UTILISATION D'UNE ENQUÊTE LÉGÈRE AUPRÈS DES NON-RÉPONDANTS POUR CORRIGER LA NON-RÉPONSE TOTALE PAR LA MÉTHODE DES GROUPES DE RÉPONSE HOMOGÈNES.

Henri BODET[1] (*)

(*) Insee, Pôle Ingénierie Statistique d'Enquête

Résumé

Dans le cas où le comportement de réponse est très différent pour une sous-population dont c'est l'objet même de l'enquête d'appréhender la taille et le contour, la méthode habituelle de correction de la non-réponse qui consiste à définir des groupes de réponses homogènes rassemblant des répondants et des non-répondants ne peut pas être mise en œuvre directement faute de savoir à quel groupe appartiennent les non-répondants. Lorsqu'on dispose des résultats d'une enquête auprès des non-répondants qui étaye l'hypothèse d'un comportement de réponse différent - et implique la nécessité de réaliser un traitement spécifique - la question se pose de savoir s'il faut intégrer les résultats de cette enquête supplémentaire et comment.

L'objet de cette communication est de présenter une méthode pour prendre en compte cette enquête auprès des non-répondants. L'idée est de scinder les groupes de réponse homogènes suivant l'appartenance ou non à la sous-population qui a un comportement spécifique et d'estimer la taille de ces groupes à l'aide de l'enquête auprès des non-répondants.

On présentera les propriétés de cette méthode dans un cadre théorique simplifié - un sondage aléatoire simple - et le résultat de simulations qui mettent en évidence le gain et les limites de la démarche. Les principaux résultats théoriques sont que la méthode permet de supprimer le biais dû à la différence de comportement de réponse au prix d'une augmentation de la variance qui peut- si l'échantillon de l'enquête auprès des non-répondants est trop petit - annihiler l'intérêt de réduire le biais.

On relatera également l'expérience d'une application de cette méthode à l'enquête sur la filière « aéronautique et spatiale » dans le grand Sud-Ouest en 2013.

Abstract

When there is a sub-population whose propensity to answer a survey is significantly different, homogeneous response groups should be built according to the fact units belong to this sus-population or not. In the case of a survey whose purpose is to assess whether units belong to this sub-population or not, non-respondents cannot be assigned to such a group. Sometimes, an additional survey among non-respondents is made to check if response probability is different from one sub-population to the other. We suggest a method to combine this additional survey and the use of the homogeneous response groups method in dealing with non-response.
In a very simple framework, we will state conditions under which the method is efficient. We will also give an account of a business survey to which this method was applied

Mots-clés

Non-réponse, groupes de réponse homogènes

[1]henri.bodet@insee.fr

Introduction

Nous nous plaçons dans le cas d'une enquête dont l'objectif est d'étudier une sous-population de la population de départ dont les membres sont *a priori* inconnus. Voici deux exemples de telles enquêtes :

- l'enquête sur les moyens et modes de gestion de l'immatériel (EMMGI) réalisée par l'Insee en 2006. Il s'agissait de savoir comment les entreprises géraient leur capital « immatériel ». Un certain nombre d'entreprises n'étaient pas concernées par le sujet de l'enquête mais il était impossible de les exclure ex-ante de la base de sondage. Un des premiers objectifs de l'enquête était de déterminer les entreprises qui avaient effectivement une politique de gestion de l'immatériel ;
- l'enquête sur la filière aéronautique et spatiale dans le grand Sud-Ouest (FAS-GSO) en 2013. L'objectif de cette enquête est d'identifier et d'étudier les entreprises qui participent à la filière « construction aéronautique et spatiale ». Ceci a conduit à interroger des entreprises qui, au vu de leur activité, travaillaient potentiellement pour la filière mais sans que l'on sache avant l'enquête si c'était effectivement le cas.

Il y a un biais dû à la non-réponse lorsque le comportement de réponse est différent selon les unités. Dans les deux cas évoqués ci-dessus, on peut penser *a priori* que les entreprises concernées par le sujet de l'enquête répondront mieux. En effet, on peut supposer que les autres auront tendance à juger leur réponse comme « inutile ». Ce biais est bien compris des utilisateurs de l'enquête - au-delà des statisticiens.

Dans les deux cas, on a procédé à une enquête légère auprès des non-répondants pour savoir si, oui ou non, ils avaient une plus ou moins grande propension que les répondants à appartenir à la population « cible ». Dans les deux cas, cette enquête a montré que c'était effectivement le cas.

Dans ces deux enquêtes la non-réponse avait été redressée par repondération. Les techniques de repondération reviennent toutes à estimer une probabilité de réponse qui sert à augmenter le poids des non-répondants. En particulier, on utilise souvent la méthode des groupes de réponse homogènes (GRH) qui consiste à découper la population enquêtée en sous-population dont on peut considérer qu'elles ont le même comportement de réponse. La probabilité de réponse de chaque groupe est alors estimée en considérant la part des répondants à l'intérieur du groupe.

Une limitation pratique de cette méthode est qu'il faut pouvoir constituer des groupes à l'aide d'une information disponible à la fois sur les répondants et les non-répondants. Cela restreint *a priori* les critères qui peuvent être utilisés pour constituer ces groupes à des données connues sans avoir recours à l'enquête. En particulier, cela empêche d'utiliser la réponse à une question de l'enquête pour constituer ces groupes.

Nous nous proposons de décrire une méthode d'intégration de l'enquête auprès des non-répondants aux groupes de réponse homogène. Cette méthode a été utilisée sur l'enquête FAS-GSO ; elle avait auparavant servi à redresser l'EMMGI sous la direction de Philippe Brion alors chef de l'unité d'harmonisation des enquêtes entreprises à l'Insee. À l'époque, cette méthode n'avait pas été documentée. L'objet de cette communication est de palier ce manque et d'examiner sur un cas simple dans quelle mesure elle peut être pertinente pour avoir des éléments plus concrets qu'une simple intuition.

