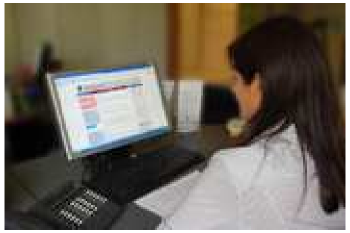
Les échantillons de réserve : éléments descriptifs et essai de modélisation simple

Marc CHRISTINE

Conseiller scientifique à la Direction de la méthodologie et de la coordination statistique et internationale

Institut national de la statistique et des études économiques (Insee), Paris, France



Insee

Mesurer pour comprendre

13 juin 2018

Plan :

1. Introduction : le contexte
2. Typologie : ce que n'est pas un échantillon de réserve
3. Quelques exemples complexes d'échantillons de réserve
4. Risques à éviter ou précautions dans l'utilisation des échantillons de réserve
5. Considérations statistiques
6. Modélisation statistique simplifiée
7. Conclusion

1. Introduction : le contexte

- Une **non-réponse** dont la prévalence est croissante dans les enquêtes

  - Diminution de la taille de l'échantillon utile
    => augmentation de la variance
  - Risque de biais : non-réponse corrélée aux variables d'intérêt de l'enquête.

- Une stratégie de recours : les échantillons de réserve.
- De plus en plus utilisée par les organisations de la statistique publique.

- **Principe** : accroître l'échantillon principal ..
  - …en lui adjoignant un échantillon additionnel (***réserve***)
  - … mobilisé si le taux de réponse à l'échantillon principal est jugé insuffisant.

- On ne parle ici que de ***non-réponse totale***

- Ou bien on assimile une non-réponse partielle à une non-réponse totale.

Premières questions

- Qu'est-ce qu'un taux de réponse insuffisant ?

- Quel est le but visé ?

  => garantir un nombre minimal de répondants.

- Pourquoi ne peut–on anticiper dès le départ ?
  - On peut disposer d'information *ex-ante* sur les taux de réponse dans des enquêtes similaires ou antérieures
  - … et ajuster a priori la taille de l'échantillon souhaitée
  - Mais le problème est le **coût.**
  - La stratégie relève du pari.

2. Typologie : ce que n'est pas un échantillon de réserve

L'appellation « échantillon de réserve » doit être bien distinguée de pratiques approchantes mais différentes :

- Remplacement déterministe d'une unité
- Remplacement aléatoire d'une unité
- Surreprésenter une strate par anticipation d'un faible taux de réponse
- Sélection a priori de couples ou de triplets d'unités (exemple PISA)
- ***Allotement*** :
  - L'échantillon principal est divisé en lots qui sont utilisés **séquentiellement et tous en totalité**
  - Convient par exemple pour un étalement temporel de la collecte
  - Ou constitution de vagues

3. Quelques exemples complexes d'échantillons de réserve

A. L'enquête sur la protection sociale complémentaire d'entreprise (obligation légale des entreprises au 1er janvier 2016).

- Volet établissements pour connaître l'offre.
- Volet salariés pour connaître le recours et l'opinion sur le dispositif.
- Échantillon de 8000 établissements.
- L'échantillon des salariés est tiré au 2ème degré parmi les salariés des établissements échantillonnés au 1er degré : ***3 salariés au plus par établissement*.**

Différents échantillons de réserve

- Deux échantillons de réserve de 2000 établissements, mobilisés si taux de réponse < 40%, puis 30 %.
- Si les échantillons de réserve sont utilisés, on y échantillonne des salariés, tous interrogés.
- Si l'établissement est non répondant, un seul salarié sera interrogé.

- Si le taux de réponse établissements est jugé suffisant mais le taux de réponse salariés insuffisant, on débloque un échantillon de réserve salarié : ***un 4ème salarié est tiré dans chaque établissement..***.
- … ***mais seulement dans les établissements de plus de 8 salariés.***

(pour des raisons de confidentialité)

- Échantillon de réserve salariés débloqué en totalité si le taux de réponse salariés est < 30%.

