Une méthode pour adapter l'échantillonnage au cours de la collecte afin de mieux répondre à un objectif

David LEVY

INSEE, Direction régionale de Bretagne

Introduction

Dans le cadre des enquêtes auprès des ménages sur les déplacements réalisées par une Direction Régionale de l'Insee, il est souvent nécessaire de recruter des enquêteurs. Si la formation et l'encadrement des enquêteurs sont très suivis, ils ne suffisent pas à garantir complètement la qualité de la collecte. L'objectif d'un contrôle post collecte est de repérer le plus grand nombre possible d'anomalies tout au long de la collecte afin de rappeler aux enquêteurs les consignes à suivre.

Cet objectif se heurte à la rareté de l'anomalie de collecte donc difficilement mesurable correctement par un échantillon de taille raisonnable.

On présentera dans ce document une stratégie d'échantillonnage permettant de répondre à ces contraintes ainsi que les estimations obtenues et leur précision associée compte tenu du plan de sondage adopté. Cette stratégie prend en compte le fait que les anomalies se concentrent généralement sur un petit nombre d'enquêteurs. Elle a été appliquée à deux enquêtes sur les déplacements effectuées par la direction régionale de l'INSEE Rhône-Alpes, à Saint-Étienne (2000-2001) et à Grenoble (2001-2002).

1. La méthode du contrôle

1.1. Modalité du contrôle

Le contrôle consiste à interroger certains ménages, qu'ils aient répondu ou non à l'enquête. Pour les ménages répondants, on cherche à s'assurer que les principales règles de collecte ont été respectées et déceler une éventuelle fraude. Pour les ménages non-répondants, on vérifie que la non-réponse n'est pas le fait de l'enquêteur. Le contrôle se fait soit par courrier, soit par téléphone.
Les taux de retour de ce type de contrôle sont de l'ordre de 60% et se décomposent de la manière suivante pour les deux enquêtes pilotes :

	Saint Etienne (contrôle par téléphone principalement)	Grenoble (contrôle par courrier principalement)
Echantillon de l'enquête	6 500	10 400
Echantillon de contrôle	1 360	2 525
Taux de réponse global au contrôle	63%	58%
Taux de réponse des enquêtés ayant refusé l'enquête	*60%*	*40%*

Les contrôles de l'enquête de Saint-Etienne ont été réalisés en grande partie par téléphone. 63 % des ménages contrôlés ont répondu au contrôle. Parmi les ménages ayant refusé de répondre à l'enquête, 60% ont quand même accepté de répondre au contrôle.

1.2. Stratégie d'échantillonnage

Dans la pratique des contrôles, on a pu constater que les anomalies étaient très souvent concentrées sur un petit nombre d'enquêteurs. Par exemple, pour Grenoble, 10% des enquêteurs concentrent 40% des anomalies. Lorsque 150 à 200 enquêteurs travaillent sur une enquête, il n'est pas facile au travers d'un échantillon de contrôle de taille raisonnable de détecter ces anomalies. C'est pour cette raison que l'on essaiera d'orienter les contrôles vers des enquêteurs susceptibles de générer des anomalies. C'est l'idée des « choix raisonnés » de ménages à contrôler, choix qui s'appuient ici sur des critères objectifs et quantitatifs.

Dans ce contexte, on utilise la technique de l'échantillonnage adapté empruntée à S. Thompson *et al.*, 1996 [1]. Elle consiste à sélectionner des enquêtes à contrôler en fonction des résultats des précédents contrôles pour se concentrer sur les enquêteurs à l'origine des anomalies. De ce fait, l'utilisation d'une telle technique implique de redéfinir des estimateurs, car les estimateurs classiques sont biaisés avec une tendance à sur-estimer les estimations des taux d'anomalies puisqu'on centre les contrôles sur des enquêtes « à risque ».

1.3. Plan de sondage

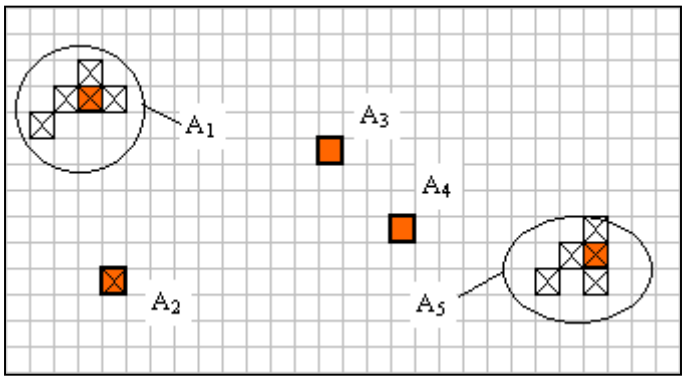
Un échantillon initial d'enquêtes à contrôler est tiré selon un plan simple sans remise. Si une anomalie est détectée, des unités "voisines" sont ajoutées à l'échantillon initial. Pour ce faire, on définit pour chaque unité un "voisinage" composé des enquêtes réalisées par le même enquêteur la même semaine, la semaine précédente et suivante. On poursuit ce procédé jusqu'à ce qu'il n'y ait plus d'anomalie dans les enquêtes ajoutées. On obtient ainsi un groupe d'anomalies.

