JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

JEAN BOUYER

Les enquêtes d'observation en épidémiologie et en santé publique : quelques problèmes et alternatives à l'échantillonnage aléatoire

Journal de la société statistique de Paris, tome 123, n° 4 (1982), p. 238-252

<http://www.numdam.org/item?id=JSFS_1982__123_4_238_0>

II

ARTICLES

LES ENQUÊTES D'OBSERVATION EN ÉPIDÉMIOLOGIE ET EN SANTÉ PUBLIQUE : QUELQUES PROBLÈMES ET ALTERNATIVES A L'ÉCHANTILLONNAGE ALÉATOIRE

Jean BOUYER

Unité de recherches épidémiologiques et statistiques sur l'environnement et la santé, INSERM U 170

Dans les enquêtes d'observation, l'échantillonnage aléatoire se heurte à des difficultés (absence de listes, présence de covariables non contrôlables). Ces problèmes peuvent être résolus par des méthodes statistiques classiques. On montre ici qu'ils peuvent l'être aussi, et parfois mieux, par un échantillonnage non aléatoire bien choisi. Pour cela, on examine successivement la question des tests — comparaison des performances de l'appariement, de l'ajustement et de l'analyse de covariance —, puis celle de l'estimation — confrontation de l'échantillonnage de l'I.N.S.E.E., des quota et d'une méthode d'échantillonnage non aléatoire « validée par les résultats obtenus ».

In observation surveys, random sampling runs into difficulties (absence of lists, existence of non verifiable covariables). These problems can be solved through conventional statistical methods. We prove here that they can also be solved, and sometimes better, through well selected non random sampling. To this end, we consider, in turn, the problem of tests — comparison between matching, adjustment and covariance analysis performances — then, the estimation problem — confrontation of the I.N.S.E.E. sampling, with quotas and a non random sampling method validated by the results obtained.

Dans les enquêtes qui sont à la base de nombreuses recherches, en particulier en épidémiologie et en santé publique, le choix des sujets étudiés est décisif. Il conditionne les capacités à résoudre le problème initial, c'est-à-dire le plus souvent répondre à des questions qu'on se pose au niveau de toute la population.

Nous supposerons que cette population de référence est bien définie. C'est tout à fait le cas dans les enquêtes, dont l'exemple type est celles que réalise l'I.N.S.E.E., dirigées vers des populations délimitées par leur lieu d'habitat (ville, région...). Cela ne va pas toujours sans difficultés dans les enquêtes épidémiologiques, par exemple quand il s'agit de cerner la population correspondant à l'ensemble des malades soignés dans un hôpital [16].

Au-delà de la définition de la population de référence, les enquêtes épidémiologiques se particularisent très souvent par leur situation d'observation de populations humaines. C'est ce point et ses conséquences sur l'échantillonnage que nous allons examiner ici.

Dans les enquêtes d'observation, il n'est pas toujours possible de trouver des échantillons aussi « purs » qu'on le souhaiterait : des variations incontrôlées peuvent accompagner le facteur étudié et rendre difficile son évaluation ou la mesure de son effet.

Journal de la Société de statistique de Paris, tome 123, n° 4, 1982.

Il arrive ainsi que des échantillons aléatoires ne soient pas les meilleurs pour mener à son terme la recherche entreprise, ou même conduisent à des biais :

- soit parce qu'il n'existe pas de liste de la population à échantillonner ou que celle-ci n'est pas à jour; cela conduit à des biais d'échantillonnage;
- soit parce qu'on ne peut pas constituer des groupes-échantillons qui ne diffèrent que par le traitement ou l'exposition. Les groupes ne sont pas alors comparables vis-à-vis de certaines covariables (qui sont intrinsèquement liées au phénomène étudié); ce qui peut aussi être à l'origine de biais.

Ces deux obstacles peuvent être surmontés par des méthodes statistiques appropriées — analyse de la covariance, pondération, redressement, ... — au moment de l'analyse des résultats. L'objet de cet article est de montrer qu'ils peuvent l'être par le choix de méthodes d'échantillonnage non aléatoire, adaptées au problème à résoudre.

Pour cela, nous allons examiner quelques procédés d'échantillonnage non entièrement aléatoire et comparer leurs « performances » avec celles du strict tirage au sort.