- **Difficulté complémentaire** : *l’échantillon de réserve salariés ne porte que sur une partie du champ.*

B. Enquête sur la surveillance médicale de l'exposition des salariés aux risques professionnels (SUMER)

- Un « échantillon » de médecins du travail (volontaires).
- Chaque médecin fait un tirage aléatoire de 30 salariés (systématique) dans la liste de ceux qu'il doit convoquer au cours d'une période donnée.
- Un échantillon de réserve de 10 salariés par médecin.
- Le médecin déclenche l'échantillon de réserve dès qu'il a 10 non-répondants.
- ***Problème*** : *comment contrôler que le médecin a fait tous les efforts nécessaires pour interroger son échantillon principal ?*

4. Risques à éviter ou précautions dans l'utilisation des échantillons de réserve

Sur les conditions d'utilisation des échantillons de réserve

- Bien définir les conditions de déclenchement
  - À partir de quels seuils de non-réponse
  - À partir de quel moment dans le processus de collecte

- Il peut y avoir des ***déclenchements séquentiels***
- Dans ce cas, il y a des lots, ceux-ci doivent être débloqués en totalité.

- Il peut y avoir des échantillons de réserve ***définis dans des strates*** a priori et déclenchés séparément et indépendamment dans chaque strate.

L’échantillon de réserve n’est pas une panacée

- Il ne réduit pas la non-réponse.

- Il faut continuer à faire porter un effort sur la collecte afin de réduire le taux de non-réponse.

- Il faut poursuivre / améliorer / perfectionner les procédures de corrections de la non-réponse et/ou de calage

- Il faut être conscient des apports et des limites :
  - Réduction de la variance
  - Mais risque de ne pas supprimer le biais : viser un nombre de répondants sans tenir compte de la non-réponse sélective.

=> on ne corrige que les inconvénients « visibles »

- Il y a un risque quand on a recours à un prestataire extérieur
  - S'il sait qu'il aura un échantillon de réserve, il peut avoir un effort moindre de collecte.
  - Il faut donc des ***clauses de contrôle strict...***
  - …avant de donner le « droit » d'utiliser les échantillons de réserve (nombre minimal de contacts, à des heures / jours / modalités de contacts variés) : ***ni prévenir, ni donner tôt***.
  - Une solution : ***tarification*** = augmenter la rémunération du questionnaire marginal pour tenir compte de la difficulté croissante de contact et éviter de se contenter des unités les plus « faciles » à contacter et enquêter.

Le traitement des échantillons de réserve doit être aussi proche que possible du traitement appliqué à l'échantillon principal :

- Modalités d'information préalable
- Modalités de contact
- Insistance et relance des enquêtés : *le prestataire doit avoir la même insistance pour les échantillons de réserve.*
- Ne pas s'arrêter quand l'objectif en termes de taux de réponse ou de nombre de répondants est atteint
- Administration du questionnaire …

5. Considérations statistiques

- La plupart du temps, on fait comme si on avait d'emblée utilisé un échantillon plus gros.
- On oublie le processus de sélection et de décision de mobilisation de l'échantillon de réserve.
- Les estimateurs utilisés ne prennent pas en compte le processus séquentiel.
- Les calculs d'espérance et de variance devraient être modifiés…
- …sinon on utilisera des formules moins appropriées négligeant le processus

  => Risque d'estimation altérée de la variance.

Réflexions sur la notion de taux de réponse

➢ Le taux de réponse : simple dénombrement ou variable d'intérêt sur la population.

- On définit les variables Y_i modélisant les comportements de réponse des unités i :

$$\boxed{Y_i = 1 \Leftrightarrow i \text{ répond, sinon : } Y_i = 0}.$$

Ces variables sont définies *a priori* sur l'**ensemble de l'échantillon**.

