicateurs de Lagrange)

La non-réponse

- Le déséquilibre de l'échantillon provient également de la non-réponse
- l'enquête emploi a un taux de non-réponse de 15%
- La non-réponse est traitée comme un dernier degré de tirage (poissonnien conditionnel de taille fixe) ...
- ... Ce qui nécessite d'estimer ρ_k, la probabilité de répondre de l'individu k
- plusieurs méthodes (modèles logistiques...)

L'approche fonctionnelle du calage :

- Modèle de non-réponse : $\rho_k = \frac{1}{f(\mathbf{x}_k'\beta)}$
- L'approche fonctionnelle (Estevao-Särndal) du calage pour prendre en compte la non-réponse : $w_k = f(\mathbf{x}_k'\lambda)d_k$
- $\lambda \to \beta$
- Argument asymptotique : si le modèle de non-réponse est vrai, asymptotiquement, $\sum_{k\in s} f(\mathbf{x}_k'\beta)d_k\mathbf{x}_k \approx \sum_{k\in U}\mathbf{x}_k$ (consistence de l'estimateur HT)

Variables de calage de l'enquête emploi

- Situation du logement (résidence principale, secondaire, vacante)
- Statut d'occupation (locataire, propriétaire)
- Type de logement (individuel, collectif)
- Nombre de pièces
- HLM ou non
- Décile de revenu du ménage
- Zonage en aires urbaines
- Age quinquennal, le sexe et la région
- Nombre de logements neufs

Enquête sur les non-répondants

- Le modèle de non-réponse doit tenir compte du statut d'activité
- D'où un ajout dans le calage du nombre d'actifs dans le ménage
- Utiliser le calage généralisé (Kott) $w_k = f(\mathbf{z}_k'\lambda)d_k$

- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey samling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : the cube method. *Biometrika*, 91, 893–912.
- (2011). Enquête Emploi en continu. *Source et méthode*.