Révision du plan de sondage pour l'enquête suisse sur la structure des salaires

Lionel Qualité

Office Fédéral de la Statistique / Université de Neuchâtel

10^{me} Colloque francophone sur les sondages | 26 octobre 2018

Enquête suisse sur la structure des salaires (LSE - Lohnstrukturerhebung)

- Collecte des informations sur les salaires, les qualifications, le taux d'occupation, la classe d'activité, etc.
- Enquête par échantillon, tous les deux ans.
- Produit des résultats sur le coût de la main d'oeuvre, la distribution des salaires.
- Données de base pour le "calculateur de salaires - Salarium".

Plan, jusqu'en 2016 - 1

- "Stratifié" par classe de taille, catégorie d'activité de l'entreprise et zone géographique (tirage Poisson depuis 2012).
- Deux degrés (les ets. livrent les données pour une partie du personnel).
- Taille d'échantillon déterminée par le budget.
- Lien pas évident avec des objectifs de précision.

Plan, jusqu'en 2016 - 2

- Allocation "à la Neyman" pour la moyenne des salaires.
- Interdépendances entre échantillon OFS et extensions cantonales.
- Tailles minimales imposées dans les "strates" (mais regroupements a posteriori).
- Paramètres estimés sur enquêtes précédentes.

Nouveauté pour 2018

- Revenus livrés par la Caisse de Compensation.
- Exhaustif mais
  1. disponibles avec deux ans de retard,
  2. sans information sur le taux d'occupation,
  3. parfois au niveau de la "tête de groupe".

Utilisation

- Comme variable proxy des salaires.
- Pour calculer taille et allocation en fonction des objectifs de précision fixés.
- *(et choisir/adapter ces objectifs en fonction du résultat)*
- Et calculer les extensions en supplément de l'échantillon OFS.
- Sans limitations ou appréhension liée à l'enquête précédente.

Plan de travail

- Allocation en fonction de nombreux objectifs de précision.
- Allocation pour estimer des médianes.
- "Validation" en utilisant les données de l'enquête 2016.

"Validation" -1

- L'occasion de réviser les estimateurs de variance de la LSE (plan de Poisson, calage, deux degrés).
- Deux calculs :
  1. Quelle variance estimée pour la LSE 2016 si l'on avait relevé des revenus AVS plutôt que des salaires ? (pour valider l'idée du proxy)
  2. Quelle variance estimée en utilisant les données de la LSE 2016 si l'on avait utilisé d'autres taux de sondage au premier et aux deuxième degré ? (pour "contrôler" que le nouveau plan a bien la précision attendue).

"Validation" -2

- Remarque : si l'on change le plan du deuxième degré, l'estimateur de la variance est la somme de 3 termes au lieu des deux termes habituels.
- Relativement simple (et estimable) avec un tirage de Poisson.
- Résultats : compatibles avec la confiance que l'on a dans l'estimation de variance, c'est à dire pas identiques mais relativement semblables.

Estimation de médianes - linéarisation

- Démarche : calculer une variable linéarisée et se ramener à une allocation pour l'estimation d'un total.
- Deville (1999) puis Osier (2009) pour le quantile q_α d'ordre α de la variable y_k :

$$z_k = -\frac{1}{Nf(q_\alpha)}\left[I(y_k \leq q_\alpha) - \alpha\right].$$

- Tillé et Vallée (2017) d'après Graf (2015) :

$$z_k = -\frac{1}{Nf(q_\alpha)}\left[\Phi\left(\frac{q_\alpha - y_k}{h}\right) - F(q_\alpha)\right],$$

où $f(u) = \frac{1}{hN}\sum_{k \in U}\phi\left(\frac{u-y_k}{h}\right)$, ϕ est un noyau, h la fenêtre, $\Phi' = \phi$, $F' = f$.

Simulations - observations

- Plan et taille de la LSE trop grand pour des simulations (cf. plus loin) $\rightarrow$ *petits* échantillons SRS, avec ou sans calage.
- Choix de h : attention aux valeurs extrêmes (ex: remplacer σ par $min\{\sigma, (q_{.75} - q_{.25})/1.34\}$ dans les formules usuelles.)
- Choix de la linéarisée correspond à un choix de l'estimateur de q_α ! (sinon biais, taux de couverture incorrect, etc).
- Médiane de la "proc surveymeans" (SAS), montre un biais et une variance plus grande que les estimateurs "à noyaux".

Exemple

Méthode	Mediane	$\mathrm{E}(\widehat{q}_{.5})$	S calculé	$\mathrm{E}(\widehat{S})$	E.T. simul.	Tx. Couv.
m0	55'948	55'974	350.92	351.08	347.80	0.9473
m1	55'958	55'963	314.88	315.60	317.72	0.9470
m2	55'956	55'959	315.86	316.50	315.40	0.9483
m3	55'958	55'960	314.88	315.41	314.81	0.9485
m4	55'956	55'951	315.85	316.78	317.77	0.9495

Table: Résultats de 10'000 simulations, échantillon simple de taille 5'000, poids calés.