Nous commencerons par décrire le problème général et par évoquer d'autres moyens de traiter cette non-réponse liée aux variables d'intérêt. Nous décrirons ensuite de façon formelle la méthode proposée avant de donner un compte rendu de son application à l'enquête FAS-GSO. Dans le cadre très simplifié d'un sondage aléaloire simple, nous examinerons les conditions de validité de cette démarche de façon théorique et à l'aide de simulations.

1. Redresser la non-réponse liée à une variable d'intérêt.

Le problème auquel on est confronté n'a rien de nouveau et plusieurs méthodes ont été envisagées pour le résoudre. La spécificité de la situation dans laquelle on suppose se trouver est le fait de disposer d'une enquête auprès des non-répondants. La revue de méthodes qui suit n'est pas exhaustive mais elle peut présenter un intérêt pour situer ce qui est envisageable. Nous tenons en outre à évoquer ces alternatives car la méthode qui fait l'objet de cette communication n'est pas une panacée et qu'elle ne s'applique qu'à des situations particulières.

1.1. Si on ne dispose pas d'une enquête auprès des non-répondants.

Nous commençons par rappeler deux approches possibles si on ne dispose pas d'une enquête auprès des non-répondants ou si on ne veut pas l'utiliser. En effet, on peut juger que l'échantillon de cette enquête est trop petit ou qu'elle est entachée d'un biais.

1.1.1. Utiliser les techniques habituelles de correction de la non-réponse.

Ceci peut paraître paradoxal, car justement elles ne s'appliquent pas en première approche. En particulier, si on se penche sur la technique des groupes de réponse homogènes, on ne peut pas les constituer selon une variable obtenue grâce à l'enquête.

Mais il faut garder à l'esprit que, en général, le plan de sondage est stratifié suivant des critères liés à l'objet de l'enquête et que cela réduira *de facto* le biais dû au lien entre la variable d'intérêt et la non-réponse.

De même, s'il existe des variables auxiliaires liées à la variables d'intérêt, elles apparaîtront comme liées à la non-réponse - ne serait-ce que par leur lien avec la variable d'intérêt et elles seront vraisemblablement incluses dans le modèle du comportement de réponse. Elles serviront ainsi à définir les groupes de réponse.

Bien entendu, cette démarche ne fera pas disparaître le biais dû au caractère non-uniforme de la non-réponse mais elle peut le réduire considérablement.

Lors du traitement de l'enquête FAS-GSO, une correction de la non-réponse a aussi été faite suivant les méthodes habituelles et elle a été comparée à la méthode utilisant l'enquête auprès des non-répondants.

1.1.2. Utiliser le calage généralisé.

En théorie, le calage généralisé peut permettre de traiter le biais dû au lien entre le comportement de réponse et la variable d'intérêt en ayant recours à une variable instrumentale. Plusieurs articles ont présenté cette méthode. On en trouvera une description dans [1] et un exemple d'utilisation en (2].

Toutefois, le résultat peut être sensible au choix de la variable instrumentale ; cette stratégie nous a donc été déconseillée. De plus, si on obtient des résultats contre-intuitifs par cette méthode, il sera difficile de les justifier.

1.2. Avec une enquête auprès des non-répondants.

Lors du traitement de l'enquête FAS-GSO, en raison de la forte présomption d'un lien entre le comportement de réponse et l'appartenance à la filière, le service enquêteur a réalisé une enquête légère auprès des non-répondants. Elle a montré une spécificité du comportement de réponse et la question s'est posée de savoir comment l'intégrer.

1.2.1. Qu'est-ce qu'une enquête légère ?

Plaçons-nous dans le cas de l'enquête FAS-GSO où on se demandait si l'appartenance à la filière influait le comportement de réponse.

On entend par « enquête légère » une enquête – ici téléphonique – auprès des non-répondants dont l'objet est uniquement de savoir si les entreprises appartiennent à la filière ou non. Il ne s'agit donc pas de faire remplir le questionnaire par téléphone mais uniquement de savoir comment les non-répondants se situent par rapport à l'appartenance à la filière. Il est important de préciser ces points.

Le fait de ne poser qu'une seule question simple d'une part évite de mettre en place un véritable outil de saisie, évite d'augmenter la charge de l'enquête et permet d'avoir quasiment tout le temps une réponse. Ceci nous place dans la situation où on n'a pas à « gérer la non-réponse de l'enquête auprès des non-répondants ». Le premier objectif de cette enquête de voir s'il y a ou pas un comportement différent entre les deux populations.

Par contre, le fait de ne disposer que de la réponse à une seule question ne permet pas d'intégrer les réponses de cette enquête auxiliaire dans celles de l'enquête initiale.

1.2.2. Ce qu'il ne faut pas faire avec les résultats de l'enquête.

Dans le cas d'une enquête comme l'enquête FAS-GSO, les unités qui n'appartiennent pas à la filière ne sont pas censées répondre à une autre question que celle-ci. On peut donc être tenté de les intégrer - et seulement elles - au fichier de résultat comme si elles étaient répondantes. Ceci contribue à augmenter artificiellement la probabilité de réponse des unités « hors filière », ce qui peut aggraver le biais dû au lien entre la réponse à l'enquête et la variable d'intérêt. Il est donc important de traiter l'information qui vient de cette enquête auxiliaire de façon spécifique.

1.2.3. Utiliser les résultats de l'enquête auprès des non-répondants pour caler le fichier.

C'est ce qui a notamment été fait dans le cadre de l'enquête emploi en continue où une enquête téléphonique auprès des non-répondants est disponible.

Cette méthode repose sur la prise en compte de cette enquête de l'enquête auprès des non-répondants pour obtenir des marges de calage sur des variables cibles. Elle est décrite dans [3]. Elle n'a pas été adoptée dans le cas de l'enquête FAS-GSO car les marges de calages ainsi obtenues semblaient entachées d'une trop grande imprécision - ou du moins d'une imprécision comparable à celle de l'enquête.