La pratique de l'échantillonnage et la notion de représentativité

Il est généralement admis qu'un échantillon est acceptable s'il est « représentatif ». Le plus souvent, ce terme est traduit par « tiré au sort ». Ce qui est tout à fait naturel puisque cela correspond au modèle probabiliste qui sous-tend la théorie des statistiques.

Cependant, dans les enquêtes effectivement réalisées, la pratique de l'échantillonnage est beaucoup plus diversifiée. L'examen des principales revues d'épidémiologie [3] montre, qu'à côté des enquêtes basées sur des échantillons aléatoires, on trouve, sans que cela puisse faire figure d'exception, des enquêtes dont le mode d'échantillonnage s'éloigne du tirage au sort : échantillon tout venant, échantillon validé par des études antérieures, échantillon empirique, ...

Les raisons de cet écart à la théorie de l'échantillonnage sont souvent à chercher dans l'obligation de s'adapter aux contraintes propres aux enquêtes d'observation : limitation des données disponibles, non-contrôle du « traitement », coût... Il est également probable que l'on retrouve là l'ambiguïté du terme « représentativité » lui-même. Dans une étude récente basée sur de nombreuses publications, dont l'ensemble des publications statistiques, Kruskal et Mosteller [10] ont ainsi dénombré neuf sens du même terme « échantillon représentatif ».

L'ensemble de ces constats invite donc à formaliser des procédés d'échantillonnage qui, dans la pratique, ont montré leurs aptitudes à surmonter certaines des difficultés inhérentes aux enquêtes d'observation.

Un pas dans ce sens peut être fait en étudiant leurs capacités à réduire le biais dû à une covariable ou à suppléer à l'absence d'une liste fiable de la population.

Il convient de traiter séparément ce qui concerne les tests et ce qui concerne l'estimation, essentiellement pour deux raisons :

- pour juger de la comparabilité de deux échantillons, dans le cas d'un test, on peut disposer de toute une série de mesures sur les deux échantillons. Cela n'est, en général, pas possible (ou de façon très grossière) dans le cas d'une estimation et de la représentativité par rapport à l'ensemble de la population;
- pour un test, le souci principal est de tenir compte de l'effet « confondant » de covariables identifiées. Pour une estimation, l'existence d'un biais est souvent le reflet d'une distorsion de l'échantillon vis-à-vis de covariables inconnues ou impossibles à préciser à cause de leur caractère multiforme.

LES TESTS

De nombreux moyens peuvent être envisagés pour prendre en compte des covariables dans la comparaison de deux groupes de sujets.

Pour apprécier dans quelle mesure il est possible d'y parvenir en intervenant dès la constitution des échantillons, nous allons examiner les méthodes les plus courantes concernant la prise en compte de covariables quantitatives : l'analyse de la covariance et l'ajustement — méthode de type purement aléatoire, puisque les échantillons sont tirés au sort dans la population de référence —, et l'appariement — qui fait appel à un certain empirisme puisque le choix des variables d'appariement est fait *a priori* et détermine la constitution des échantillons (*).

Jusqu'alors, ces trois techniques n'ont pas été comparées globalement sur un ensemble commun de critères. Nous allons, dans ce qui suit, en envisager trois : l'efficacité pour diminuer le biais, la précision et la facilité d'utilisation.

Notations

La variable observée est notée Y et la covariable X. Les populations comparées sont P_1 et P_2.

On supposera dans la suite que X est distribuée normalement : $X = N(m_i, \sigma_i^2)$ dans P_i $(i = 1,2)$ avec, sans perte de généralité, $m_1 + m_2 = 0$ et $\sigma_1^2 + \sigma_2^2 = 2$.

La régression de Y sur X, supposée la même dans les deux populations, est notée :

$$Y = \alpha_i + u(x) + \varepsilon \quad (i = 1, 2) \text{ où } \varepsilon = N(O, \sigma_\varepsilon^2)$$

La différence cherchée entre les populations P_1 et P_2 est ainsi : $d = \alpha_1 - \alpha_2$.