Le groupe est constitué d'une enquête de l'échantillon de contrôle initial et de l'ensemble des unités voisines ajoutées le cas échéant. Si l'unité initiale est sans anomalie, le groupe est de taille1. Sinon, sa taille est supérieure à 1.

Cependant, certaines unités peuvent être ajoutées plusieurs fois, dès lors qu'elles appartiennent à plusieurs groupes. La notion de groupe doit donc être redéfinie si l'on veut obtenir une partition de la population.

Parmi toutes les unités tirées et ajoutées, on appelle unités marginales des unités ne présentant pas d'anomalie (unités saines). Chaque groupe est ainsi délimité par des unités marginales.

Graphique 1 : Constitution des groupes

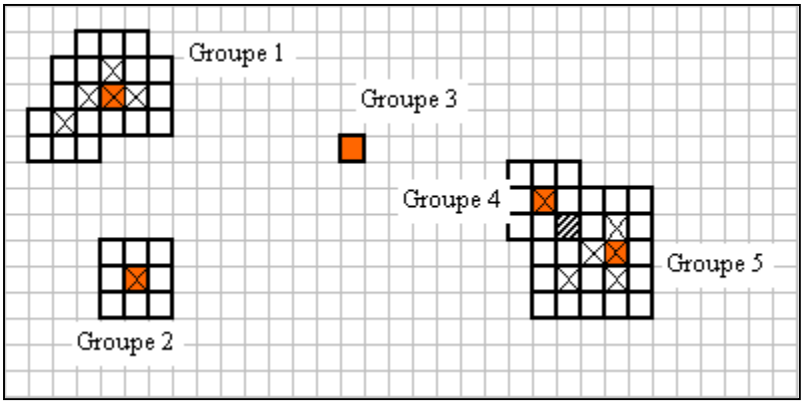


Note de lecture : Chaque carré représente une enquête, un carré est gris s'il fait partie de l'échantillon initial des contrôles à réaliser. Les croix représentent une anomalie détectée. Les carrés hachurés symbolisent des unités appartenant à plusieurs groupes.

Sur le graphique 1, le groupe 4 est constitué de 9 enquêtes dont une anomalie. Le groupe 5 est constitué de 22 enquêtes dont 5 anomalies. De plus, les groupes 4 et 5 ont une unité marginale commune représentée par des hachures. Les groupes ne constituent pas une bonne partition de la population puisqu'ils peuvent se chevaucher.

Pour remédier au problème posé par les groupes, on utilise la notion de réseau, défini par un groupe de taille supérieure à 1 auquel on a enlevé les unités marginales. Si la taille du groupe est de 1, le réseau se limite à une enquête sans anomalie. La taille minimale d'un réseau est 1 (une enquête sans anomalie ou une enquête avec anomalie isolée). De plus, la sélection d'une unité d'un réseau implique nécessairement la sélection de toutes les unités de ce réseau. Ainsi, l'ensemble Ai des réseaux constitue une partition de la population, contrairement aux groupes, comme l'illustre le graphique 2.

Graphique 2 : Constitution des réseaux



Constitution des réseaux d'unités à partir des enquêtes contrôlées initialement (carré gris) et des contrôles supplémentaires et positifs (croix).

1.4. Formalisation du problème - Estimateur utilisé

1.4.1. Estimateur

Pour ne pas avoir d'estimateur biaisé, on considère comme unité les réseaux et non plus les enquêtes. L'estimateur d'Horvitz-Thompson est alors :

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{k} \frac{y_i}{\alpha_i}$$

où :

y_i est le nombre d'anomalies dans le réseau i

α_i est la probabilité d'inclusion des unités du réseau i

N est la taille de la population

k est le nombre total de réseaux dans l'échantillon

On note par ailleurs :

m_i, le nombre d'unités du réseau i

a_i, le nombre d'unités des réseaux pour lesquels i est une unité marginale

n_1, la taille de l'échantillon initial

Le problème consiste maintenant à définir les probabilités d'inclusion. Pour cela, on connaît le nombre total d'échantillons ne contenant pas i, donné par :

$$C_{N-m_i-a_i}^{n_1}, \text{ où } C_N^n = \frac{N!}{n!(N-n)!}$$

La probabilité d'inclusion est alors définie par :