2. La méthode employée pour intégrer les résultats d'une enquête légère auprès des non-répondants.

2.1. Le principe

Comme il l'a été expliqué plus haut, la seule information dont on dispose pour les unités interrogées lors de l'enquête auprès des non-répondants (ENR) est leur appartenance ou non à une sous-population . (La filière pour l'enquête FAS-GSO). On suppose que la méthode utilisée pour redresser l'enquête est celle des groupes de réponse homogènes (décrite par exemple dans [4] pp 259-260). Les résultats de l'enquête auprès des non-répondants devraient pousser à construire ces groupes suivant l'appartenance ou non à cette sous-population. (appartenance ou non à la filière par exemple)

L'application de cette méthode des groupes de réponse homogènes pour traiter la non-réponse suppose en théorie de pouvoir réunir dans un groupe les répondants et les non-répondants ; on estime ensuite leur probabilité de réponse par un calcul comme : $\frac{nombre\ de\ répondants\ dans\ le\ groupe}{taille\ du\ groupe}$. Or la taille du groupe s'écrit comme la somme du nombre de répondants et de non-répondants. On n'a donc pas véritablement besoin de savoir qui sont les non-répondants dans le groupe mais de connaître leur nombre. L'idée de la méthode est tout simplement d'estimer ce nombre à l'aide de l'enquête auprès des non-répondants.

En pratique, on pondère les estimations par leur poids de sondage - ce qui permet de retrouver la taille de la population initiale. Ceci ne change rien à la méthode, l'enquête auprès des non-répondants permet aussi d'estimer la somme des poids des non-répondants.

2.2 Description formelle de la méthode

Pour décrire la méthode de manière plus formelle, nous noterons :

- w_i le poids de sondage affecté à l'unité *i* dans l'enquête principale ;
- y_i la variable d'intérêt ;
- R l'ensemble des répondants et M celui des non-répondants
- E l'ensemble des unités contactées lors de l'enquête auprès des non-répondants ;
- g_j un poids de sondage propre à l'enquête auprès des non-répondants ;
- G_h un des H groupes de réponse homogènes (construits à l'aide d'une réponse à l'enquête).

On souhaite estimer la probabilité de réponse dans un groupe par : $\frac{\sum_{répondants} w_i}{\sum_{non-répondants} w_i + \sum_{répondants} w_i}$.

Le terme inconnu $\sum_{non-répondants} w_i$ est estimé par $\sum_{non-répondants\ contactés} w_i g_i$.

En notant $\tilde{\tilde{Y}}$ l'estimateur utilisant l'enquête auprès des non-répondants, on a

$$\tilde{\tilde{Y}} = \sum_{h=1\cdots H} \left(1 + \frac{\sum_{j \in G_h \cap E} w_j g_j}{\sum_{i \in G_h \cap R} w_i} \right) \sum_{i \in G_h \cap R} w_i y_i \qquad (E)$$

2.3 Cas d'un estimateur qui ne porte que sur une sous-population

On se place dans la situation où la population se partage en deux sous-population indexée par "0" et "1" ; dans le cas où il n'y a qu'un seul groupe de réponse homogène, l'estimateur s'écrit :

$$\tilde{Y} = \left(1+\frac{\sum_{j\in G_1\cap E} w_j g_j}{\sum_{i\in G_1\cap R} w_i}\right)\sum_{i\in G_1\cap R} w_i y_i + \left(1+\frac{\sum_{j\in G_0\cap E} w_j g_j}{\sum_{i\in G_0\cap R} w_i}\right)\sum_{i\in G_0\cap R} w_i y_i .$$

Nous nous placerons dans la suite du texte dans le cas particulier où on cherche à estimer un total qui ne porte que sur la sous-population 1. C'est-à-dire où y_i est nul en dehors de cette population. Ce cadre est en fait assez naturel : dans le contexte de l'enquête filière, on voulait principalement estimer le chiffre d'affaires de la filière, le nombre de salariés dont l'activité dépend de la filière et décrire la filière.

Dans ce cas le second terme disparaît. Or, ce second terme, est, du fait de l'introduction de l'enquête auprès des non-répondants fortement corrélé au premier. On voit bien qu'une sous-estimation du nombre de non-répondants dans le groupe 1 par cette enquête entraîne une surestimation de leur nombre dans le groupe 0.

Du fait que la variable d'intérêt est nulle sur le groupe 0 cette corrélation disparaît.

2.4 Absence de biais de l'estimateur intégrant l'enquête auprès des non-répondants.

Si on suppose que le système de poids de l'enquête auprès des non-répondants (ENR) conduit à des estimateurs sans biais, l'expression (E) permet de voir que l'espérance de l'estimateur $\tilde{\tilde{Y}}$ par rapport à cette enquête vaut :

$E_{ENR}\left(\tilde{\tilde{Y}}\right) = \sum_{h=1\ldots H}\left(1+\frac{\sum_{i\in M\cap G_h} w_i}{\sum_{i\in R\cap G_h} w_i}\right)\sum_{i\in R\cap G_h} w_i y_i$ c'est-à-dire l'estimateur que l'on aurait eu avec la méthode des groupes de réponse homogènes si on avait eu l'information pour les constituer.

Cet estimateur étant construit pour être approximativement sans biais, l'estimateur $\tilde{Y}$ possède cette propriété.

3. La méthode peut-elle fonctionner ?

L'estimateur intégrant l'enquête auprès des non-répondants est approximativement sans biais, son utilisation semble donc corriger parfaitement le problème dû au lien entre la non-réponse et la variable d'intérêt. Cependant, en le construisant, on a introduit une troisième phase de hasard - après l'échantillonnage et le comportement de réponse - qui est la sélection des unités interrogées. Intuitivement, on voit bien que cette phase conduit à un estimateur dont la variance est plus élevée. La question qui se pose est : cette variance supplémentaire ne vient-elle faire disparaître le gain que l'on a eu en supprimant le biais ?