Si on l'estime par $\delta = \bar{y}_1 - \bar{y}_2$, il y a un biais (initial) qui vaut, en moyenne :

$$B_1 = \int u(x)[\Phi_1(x)]\,dx \text{ où } \Phi_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\frac{(x - m_i)^2}{\sigma_i^2}\right]$$

Pour diminuer, ou éliminer ce biais, on peut envisager deux possibilités :

- estimer d par une quantité autre que δ qui tienne compte de la différence entre les distributions de X dans les deux populations. Ce seront les méthodes d'analyse de la covariance et d'ajustement;
- ne pas prendre des échantillons aléatoires de P_1 et P_2, mais les choisir de sorte que les distributions de X aient, dans ces échantillons, des densités de probabilités Φ'_1 et Φ'_2 et que le biais final, $B_F = \int u(x)[\Phi'_1(x) - \Phi'_2(x)]\,dx$, soit plus petit. Ce sera la méthode d'appariement : on constitue d'abord un échantillon aléatoire E_1, de taille N, de la population P_1 et un échantillon aléatoire E_2, de taille rN $(r > 1)$, de la population P_2. Puis on extrait de E_2 un sous-échantillon $E_{2'}$ de taille N. d est estimé par $\bar{y}_1 - \bar{y}_{2'}$. [15].

Dans chaque cas, le gain obtenu sera mesuré par le pourcentage de réduction de la valeur absolue du biais :

$$\theta = 100\left[1 - \frac{|B_F|}{|B_I|}\right]$$

Une valeur négative de θ indique donc une augmentation du biais.

(*) On trouvera une description de ces méthodes dans : (11) pour l'analyse de la covariance, (7) pour l'ajustement et (15) pour l'appariement.

Efficacité pour diminuer le biais

Dans les enquêtes d'observation, c'est une des premières qualités demandées aux méthodes d'échantillonnage et d'analyse. Elle est mesurée ici par θ, dont les valeurs numériques sont calculées en fonction des paramètres suivants :

- $B = \dfrac{m_1 - m_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} = 2\ m_i$ et σ_1^2/σ_2^2 qui reflètent l'écart entre les distributions de X dans les deux populations;
- la forme de la régression : u (x) = kx (linéaire) et u (x) = e^{ax} avec a = $\pm$ 1/2 et a = $\pm$ 1/2 (exemples de régressions non linéaires).
- dans le cas de l'appariement, le rapport r des tailles des échantillons initiaux E_1 et E_2;
- dans le cas de l'ajustement, le nombre c des classes constituées.

TABLEAU I

Pourcentages de réduction au biais avec l'appariement — 1. u(x) = kx

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$			$\sigma_1^2/\sigma_2^2 = 2$		
	B / r	1/4	1/2	1	1/4	1/2	1	1/4	1/2	1
N = 25	2	97	94	80	87	82	66	63	60	48
	3	99	98	93	94	91	81	77	72	61
	4	99	99	97	95	95	88	81	79	68
N = 50	2	99	98	84	92	87	69	66	59	51
	3	100	99	97	96	95	84	79	75	63
	4	100	100	99	98	97	89	86	81	71
N = 100	2	100	99	86	95	90	69	67	59	49
	3	100	100	98	99	96	86	81	75	64
	4	100	100	99	99	98	90	85	81	71

TABLEAU II

Pourcentages de réduction du biais avec l'appariement — 2. u(x) = e^{ax}; N = 50

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$			$\sigma_1^2/\sigma_2^2 = 2$		
	B / r	1/4	1/2	1	1/4	1/2	1	1/4	1/2	1
Exp. (x)	2	94	93	69	70	60	39	35	30	16
	3	97	94	90	79	75	55	51	43	28
	4	98	97	94	87	84	65	55	48	29
Exp. $\left(\frac{x}{2}\right)$	2	95	96	76	83	74	53	48	45	31
	3	99	98	94	90	87	70	66	60	45
	4	99	99	97	94	92	79	70	65	50
Exp. $\left(-\frac{x}{2}\right)$	2	99	99	91	99	94	82	79	81	67
	3	100	100	98	99	98	92	61	89	79
	4	100	100	99	99	99	96	80	94	84
Exp. (− x)	2	100	99	96	94	98	81	− 50	77	83
	3	100	100	99	96	100	97	− 48	82	92
	4	100	100	100	97	100	99	01	74	94

Les tableaux I et II (tirés de Rubin [15]) donnent les valeurs de θ obtenues par appariement, suivant la forme de la régression.

Les résultats sont d'autant meilleurs que N et r sont plus grands et que B et σ_1^2/σ_2^2 sont plus petits. Dans le cas, fréquent en pratique, où $\sigma_1^2/\sigma_2^2 = 1$, on voit que : dès que $r \geq 3$, on a $\theta \geq 80$ % dans le cas linéaire et $\theta \geq 70$ % quand l'écart à la linéarité n'est pas trop fort ($a = \pm 1/2$) ou quand la différence entre m_1 et m_2 est assez petite.