- Pour un échantillon de taille n :

taux de réponse = proportion empirique de répondants parmi les unités de l'échantillon :

$$\frac{1}{n}\sum_{i \in S} Y_i \,.$$

- Mais on peut aussi considérer que le comportement de réponse est une **caractéristique de la population**, qui pourrait donc être définie sur celle ci tout entière.

Paramètre d'intérêt dans la population = proportion : $\frac{1}{N}\sum_{i=1}^{N} Y_i$.

- Les comportements de réponse sont effectivement (mais seulement) observés sur tout l'échantillon (***sans non-réponse*** !),

- $\Rightarrow$ Estimation sans biais du taux de réponse : $\frac{1}{N}\sum_{i \in S}\frac{Y_i}{\pi_i}$

ou estimateur de HAJEK : $\frac{\sum_{i \in S}\frac{Y_i}{\pi_i}}{\sum_{i \in S}\frac{1}{\pi_i}}$.

Hypothèse implicite : le comportement de réponse ne dépend que des caractéristiques de l'individu et non du fait qu'il ait été sélectionné ou pas.

- $\tau = \frac{1}{N}\sum_{i=1}^{N} Y_i$: **vrai taux de réponse intrinsèque** dans la population (évidemment inconnu)

- $\hat{\tau}(S)$: statistique de taux de réponse calculée dans l'échantillon S,

  estimateur sans biais de τ :

$$\hat{\tau}(S) = \left| \begin{array}{c} \frac{1}{N}\sum_{i\in S}\frac{Y_i}{\pi_i} = \frac{1}{N}\sum_{i\in R}\frac{1}{\pi_i} \\ ou : \frac{\sum_{i\in R}\frac{1}{\pi_i}}{\sum_{i\in S}\frac{1}{\pi_i}} \end{array} \right.$$

Cadre théorique

Plusieurs formalisations possibles :

- Deux échantillons indépendants

- Échantillon de réserve disjoint de l'échantillon principal

  Problème sur la préservation des propriétés d'équilibrage si l'on utilise la réunion de ces deux échantillons

  (*cf. échantillons conditionnels successifs*).

- **Cadre retenu** : échantillonnage en plusieurs phases

  Préserve les conditions d'équilibrage à condition de tirer les échantillons de 2nde phase sur des variables ad hoc.

***Rappel* : Pour avoir un équilibrage de l’échantillon de 2nde phase sur Z :**

- Équilibrage de l’échantillon de 1ère phase sur Z.

- Équilibrage lors du tirage conditionnel de l’échantillon de 2nde phase sur : $\dfrac{Z}{\pi^1}$

- Probabilités conditionnelles adaptées lors de ce tirage :

$$\pi_i^{2/S_1} = \frac{\pi_i^2}{\pi_i^1}\boldsymbol{1}_{i \in S_1}$$

6. Modélisation statistique simplifiée

- Univers U, taille N
- 1er échantillon sans remise, S_0, dit ***complet***
  - Probabilités d'inclusion π_i^0
  - Taille $n(S_0)$
- Échantillon de 2nde phase, S, dit ***échantillon de référence*** ou ***principal***
  - ***Tirage aléatoire simple au sein de S, de taille n fixée.***
  - Probabilités d'inclusion conditionnelles : $\boxed{\pi_i^{2/S_0} = \dfrac{n}{n(S_0)} \mathbf{1}_{i \in S_0}}$.
  - Probabilités finales d'inclusion : $\boxed{\pi_i = E\pi_i^{2/S_0} = n\, E\left[\dfrac{\mathbf{1}_{i \in S_0}}{n(S_0)}\right] = \dfrac{n}{n_0}\pi_i^0 \; [si \; n(S_0) = n_0]}$.