Pour y répondre, nous nous placerons dans un cadre très simplifié par rapport à une enquête réelle mais suffisamment simple pour qu'on puisse y mener des calculs ou faire des simulations. Nous verrons que, si le comportement de réponse est effectivement différent entre les deux populations, l'intégration de l'enquête auprès des non-répondants peut se révéler bénéfique si la taille de l'échantillon est suffisament grande.

3.1 Cadre théorique simplifié et notations

Les enquêtes dont il est question ont des plans de sondages stratifiés. Toutefois, l'introduction de groupes de réponse homogènes crée de fait un lien entre les strates qui fait qu'on ne peut pas les étudier séparément.

Il existe des résultats pour calculer la variance de l'ensemble du processus échantillonnage stratifié puis repondération par les groupes de réponses homogènes. - voir par exemple [5]. Cependant, la prise en compte de l'aléa supplémentaire introduit par l'enquête auprès des non-répondants rend les calculs très lourds.

Pour obtenir des résultats théoriques et effectuer des simulations, nous nous placerons dans la situation suivante :

- on a réalisé un sondage aléatoire simple de n unités dans une population U de N unités ;
- la population est partagée en deux sous population U_0 et U_1 de tailles respectives N_0 et N_1 (l'inverse clarifierait la lecture !)
- on ne connaît l'appartenance à la population U_0 ou U_1 que pour les unités qui ont répondu à l'enquête ;
- chaque unité répond ou pas à l'enquête de façon indépendante ;
- on ne dispose pas d'information auxiliaire pour corriger la non-réponse ;
- la probabilité de réponse est p_0 sur la population U_0 et p_1 sur la population U_1 ;
- on interroge, via un sondage aléatoire simple, *k* non-répondants sur lesquels on a l'information de l'appartenance à U_0 ou U_1 ;

Précisons quelques notations :

- on note S l'échantillon initial, R l'ensemble des répondants, M celui des non-répondant et E celui de l'enquête auprès des non-répondants ;
- on note r et m le nombre de répondants et de non-répondants dans l'échantillon

.

- on indice par 0 les valeurs relatives à la population U_0 (ainsi, par exemple, n_0 est la taille de l'échantillon tombant dans la population 0), de même pour U_1 ;
- on note f_i la proportion de la population *i* dans la population totale ;
- enfin, on note $\hat{Y}$ l'estimateur classique sans intégrer l'enquête auprès des non-répondants et $\tilde{Y}$ celui qui la prend en compte.

On suppose de plus que l'on s'intéresse à la proportion d'unité de la population qui appartient à la population U_1 (ce qui revient à estimer le total d'une variable d'intérêt égale à $\frac{1}{N}$ sur la population U_1 et à zéro ailleurs). Pour peu que p_0 et p_1 soient différents, on se trouve donc mécaniquement dans la situation où le comportement de réponse est lié à la variable d'intérêt.

Dans ce cas, l'estimateur sans intégrer l'enquête auprès des non-répondants prend la forme suivante : $\hat{Y} = \frac{r_1}{r}$.

Comme l'enquête auprès des non-répondant est un sondage aléatoire simple de k unités parmi les m non-répondants, les poids g_j qui figurent dans l'expression (E) valent tous $\frac{m}{k}$.

L'expression (E) donne pour l'estimateur tenant compte de l'enquête auprès des non-répondants la forme suivante : $\tilde{Y} = \left(1 + \frac{\left(m\frac{k_1}{k}\right)}{r_1}\right)\frac{r_1}{n}$ ce qui conduit après simplification à $\tilde{Y} = \frac{r_1}{n} + \frac{m}{n}\frac{k_1}{k}$.

Si on adopte le point de vue que le comportement de réponse est un préalable à l'enquête, le premier terme s'interprète comme l'estimation de la proportion de répondants appartenant à la population sur la population U_1 (estimée par l'enquête principale) dans la population totale et le second comme la proportion de non-répondants appartenant à la population sur la population U_1 (estimée par un sondage à deux phases : l'enquête principale puis l'enquête sur les non-répondants) dans la population totale. La proportion de membres de la population U_1 s'obtient comme la somme de ces deux estimations.

3.2 Biais de l'estimateur sans intégrer l'enquête auprès des non-répondants.

Notons B le biais de l'estimateur classique - qui n'intègre pas l'enquête auprès des non-répondants.

Si on introduit le terme $\delta = \frac{p_1}{p_0} - 1$, terme qui ne dépend que du rapport des taux de réponse, le biais sur la proportion d'unités dans la population 1, peut être approché par $B \approx \frac{f_1(1-f_1)\delta}{f_1\delta + 1}$ (1)

Ce biais ne dépend pas de la taille de l'échantillon, il ne dépend que du comportement de réponse et de la variable d'intérêt.

C'est ce biais que l'introduction de l'enquête auprès des non-répondant cherche à réduire. Examinons maintenant à quel prix.

3.3 Variance supplémentaire introduite par l'enquête auprès des non-répondants.

Conditionnellement à l'échantillon et au comportement de réponse, l'estimateur $\tilde{Y}$ possède une variance qui est due à la réalisation de l' enquête auprès des non-répondants elle peut se calculer comme :

$$V\left(\tilde{Y}\middle|R,S\right)=\left(\frac{m}{n}\right)^2\left(\frac{1}{k}-\frac{1}{m}\right)\left(\frac{m_1}{m}\right)\left(1-\frac{m_1}{m}\right)\frac{m}{m-1} \quad \text{si} \quad k\le m\,.$$

Si l'enquête auprès des non-répondants est exhaustive ou quasiment exhaustive, la variance de l'estimateur qui l'intègre peut être inférieure à celle de l'estimateur direct. Nous supposerons que nous ne sommes pas dans ce cas où l'utilisation de l'enquête ne pose pas question.

Nous supposerons en outre que le carré du biais est prépondérant devant la différence de variance induite par la non-réponse. Dans ces conditions, on peut assimiler la variance conditionnelle ci-dessus à la variance supplémentaire induite par la prise en compte de l'enquête auprès des non-répondants.