L'analyse de la covariance élimine totalement le biais quand la régression est linéaire. S'il y a écart à la linéarité, on peut montrer (voir annexe A), en prenant comme précédemment $u(x) = e^{ax}$, que :

$$\theta = 100\left[1 - \frac{|e^v(1 - e^{a^2 - 2v}) - bB|}{|e^v(1 - e^{a^2 - 2v})|}\right]$$

où

$$v = \frac{1}{2}\left[aB + 2a^2\frac{\sigma_1^2/\sigma_2^2}{1 + \sigma_1^2/\sigma_2^2}\right]$$

Le tableau III donne les valeurs numériques de θ et montre que celles-ci s'éloignent sensiblement de 100 % dès que σ_1^2 et σ_2^2 sont différents.

$$b = \left[a\frac{\sigma_1^2}{\sigma_2^2}e^v + a\,e^{aB + v}\right]\frac{1}{1 + \sigma_1^2/\sigma_2^2}$$

TABLEAU III

Pourcentages de réduction du biais par analyse de la covariance — $u(x) = e^{ax}$

	B	Exp. (x)	Exp. $\left(\frac{x}{2}\right)$	Exp. $-\left(\frac{x}{2}\right)$	Exp. (− x)
$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$	1/4	− 304	− 98	62	48
	1/2	− 92	54	80	72
	1	61	87	96	97
$\sigma_1^2/\sigma_2^2 = 1$	1/4	99	100	100	99
	1/2	98	99	99	98
	1	92	98	98	92
$\sigma_1^2/\sigma_2^2 = 2$	1/4	48	62	− 98	− 304
	1/2	72	80	54	− 92
	1	97	96	87	61

Enfin, dans les tableaux IV et V figurent les valeurs de θ obtenues par ajustement (l'expression analytique de θ est donnée en annexe B). Lorsque la régression est linéaire, et pour c fixé, les valeurs de θ sont quasiment indépendantes de B et de σ_1^2/σ_2^2. . Elles sont comparables à celles obtenues par appariement quand $c = 6$ et $\sigma_1^2/\sigma_2^2 = 1$, . Mais, en dehors de ce cas, les résultats sont très irréguliers et font de l'ajustement une méthode peu stable.

TABLEAU IV

Pourcentages de réduction du biais par ajustement — $u(x) = k(x)$

B / σ_1^2/σ_2^2		1/4	1/2	1
$\frac{1}{2}$	C = 2	64	63	62
	C = 4	86	85	82
	C = 6	92	91	91
1	C = 2	64	63	62
	C = 4	87	86	85
	C = 6	92	92	92
2	C = 2	63	63	61
	C = 4	86	86	84
	C = 6	91	92	90

TABLEAU V

Pourcentages de réduction du biais par ajustement — $u(x) = e^{ax}$

	B	Exp. (x)		Exp. $\left(\frac{x}{2}\right)$		Exp. $-\left(\frac{x}{2}\right)$		Exp. (− x)	
		C = 2	C = 6	C = 2	C = 6	C = 2	C = 6	C = 2	C = 6
$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$	1/4	− 184	− 68	10	34	37	75	27	67
	1/2	21	37	93	27	47	82	37	74
	1	87	92	74	99	53	85	46	80
$\sigma_1^2/\sigma_2^2 = 1$	1/4	54	85	61	91	61	89	55	84
	1/2	53	84	60	90	61	90	55	85
	1	51	84	59	90	60	89	55	86
$\sigma_1^2/\sigma_2^2 = 2$	1/4	20	56	35	72	22	31	− 147	− 74
	1/2	28	62	44	77	89	89	53	43
	1	33	65	49	77	66	96	73	100

Il apparaît qu'aucune des méthodes n'est supérieure aux autres dans tous les cas envisagés. Même si, chacune dans « son domaine » élimine la totalité ou la plus grande partie du biais initial et surpasse les autres : l'analyse de la covariance quand la régression de Y sur X est linéaire ou quand $\sigma_1^2/\sigma_2^2 = 1$. l'appariement dans les autres cas.

Toutefois, en l'absence d'une certitude de la linéarité de cette régression, l'appariement est la méthode qui, « en moyenne », donne les meilleurs résultats sans être jamais catastrophique.