Statut de ces deux échantillons :

- L'**échantillon de référence** (*utile*) est l'échantillon S, tiré en 2nde phase
- Si l'on décide de recourir à un échantillon de réserve, l'échantillon utilisé sera l'échantillon complet S_0.
- L'**échantillon de réserve** sera alors : $\boxed{S_{Rv} = S_0 \setminus S}$

Dans la pratique :

Échantillon complet de taille fixe n_0

Taille de l'échantillon de réserve = fraction α (= 5, 10%, ..) de la **taille *n* de l'échantillon de référence**

Relations entre les tailles et les probabilités d'inclusion :

$$n = \frac{n_0}{1+\alpha}$$

$$\begin{cases} \pi_i^{s_0} = \dfrac{1}{1+\alpha} 1_{i \in s_0} \\ \pi_i = \dfrac{\pi_i^0}{1+\alpha} \end{cases}$$

Non-réponse dans l'échantillon principal

- Sous-échantillon des répondants noté $R(\subset S)$
- **Echantillon poissonnien de 3ème phase tiré dans** S
- **Probabilités conditionnelles :**

$$\boxed{\pi_i^{3/S} = p_i \mathbf{1}_{i \in S}}.$$

Formules d'estimation d'un total T(X) à partir de l'échantillon principal :

- ***Sans*** non-réponse à partir de l'échantillon de référence S :

$$\hat{T}_S(X) = \sum_{i \in S} \frac{X_i}{\pi_i}$$

- ***Avec*** non-réponse à partir de l'échantillon de référence S :

$$\hat{T}_3(X) = \sum_{i \in R} \frac{X_i}{\pi_i \, p_i} \quad (p_i \ \textit{inconnues, à estimer})$$

Mobilisation de l'échantillon de réserve

On mobilise l'échantillon de réserve si :

$$\boxed{\hat{\tau}(S) < \beta}$$

Échantillon de réserve : $\boxed{S_{Rv} = S_0 \setminus S}$

Échantillon utilisé au final = échantillon complet S_0.

- Ainsi, ***conditionnellement à l'événement*** $\{\hat{\tau}(S) < \beta\}$, on devrait prendre comme estimateur du total de X l'estimateur de HORWITZ-THOMSON sur échantillon complet :

$$\hat{T}_0(X) = \sum_{i \in S_0} \frac{X_i}{\pi_i^0} = \sum_{i \in S} \frac{X_i}{\pi_i^0} + \sum_{j \in S_0 \setminus S} \frac{X_j}{\pi_j^0}.$$

Non-réponse dans l'échantillon de réserve

- Sous-échantillon des répondants noté $R_{Rv} \ (\subset S_{Rv})$
- **Echantillon poissonnien de 3ème phase tiré dans** $S_{Rv} = S_0 \setminus S$
- **Probabilités conditionnelles :**

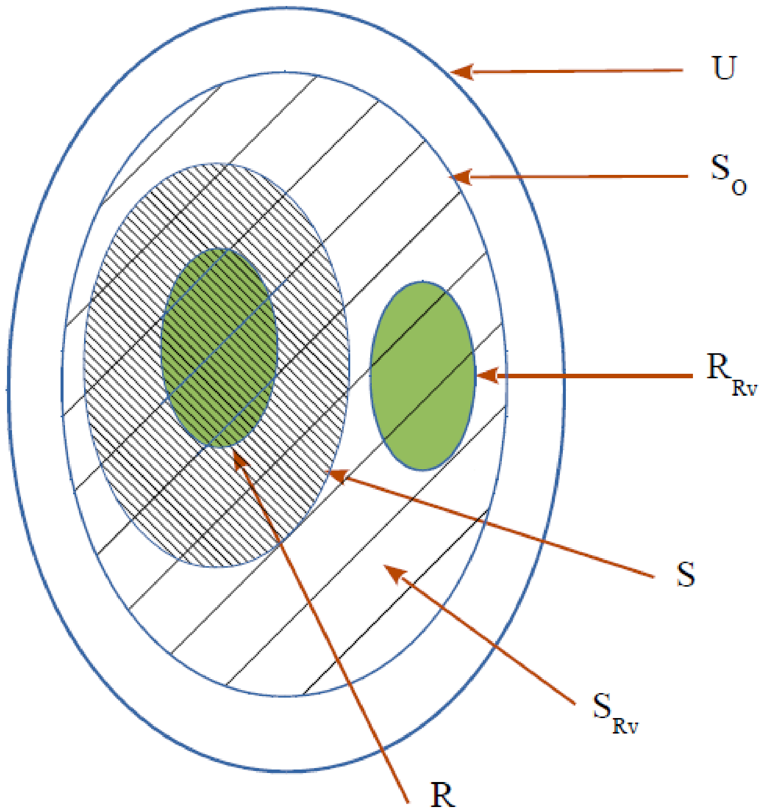
$$\boxed{\pi_j^{3/S_{Rv}} = q_j \, \mathbf{1}_{i \in S_{Rv}} = q_j \, \mathbf{1}_{i \in S_0} \, \mathbf{1}_{i \notin S}}.$$