Nous proposons d'obtenir un ordre de grandeur de cette quantité en remplaçant chaque terme aléatoire par son espérance par rapport au comportement de réponse et au tirage de l'échantillon, cela conduit à :

$$V(\bar{y}|S,R)\approx f_1q_1(1-f_1)q_0\left(\frac{1}{k}-\frac{1}{n(f_1q_1+(1-f_1)q_0)}\right)$$ (2). où $q_0=1-p_0$ - de même pour q_1

Cette expression n'a de sens que si $k\le n\left(f_1q_1+(1-f_1)q_0\right)$ où $n\left(f_1q_1+(1-f_1)q_0\right)$ représente le nombre moyen de non-répondants. Si ce n'est pas le cas, cela signifie que l'enquête sur les non-répondants sera exhaustive un nombre non-négligeable de fois.

Les simulations montrent que cette formule (2) fournit une estimation correcte de l'ordre de grandeur du supplément de variance tant que k n'est pas trop grand et si le biais est assez grand.

3.4 Quel échantillon minimal pour l'enquête auprès des non-répondants ?

L'introduction de l'enquête auprès des non-répondants permet de supprimer un biais donné par l'expression (1) moyennant l'introduction d'une variance supplémentaire approximativement estimée par l'expression (2). Si on considère le critère de la minimisation de l'erreur quadratique moyenne, l'introduction de l'enquête auprès des non-répondant est bénéfique si la variance supplémentaire est plus petite que le carré du biais. En comparant les expressions (1) et (2), on voit que cela revient à la condition $k\ge k^*$ où k^* s'exprime par $k^*=\left[\frac{1}{n(f_1q_1+(1-f_1)q_0)}+\frac{B^2}{f_1q_1(1-f_1)q_0}\right]^{-1}$ (3). Du fait des approximations en particulier dans l'expression (2), la valeur fournie par (3) ne doit pas être interprétée rigoureusement mais fournit un ordre de grandeur de la taille d'une enquête auprès des non-répondants qui améliore l'estimateur. Si le biais est nul, on trouve pour k^* le nombre moyen de non-répondants, ce qui s'interprète comme le fait que l'enquête sur les non-répondantes est exhaustive.

Voici quelques exemples de valeurs prises par B et k^*

n	f_1	p_1	p0	B	k^*
500	0,15	0,4	0,8	- 0,06	3
500	0,15	0,75	0,69	-0,02	54

3.5 Résultats de simulations

Des simulations ont été réalisées à l'aide du logiciel R. La situation simulée est toujours celle qui est décrite au paragraphe 3.1. On a supposé que l'on avait une population de 2 000 unités et que la taille de la population 1 est de 300 unités. On a fixé en outre la taille de l'échantillon à 500 unités. Par contre, l'estimateur calculé est le nombre d'unités la population 1 et non pas leur proportion. La valeur cible est donc de 300 unités.

On s'est placé dans trois situations qui conduisent à un taux de réponse moyen de 0,74 et où les unités de la population cible ont moins de chance de répondre :

Situation 1 : fort biais de non-réponse (variation du simple au double des probabilités de réponse)

Dans cette situation $p_1 = 0,4$ et $p_0 = 0,8$. La probabilité de répondre est deux fois moins élevée dans la population cible

Situation 2 : biais de non-réponse modéré (écart de 9 % entre les probabilités de répondre).

Dans cette situation $p_1 = 0,69$ et $p_0 = 0,75$. La probabilité de répondre est significativement moins élevée dans la population cible mais l'écart est plus faible.

Situation 3 : absence de biais de non-réponse

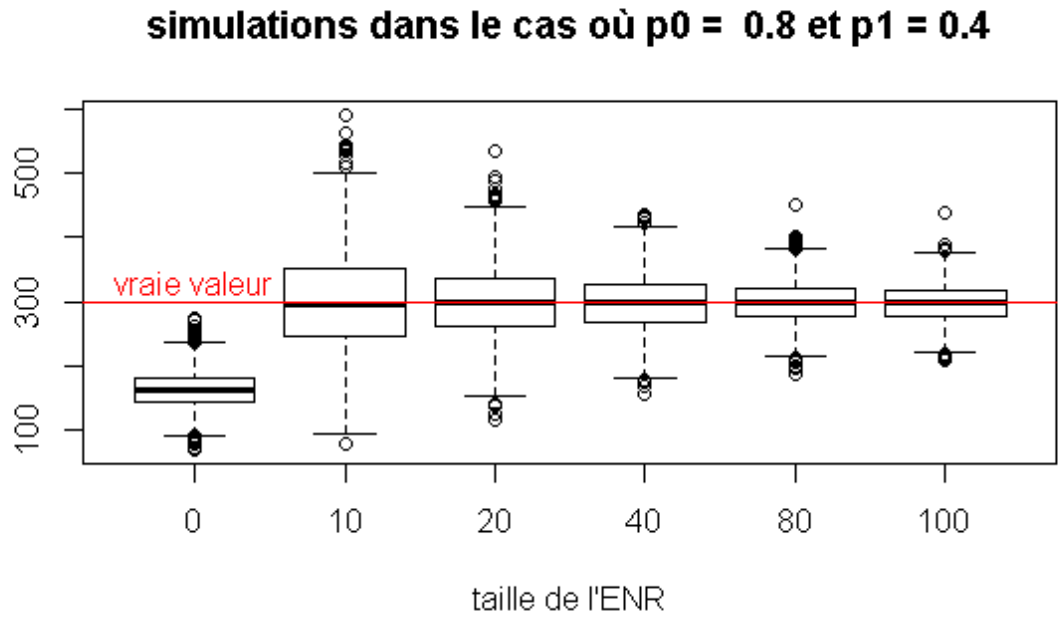
Dans cette situation $p_1 = 0,74$ et $p_0 = 0,74$. Il s'agit du cas où il n'y a pas normalement de raison de recourir à une enquête auprès des non-répondants.