Allocation - 1

- Taux de sélection uniforme π_h dans "strates" h, croisements de classe de taille, activité et zone géographique (idem 2016).
- Variance de la forme

$$V = \sum_{h \in H} \frac{d_h^2}{n_h} - f_h,$$

  où $n_h = N_h \pi_h$.
- Les n_h ne sont pas nécessairement entiers, ni écartés de 0.
- d_h et f_h fonction des données, des taux de réponse prévus, du taux de sondage au deuxième degré, du calage prévu, etc.

Allocation - 2

- Optimale pour un total, sous contrainte de coût, ou de coût minimal pour une variance donnée.
- Avec n_h compris entre une valeur minimale a_h et une valeur maximale b_h : nombreuses références (Neyman, Aeberhardt et Marcus, Koubi et Mathern à l'Insee, mais aussi Gabler et al., etc.)
- Remarquent que l'ordre des éléments de $A = \{a_h\sqrt{c_h}/d_h, b_h\sqrt{c_h}/d_h; h = 1, \ldots, H\}$ donne un ordre "d'activation" des contraintes (c_h: coût).
- Pour un plan de Poisson, $a_h = 0$ acceptable mais j'ai besoin de pouvoir choisir a_h plus loin.

Allocation - 3

- Résultat de Gabler et al. semble faux.
- J'ai (je pense) reproduit la même chose que Aeberhardt et Marcus.
- Et *montré* que c'était la solution du problème.
- Difficultés de programmation : égalités dans A ; $d_h = 0$; comparaisons de réels ; variances infinies ($n_h = 0$) ; etc.

Allocation - 4

- Dans la LSE : *Limite de publication* pour un CV de 5%. Cible pour le CV : 3%.
- Statistiques d'intérêt (principales) : coût total de la main d'oeuvre par Section d'activité (niveau 1 de la classification), par classe de taille ; salaires médians par croisement de grande région et 39 domaines d'activité (composés à partir du niveau 3 de la classification). Extensions : mêmes objectifs que pour les grandes régions.
- 342 objectifs OFS, 345 objectifs extensions.
- Budget 2016 : 56'000 entreprises y compris extensions (dont toutes les ets. $\geq$ 50 employés, le "profiling").
- Attentes 2018 : réduire, mais prudemment.

Allocation - 5

- Problème d'optimisation convexe, mais solution rarement dans l'intérieur du domaine.
- *On* m'a dit que les algorithmes de résolution auraient du mal.
- Ma solution sous-optimale : allouer séquentiellement (un objectif, puis le suivant en complément, etc.)
- Et chercher un ordre pour ces objectifs.

Allocation - 6

- Calcul des objectifs : si le CV minimum dépasse 3 ou 5%, vouloir le meilleur CV revient beaucoup trop cher.
- Si le CV minimum est proche de la cible, cela peut aussi couter cher.
- Solution retenue : ajouter 0.5 ou 1 pt au CV minimum s'il est proche ou dépasse la cible.
- Le calage entraine une dépendence entre domaines même disjoints !
- Pour les extensions :
  - allocation à partir de l'échantillon OFS,
  - uniquement dans le canton ou la commune qui commande l'extension.

Allocation - 7

- Résultat pour LSE 2018 : 46'000 ets.
- *Validation* à l'aide d'une petite simulation (env. 2 jours pour 1'000 répétitions).
- CV compatibles, sauf dans domaines où une ou quelques ets.
  - représentent une grande part des emplois,
  - ont des revenus différents de ceux du reste du domaine.
- Recommandation : obtenir les réponses de ces ets.

Références - 1

- Aeberhardt, R. et Marcus, V. (2006). *Mesure et Contrôle de la Précision dans un Plan de Sondage Complexe. Cas de l'Enquête sur la Structure des Salaires de 2006.* Présentation à un atelier méthode de la DSE, INSEE.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25, pp. 193-204.
- Gabler, S., Ganninger, M. et Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75, pp. 151-161.

Références - 2

- Graf, M. (2017). A simplified approach to linearization variance for surveys. *Document de travail.*
- Koubi, M. et Mathern, S. (2009). *La nouvelle méthode d'échantillonnage de l'enquête trimestrielle ACEMO depuis 2006. Amélioration de l'allocation de Neyman.* Document d'études de la Direction de l'animation de la recherche, des études et des statistiques, 146.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558-606.

Références - 3

- Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, pp. 167-195.

- Tillé, Y. et Vallée, A.-A. (2017). Variance Estimation by Linearization via the Sampling Indicators With Application to Nonresponse. *Document de travail.*