En général, sous l'hypothèse : $q_j = p_j$

$$P\{i \in R_{Rv} / i \in S_{Rv}\} = P\{i \in R / i \in S\}$$

= Probabilités définies ex-ante sur toute la population, indépendamment du tirage, inconnues mais estimables.

Schéma d’imbrication des échantillons



Formules d’estimation d’un total à partir de l’échantillon *complet* en présence de non-réponse :

$$\hat{T}_{Rv}(X) = \sum_{i \in R} \frac{X_i}{p_i \pi_i^0} + \sum_{j \in R_{Rv}} \frac{X_j}{q_j \pi_j^0}$$

- C’est l’estimateur que l’on considèrerait si l’on mobilisait ***d’emblée*** la réserve (quelle que soit l’appréciation sur le taux de réponse dans l’échantillon de référence).

Dans le cas d'un processus de mobilisation ***aléatoire*** d'un échantillon de réserve, **l'estimateur réellement utilisé (« composite » ou « final »)** :

$$\boxed{\begin{aligned}\hat{T}_f(X) &= \hat{T}_3(X)\,1_{\hat{\tau}(S)\geq\beta} + \hat{T}_{Rv}(X)\,1_{\hat{\tau}(S)<\beta}\\ &= \hat{T}_3(X) + [\hat{T}_{Rv}(X) - \hat{T}_3(X)]\,1_{\hat{\tau}(S)<\beta}\end{aligned}}$$

Expression explicite :

$$\hat{T}_f(X) = \hat{T}_3(X) + \left[\underbrace{\sum_{i \in R} \frac{X_i}{p_i \pi_i^0} + \sum_{j \in R_{Rv}} \frac{X_j}{q_j \pi_j^0}}_{=\hat{T}_{Rv}(X)} - \underbrace{\sum_{i \in R} \frac{X_i}{p_i \pi_i}}_{=\hat{T}_3(X)} \right] 1_{\hat{\tau}(S)<\beta}$$

$$= \hat{T}_3(X) + \left[\sum_{i \in R} \frac{X_i}{p_i} \left(\frac{1}{\pi_i^0} - \frac{1}{\pi_i} \right) + \sum_{j \in R_{Rv}} \frac{X_j}{q_j \pi_j^0} \right] 1_{\hat{\tau}(S)<\beta}$$

Avec :

$$\begin{cases} R \subset S \subset S_0 \\ R_{Rv} \subset S_{Rv} = S_0 \setminus S \end{cases}$$

- On a donc trois estimateurs à comparer :
  - Estimateur sur échantillon principal seul (avec non-réponse) :

$$\hat{T}_3(X) = \sum_{i \in R} \frac{X_i}{\pi_i \, p_i}$$

  - Estimateur sur échantillon complet incluant la réserve (avec non-réponse)

$$\hat{T}_{Rv}(X) = \sum_{i \in R} \frac{X_i}{p_i \, \pi_i^0} + \sum_{j \in R_{Rv}} \frac{X_j}{q_j \, \pi_j^0}$$

  - Estimateur « final », tenant compte du processus aléatoire de choix des échantillons retenus.

$$\hat{T}_f(X) = \hat{T}_3(X) + [\hat{T}_{Rv}(X) - \hat{T}_3(X)] \mathbb{1}_{\hat{\tau}(S) < \beta}$$

- Estimateurs sans biais, asymptotiquement pour l'estimateur « final » (*du moins si les probabilités de réponse sont connues et exactes*).