Dans toutes ces situations, il y a en moyenne 130 non-répondants. On a simulé la réalisation d'enquêtes auprès des non-répondants portant sur 10, 20, 40, 80 et 100 unités.

En pratique, on a simulé mille fois de manière indépendante le comportement de réponse et le tirage d'échantillon. Et pour chacune de ces mille simulations, on a simulé la réalisation d'enquêtes auprès des non-répondants de 10, 20, 40, 80 et 100 unités.

Dans les simulations qui suivent, le supplément de variance « constaté » est obtenu en soustrayant la variance de l'estimateur intégrant une enquête auprès des non-répondants à celle de la variance de l'estimateur qui ne la prend pas en compte.

3.5.1 Résultat en cas de fort biais de non-réponse



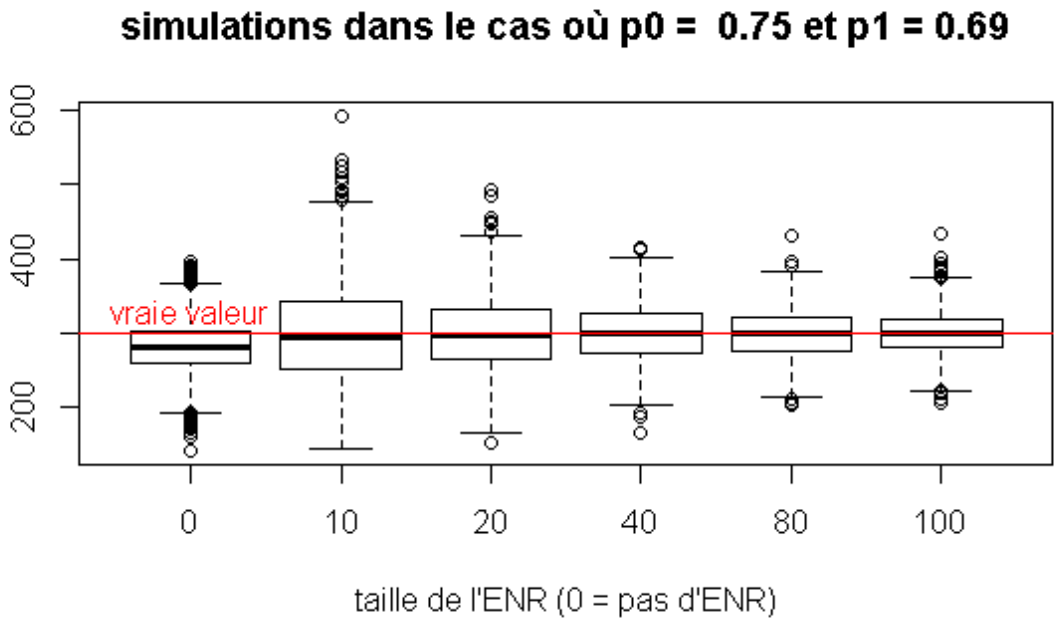
On voit que dans ce cas, l'estimateur sans enquête auprès des non-répondants est systématiquement moins bon. Ceci est conforme aux calculs effectués plus haut qui donnait un ordre de grandeur de 3 pour la taille minimal de l'enquête auprès des non-répondants.

Ces simulations permettent également d'évaluer empiriquement le biais et la variance des différents dispositifs. L'erreur moyenne est la racine carrée de l'erreur quadratique moyenne rapportée à la valeur à estimer.

taille de l'enquête auprès des non-répondants	**Écart-type**	**espérance de l'estimateur**	**Erreur moyenne**	**supplément de variance dû à l'ENR**	
				Constaté sur les simulations	**valeur donnée par la formule (2)**
0 (pas d'enquête)	27	162	47%	///	
10	80	299	27%	5626	5649
20	57	300	19%	2558	2589
40	44	299	15%	1209	1059
80	33	299	11%	369	294
100	30	298	10%	176	141

3.5.2 Résultats en cas de faible biais de non-réponse

On se place ici dans la situation 2 décrite plus haut.



On voit que dans ce cas une enquête auprès des non-répondants de taille trop petite supprime certes le biais dû à la différence de comportement de réponse.au prix d'une augmentation de la variance qui en fait perdre tout le bénéfice.

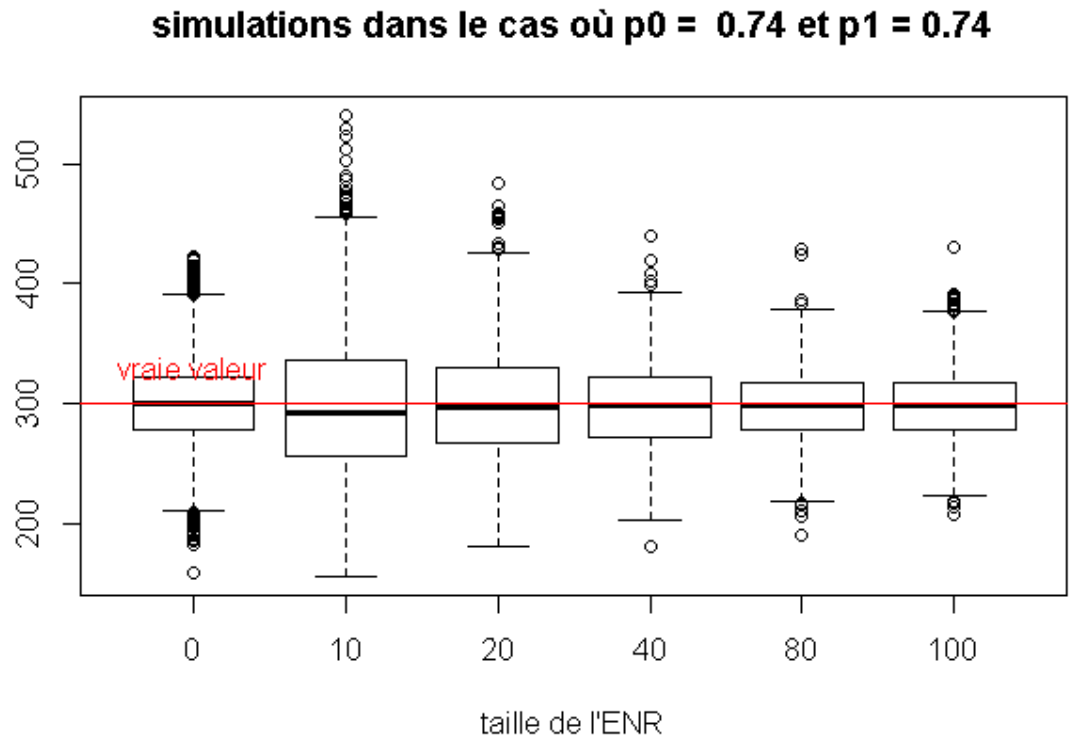
taille de l'enquête auprès des non-répondants	Écart-type	espérance de l'estimateur	Erreur moyenne	supplément de variance dû à l'ENR	
				Constaté sur les simulations	valeur donnée par la formule (2)
0 (pas d'enquête)	33	280	13%	///	
10	68	299	23%	3538	3647
20	50	299	17%	1402	1671
40	38	298	13%	368	683
80	31	298	10%	-102	189
100	29	298	10%	-218	90