$$\pi_i = 1 - \left[C_{N-m_i-a_i}^{n_1} \frac{1}{C_N^{n_1}} \right]$$

Si tous les m_i sont connus, en revanche les a_i ne sont pas tous connus sur l'échantillon. En effet, toutes les unités pour lesquelles une enquête est une unité marginale ne sont pas forcément échantillonnées. Il n'est donc pas possible de calculer tous les π_i. Thompson et Seber contournent le problème en proposant d'utiliser les probabilités d'inclusions "partielles" suivantes :

$$\boxed{\alpha_i = 1 - \left[C_{N-m_i}^{n_1} \frac{1}{C_N^{n_1}} \right]}$$

Ces probabilités peuvent s'interpréter comme la probabilité d'intersection entre l'échantillon initial et le réseau A_i.

1.4.2. Variance estimée

Un estimateur de la variance de l'estimateur est donné par C.E. Särndal et al., 1992 [2]. En utilisant les probabilités d'inclusion définies plus haut, l'estimation de cette variance est :

$$\hat{V}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{i}^{k} y_i^2 \frac{1-\alpha_i}{\alpha_i^2} + \sum_{i=1}^{k}\sum_{j\neq 1}^{k} y_i y_j \frac{\alpha_{ij}-\alpha_i\alpha_j}{\alpha_i\alpha_j\alpha_{ij}}\right]$$

ou encore :

$$\hat{V}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{y_i y_j}{\alpha_{ij}}\left(\frac{\alpha_{ij}}{\alpha_i\alpha_j} - 1\right)\right]$$

Pour calculer cet estimateur on a besoin des probabilités d'inclusion d'ordre deux :

$$\alpha_{ij} = \alpha_i + \alpha_j - \left(1 - \Pr[i \cap j]\right)$$

En adoptant la même approximation que celle utilisée pour le calcul de α_i, on a :

$$\alpha_{ij} = 1 - \left[C_{N-m_i}^{n_1} + C_{N-m_j}^{n_1} - C_{N-m_i-m_j}^{n_1}\right]\frac{1}{C_N^{n_1}}$$

2. Applications

2.1. L'enquête déplacements de Saint-Étienne

Un échantillon initial de 1 360 enquêtes à contrôler est tiré avant le début de la collecte selon un plan de sondage aléatoire simple avec un taux de sondage de 20%. Chaque semaine, les contrôles par téléphone sont déclenchés, en fonction de la réalisation des enquêtes. Les contrôles portent ainsi aussi bien sur les ménages répondants que sur les non-répondants et sur les logements hors champ (résidence secondaire et vacante). Dans ce dernier cas, le contrôle se fait directement sur le terrain.

Dès qu'une anomalie suffisamment importante est détectée et confirmée par les responsables de la collecte, on déclenche automatiquement le contrôle des enquêtes réalisées par le même enquêteur la semaine précédente, la semaine en cours et la semaine suivante (un enquêteur réalise en moyenne 3 enquêtes par semaine). Au total, près de 150 ménages supplémentaires ont été contrôlés parmi lesquels 92 ont répondu.

Sur les 1 360 enquêtes à contrôler initialement, seul 851 ménages ont accepté de répondre. En l'absence de réflexion plus poussée sur la non-réponse, on considère par la suite que ces répondants forment également un échantillon aléatoire.

Le taux d'anomalie simple est de 8,7% sur l'échantillon initial et de 17% sur l'échantillon supplémentaire. L'estimateur adapté du taux d'anomalie donne une estimation de 6,6%. Pour comparer la performance de cet estimateur, on le compare à l'estimateur simple calculé sur l'échantillon initial, comme s'il n'y avait pas eu d'adaptation.