- Variances calculables et comparables (en tenant compte des processus de sélection aléatoire des phases 2 et 3) :

$$\boxed{V[\hat{T}_3(X)] - V[\hat{T}_{Rv}(X)] \approx \frac{\alpha}{1+\alpha} \sum_{i=1}^{N} \frac{(X_i)^2}{\pi_i \, p_i}}$$

- Pour l'estimateur « final » :

  - Espérance : $E\hat{T}_f(X) = T(X) + Cov[\hat{T}_{Rv}(X) - \hat{T}_3(X), \mathbf{1}_{\hat{\tau}(S)<\beta}]$

  - Variance exacte difficile à calculer.

$$\boxed{\begin{aligned} V\hat{T}_f(X) &= (1-\mu)V[\hat{T}_3(X)/\hat{\tau}(S) \geq \beta] + \mu V[\hat{T}_{Rv}(X)/\hat{\tau}(S) < \beta] \\ &\quad + \mu(1-\mu)\left[E[\hat{T}_3(X)/\hat{\tau}(S) \geq \beta] - E[\hat{T}_{Rv}(X)/\hat{\tau}(S) < \beta]\right]^2 \\ &avec : \mu = P\{\hat{\tau}(S) < \beta\} \end{aligned}}$$

- Une approximation grossière :

$$\boxed{\mathbf{1}_{\hat{\tau}(S)\geq\beta} \approx \mathbf{1}_{\tau\geq\beta} = \left| \begin{array}{l} 1 \ si \ \tau \geq \beta \\ 0 \ si \ \tau < \beta \end{array} \right.}$$

- Approximation plus sophistiquée :

$$\boxed{E[\hat{T}_f(X)-T(X)]^2 \approx V[\hat{T}_{Rv}(X)] + 2E\left[[\hat{T}_{Rv}(X)-T(X)][\hat{T}_3(X)-\hat{T}_{Rv}(X)]\mathbf{1}_{\hat{\tau}(S)\geq\beta}\right]}$$

7. Conclusion

- Des travaux à poursuivre sur le plan théorique.
- Les conforter par des simulations.
- Le cadre théorique proposé reste très réducteur => l'élargir à d'autres modes de constitution et de mobilisation des échantillons de réserve.
- Des hypothèses à discuter :
  - Existe-t-il une propension à répondre « universelle » (définie ex-ante) ?
  - Peut-on la mesurer (par une probabilité de réponse) ?
  - Peut-on définir un taux de réponse « intrinsèque » dans la population ?
  - Ce taux est « manipulable ».

7. Conclusion

- Une pratique relativement simple à mettre en œuvre mais qui ne doit pas faire illusion quant aux gains réels qu'elle procure.

- Des conditions de mobilisation qui doivent être strictement contrôlées

  - tant en amont (prescription du responsable d'enquête)
  - …qu'en aval (collecte terrain)
  - …ainsi qu'**au niveau de l'exploitation statistique.**


jms
13es
Journées de Méthodologie Statistique de l'Insee
du 12 au 14 juin
2018
Journées de Méthodologie Statistique
Informations & inscriptions : jms-insee.fr
Contact : jms2018@jms-insee.fr

Merci pour votre attention !

marc.christine@insee.fr

Insee
88 Avenue Verdier – CS 70058
92541 Montrouge Cedex
www.insee.fr

Informations statistiques :
www.insee.fr / Contacter l'Insee
09 72 72 4000
(coût d'un appel local)
du lundi au vendredi de 9h00 à 17h00