Les calculs précédents indiquaient un seuil à partir duquel l'enquête auprès de non-répondants devient bénéfique à partir d'un seuil situé autour de 50 unités. On voit ici la limite des approximations faites pour établir ce résultat : la formule (2) ne peut pas, par construction, prédire les cas où la variance de l'estimateur intégrant l'enquête devient plus faible que celle de l'enquête auprès des non-répondants.

Étant donnée la faiblesse du biais, on voit que dans ce cas, l'utilisation de l'enquête est risquée. Effectivement, en situation réelle, on ne connaîtrait pas les paramètres servant à utiliser la formule (2) ; on ne pourrait que les estimer d'après l'enquête ou à dire d'expert.

3.5.3 Résultat en l'absence de biais de non-réponse

On se place maintenant dans la situation où il n'y a pas de biais. Sans surprise, les simulations confirment que dans ce cas, l'intégration de l'enquête auprès des non-répondants ne peut être que néfaste : la variance supplémentaire n'est compensée par aucune réduction de biais.



L'estimateur intégrant l'enquête auprès des non-répondants ne peut que se rapprocher de l'estimateur direct si l'échantillon de l'enquête auprès des non-répondants est assez important.

taille de l'enquête auprès des non-répondants	Écart-type	espérance de l'estimateur	Erreur moyenne	supplément de variance dû à l'ENR	
				Constaté sur les simulations	valeur donnée par la formule (2)
0 (pas d'enquête)	34	300	11%	///	
10	62	299	21%	2698	3182
20	48	300	16%	1166	1459
40	36	297	12%	173	597
80	30	298	10%	-235	166
100	29	298	10%	-277	80

4. Expérience d'utilisation de la méthode sur l'enquête « filière aéronautique et spatiale dans le grand sud-ouest »

4.1 L'enquête « filière aéronautique et spatiale dans le Sud-Ouest ».

L'objet de ces enquêtes est d'évaluer le poids de la filière aéronautique et spatiale parmi les entreprises du Grand Sud-Ouest (une appellation qui réunit les régions Aquitaine et Midi-Pyrénées) et de décrire les entreprises qui y appartiennent. Les principaux agrégats sont le chiffre d'affaires dans la filière et le nombre d'emplois qui y sont liés - qu'ils soient situés dans une entreprise totalement intégrée à la filière ou dans une entreprise pour qui la construction aéronautique et spatiale n'est qu'une activité.

Pour bien situer le problème, l'appartenance à la filière était définie comme le fait de participer à la fabrication de produit « aéronautique et spatiale ». Le secteur d'activité donne un indice d'appartenance à la filière mais ne suffit pas à la déterminer. Par exemple les entreprises dont l'activité est 25.62B – Mécanique industrielle peuvent ou travailler complètement pour la filière ou pas du tout ou ne travailler que partiellement.

La seule solution consistait à réaliser une enquête. Non seulement pour savoir quelles sont les entreprises qui travaillent pour la filière mais aussi pour avoir des informations sur leur intégration à cette filière et leurs perspectives. Il faut interroger également des entreprises dont on ne sait pas si elles sont membres de la filière mais qui peuvent potentiellement y appartenir. Les entreprises « potentiellement » dans la filière étant trop nombreuses pour toutes être interrogées exhaustivement, le recours à un échantillon était nécessaire.

Bien entendu, certaines entreprises ne répondront pas à l'enquête - même si le taux de réponse est très bon pour une enquête de ce type. Des techniques de correction de la non-réponse par repondération étaient envisagées.

Dès le début, un doute a été émis sur le fait que la propension à répondre des entreprises pouvait ne pas être la même si elles appartenaient ou non à la filière. L'idée pressentie étant qu'une entreprise hors de la filière serait moins portée à répondre car elle considérerait que le sujet de l'enquête ne la concerne pas. Ce qui conduirait à surestimer le nombre d''entreprises appartenant à la filière.

Le plan de sondage comprenait deux ensembles :

- une strate « cœur de la filière » interrogée exhaustivement composée des grandes entreprises (plus de 50 salariés) ainsi des entreprises dont l'activité était par nature incorporée à la filière (construction d'instruments de navigation) ;
- des strates « potentiellement dans la filière » échantillonnées composées des petites entreprises dont l'activité était potentiellement dans la filière (mécanique industrielle par exemple).

Beaucoup d'entreprises de la strate exhaustive avaient en outre l'habitude d'être interrogées dans une enquête auprès des fournisseurs et sous-traitants des constructeurs du secteur aéronautique et spatial. Cette enquête ressemblait beaucoup à l'enquête « filière aéronautique et spatiale » ; on pouvait donc penser qu'elles répondraient plus facilement.