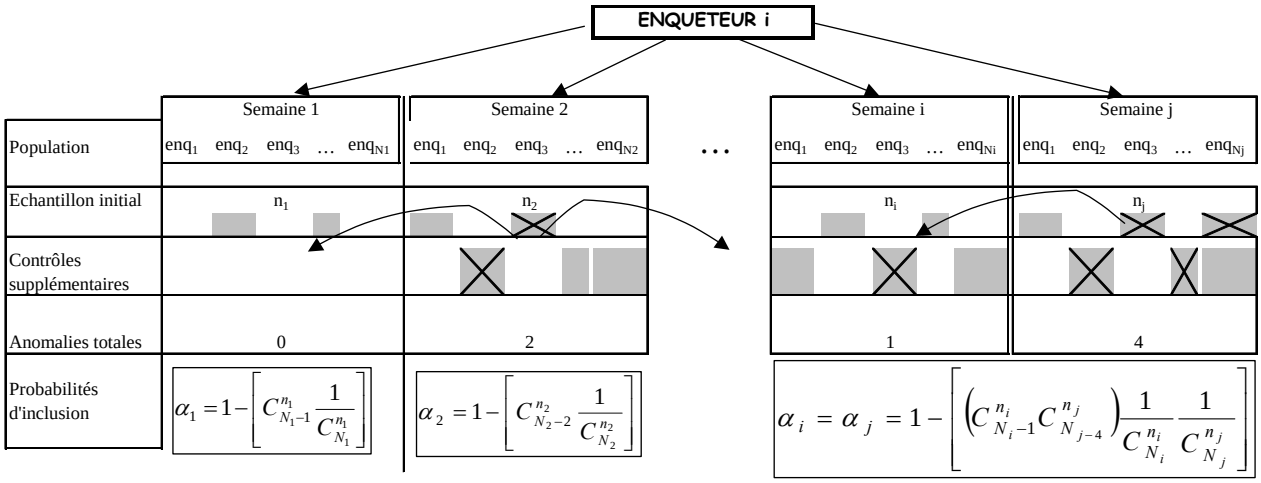
	Estimateur simple	Estimateur adapté
Nombre de répondants	851	943
Taux d'anomalie estimé	8,7%	6,7%
Intervalle de confiance à 95%	[6,5 ; 10,9]	[6,5 ; 6,9]

2.2. L'enquête déplacements de Grenoble : échantillonnage adapté et plan stratifié

A la différence de l'enquête précédente, le plan de sondage des contrôles a été **préalablement stratifié** par enquêteur et par semaine d'enquête car les anomalies sont constatées principalement lors des premières semaines d'enquêtes. On a donc contrôlé davantage d'enquêtes les premières semaines de chaque enquêteur. Comme les enquêteurs ne commencent pas tous en même temps la collecte en raison des démissions/recrutements qui interviennent tout au long de la collecte, il est important de gérer les contrôles par enquêteur.

Ceci introduit une contrainte supplémentaire dans l'organisation du contrôle : l'échantillon doit être tiré chaque semaine. La stratégie pour adapter l'échantillon est la même que celle utilisée précédemment. Cette stratégie conduit à créer des réseaux d'anomalies à cheval sur plusieurs strates.

Graphique 3 : Exemple d'adaptation pour un enquêteur i



En semaine 2, une anomalie est détectée dans l'échantillon initial. Les enquêtes voisines - celle réalisée dans les semaines 1, 2 et 3 - sont toutes contrôlées. Une seule anomalie est détectée dans cet échantillon supplémentaire, elle appartient à la semaine 2.

En semaine j, 2 anomalies sont détectées, entraînant le contrôles des enquêtes voisines. 3 anomalies sont de nouveau détectées, l'une en semaine i - donc dans une autre strate - et 2 en semaine j.

Sur les 2 525 ménages contrôlés dans l'échantillon initial, 1 463 ont répondu avec un taux d'anomalies observées de 11%. 91 ménages supplémentaires ont été contrôlés parmi lesquels 52 ont répondu avec un taux d'anomalies de 25%.

Le taux de contrôles supplémentaires est beaucoup plus faible que précédemment (3,5% contre 11%) en raison d'une définition des anomalies plus restrictive. Cependant, le plan de sondage semble plus efficace puisque le taux d'anomalies de l'échantillon supplémentaire est supérieur (25% contre 17%). Autrement dit, on a mieux repéré les concentrations d'erreurs. Les résultats des estimateurs sont donnés dans le tableau suivant :

	Estimateur simple	Estimateur adapté
Nombre de répondants	1 463	1 515
Taux d'anomalie estimé	11,3%	10,1%
Intervalle de confiance à 95%	[9,8 ; 12,8]	[9,6 ; 10,5]

Conclusion

Dans les deux exemples, la stratégie pour orienter l'échantillon à permis de répondre à l'objectif fixé, à savoir repérer le plus grand nombre possible d'anomalies et fournir une estimation fiable de ce taux.
Cette stratégie est d'autant plus justifiée et efficace que le taux d'anomalies des échantillons supplémentaires est bien supérieur à celui des échantillons initiaux, c'est à dire quand il y a une forte concentration des anomalies graves ou répétées sur quelques enquêteurs.

La plupart des anomalies repérées ont été corrigées et ont donné lieu à des échanges avec les enquêteurs sur les consignes de collecte. De plus, cette procédure de contrôles a permis de réduire les problèmes de collecte, si bien qu'ils sont devenus très rares en fin de collecte.

Bibliographie

[1] THOMPSON, S., SEBER, G. (1996), Adaptive Sampling, Wiley, New York.

[2] SÄRNDAL, C.E., SWENSON, B., and WRETMAN, J. (1992), *Model Assisted Survey Sampling,* Springer, New York.