4.2 La collecte et l'enquête auprès des non-répondants.

Une enquête téléphonique auprès des non-répondant a eu lieu pour déterminer s'ils étaient ou non dans la filière. Les résultats ont été traités séparément suivant que l'entreprise appartenait ou pas au "cœur de la filière".

Résultats sur le cœur de la filière

Cœur de la filière	Situation par rapport à la filière			Taux apparent d'appartenance à la filière.
Situation par rapport à l'enquête	Dans la filière	hors de la filière	Situation inconnue	
Répondantes	617	203	///	75 %
Non répondantes	196	80	280	71 %
Taux de réponse estimé (en répartissant les unités dont la situation est inconnue)	61 %	56 %	///	

Sur le cœur de la filière, on observe que les répondantes ou les non-répondantes appartiennent à la filière dans les mêmes proportions - peut-être plus fréquemment chez les répondants. Ce qui laisse penser que le taux de réponse dans la filière est légèrement supérieur, conformément à ce qui était anticipé.

Résultats sur les entreprises potentiellement dans la filière.

potentiellement dans la filière	Situation par rapport à la filière			Taux apparent d'appartenance à la filière.
Situation par rapport à l'enquête	Dans la filière	hors de la filière	Situation inconnue	
Répondantes	73	708	///	9 %
Non répondantes	39	222	180	15 %
Taux de réponse estimé (en répartissant les unités dont la situation est inconnue)	53 %	66 %	///	

On voit que pour ces entreprises, le comportement de réponse semble très différent entre les unités hors et dans la filière. Et surtout, les unités « hors de la filière » ont un taux de réponse beaucoup plus élevé que les unités « dans la filière ». Ce constat va à l'encontre de l'idée initiale comme quoi les entreprises « hors de la filière » se sentirait moins concerné par l'enquête mais il semble évident lorsqu'on met en avant la différence de temps de remplissage de questionnaires (plus d'une demi-heure si on est dans la filière et deux minutes sinon), sans compter le fait que ces entreprises n'avaient pas l'habitude d'être interrogées contrairement à celles du cœur de la filière. Il est à noter qu'un des mérites de l'enquête auprès des non-répondants a été de mettre en évidence un comportement de réponse contraire à ce que l'on attendait a priori.

Au vu de ces résultats, l'enquête sur les non-répondantes a donc été intégrée aux résultats pour les entreprises « potentiellement dans la filière ».

4.3 les redressements effectués.

Lors des redressements on a comparé une correction de la non-réponse par un système de groupes de réponse homogènes « classique » et un système de groupes intégrant l'enquête auprès des non-répondants.

Le fait d'intégrer l'enquête auprès des non-répondants contraint les groupes de réponses homogènes à utiliser moins d'information car il faut qu'ils soient assez « grands » pour être scindés en deux.

Ensuite, on a procédé à un calage sur marge des deux systèmes de poids - en retenant les mêmes marges.

La plus grande partie des agrégats vient des unités dans le « cœur de filière », nous extrayons ici la partie « potentiel » pour montrer l'impact des différentes méthodes de redressement.

	Méthode de redressement	**Nombre de salariés dédiés à la filière**	**Chiffre d'affaires réalisé avec la filière (millions d'euros)**	**Ecart entre les méthodes de redressement**	
				Sur le nombre de salariés	**Sur le chiffre d'affaires**
Avant calage sur marges	GRH sans prise en compte de l'ENR	23 464	3 590	3,7 %	4,0 %
	GRH avec prise en compte de l'ENR	24 325	3 733		
Après calage sur marges	GRH sans prise en compte de l'ENR	27 841	4 394	2,4 %	3,9 %
	GRH avec prise en compte de l'ENR	28 502	4 566		

On voit que l'impact de la méthode de redressement est moindre après réalisation d'un calage sur marge, surtout sur le nombre de salariés. Ce phénomène se reproduisait sur plusieurs variables d'intérêt. En tenant compte de la partie « cœur de la filière », la différence entre les deux méthodes était de l'ordre de 1 % seulement après calage sur marges.

Conclusion

La méthode proposée pour intégrer l'enquête sur les non-répondants - s'en servir pour évaluer le nombre de non-répondants dans des groupes de réponse homogènes construits sur une variable issue de l'enquête - présente les propriétés suivantes :

- elle supprime le biais dû au lien entre les variables d'intérêt et le comportement de réponse ;
- elle augmente la variance de l'estimateur - sauf si l'enquête auprès des non-répondants est quasiment exhaustive.

Dans le cadre d'une situation simplifiée, nous avons pu montrer qu'effectivement, il y a un seuil à partir duquel l'intégration de l'enquête auprès des non-répondants est bénéfique.

De la mise en œuvre dans l'enquête « filière aéronautique et spatiale dans le Grand Sud-Ouest », les points suivants méritent d'être retenus ;

- la recherche de groupes de réponse homogènes suffisamment importants pour être scindé suivant une variable d'intérêt oblige à restreindre les variables utilisées ;
- même si on ne l'intègre pas dans les résultats, l'enquête auprès des non-répondants à le mérite de montrer qu'il y a ou non une différence de comportement de réponse et de dire dans quel sens elle joue ;
- la prise en compte du calage sur marges réduit l'impact du choix de la méthode de redressement.

Bibliographie

[1] Deville J.C., « La correction de la non-réponse par calage généralisé », *Les actes des journées de méthodologie statistique*, 2002.
[2] Osier, G. « Traitement de la non-réponse non-ignorable par calage généralisé : une simulation à partir de l'enquête Budget des ménages au Luxembourg » *Les actes des journées de méthodologie statistique*, 2012..
[3] Durier S., Prost C, « L'enquête auprès des non-répondants à l'enquête Emploi », Sébastien Durier et Corinne Prost, *Les actes des journées de méthodologie statistique*, 2009
[4] Ardilly P. Les techniques de sondages , *Éditions Technip, 1994*
[5] Tillé Y. Théorie des sondages : échantillonnage et estimation en population finie, *Edition Dunod2001*