D'autre part, la complémentarité des « performances » de l'appariement et de l'analyse de la covariance invite à combiner ces deux méthodes. Les résultats obtenus sont donnés dans le tableau VI.

TABLEAU VI

Pourcentages de réduction du biais avec analyse de la covariance sur des échantillons appariés

$u(x) = e^{ax}$; $N = 50$

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$			$\sigma_1^2/\sigma_2^2 = 1$			$\sigma_1^2/\sigma_2^2 = 2$		
	B / r	1/4	1/2	1	1/4	1/2	1	1/4	1/2	1
Exp. (x)	2	96	92	82	100	94	87	90	93	73
	3	99	99	95	100	98	91	92	95	81
	4	100	100	98	100	99	94	94	96	81
Exp. $\left(\frac{x}{2}\right)$	2	98	100	95	99	98	96	90	100	91
	3	100	100	99	100	100	97	93	100	94
	4	100	100	99	100	100	98	95	100	95
Exp. $\left(-\frac{x}{2}\right)$	2	100	100	99	97	100	99	23	89	99
	3	100	100	99	99	100	100	51	92	100
	4	100	100	100	99	100	100	60	93	99
Exp. (− x)	2	100	100	99	92	99	99	− 99	29	96
	3	100	100	100	97	99	100	− 25	53	96
	4	100	100	100	98	99	99	− 05	74	98

Précision

Pour étudier ce point, directement en relation avec la puissance des tests ultérieurs, il faut calculer la variance de δ pour chacune des méthodes.

Dans ce qui suit, n est la taille des échantillons sur la base desquels est fait le test.

a) X n'est pas prise en compte

$\text{var } \delta = \frac{2}{n} \sigma_y^2$ (où σ_y^2 est la variance de Y, supposée la même dans les populations P_1 et P_2).

b) Appariement

$\text{var } \delta = \frac{2}{n} \sigma_{y.x}^2 = \frac{2}{n} \sigma_y^2 (1 - \rho^2)$ où ρ est le coefficient de corrélation entre X et Y.

c) Analyse de la covariance

ici $\delta = \bar{y}_1 - \bar{y}_2 - b (\bar{x}_1 - \bar{x}_2)$. On obtient donc (9) :

$$\text{var } \delta = \frac{2}{n} \sigma_{y.x}^2 \left(1 + \frac{n (x_1 - x_2)^2}{2 \Sigma (x - \bar{x})^2}\right)$$

La moyenne de cette expression sur des échantillons successifs et pour n grand s'écrit (6) :

$$\overline{\text{var } \delta} = \frac{2}{n} (1 - \rho^2)^2 \left(1 + \frac{B^2}{4}\right) \sigma_y^2$$

d) Ajustement

$\text{var } \delta = \frac{2}{n} \sigma_y^2 (1 - \rho^2) K$ (voir annexe B).

Pour les différentes valeurs de B, σ_1^2/σ_2^2 et c envisagées dans les calculs précédents de θ, les termes K et $\left(1 + \frac{B^2}{4}\right)$ ont des valeurs numériques sensiblement égales.

e) Combinaison analyse de la covariance - appariement

$$\overline{\text{var } \delta} = \frac{2}{n} \sigma_y^2 (1 - \rho^2)^2 \left(1 + \frac{B^2}{4}\right)$$

L'appariement apparaît donc plus précis que les deux autres méthodes. On peut cependant penser qu'en pratique les différences de précision sont faibles. En effet, le rapport de précision, $\left(1 + \frac{B^2}{4}\right)$, n'est notablement supérieur à 1 que si B est grand (c'est-à-dire si les distributions de la covariable diffèrent de façon importante d'une population à l'autre), cas qu'on cherche à éviter en pratique.

Difficultés de mise en œuvre

Ce critère est tout à fait décisif, car la supériorité théorique d'une méthode est d'un piètre intérêt si les difficultés à la mettre en œuvre sont trop grandes.

L'appariement peut paraître plus contraignant car il faut choisir *a priori* les covariables intéressantes, sans disposer des résultats complets de l'enquête. (Il ne faut cependant pas oublier que, pour prendre en compte une covariable par ajustement ou analyse de la covariance, il faut avoir pensé à la mesurer dès le début de l'enquête.) Mais surtout, un mauvais choix de la variable d'appariement est difficilement rectifiable à l'analyse et peut conduire à des conclusions erronées (c'est ce que soulignent Breslow et Day (4) en mettant en garde contre « l'over-matching »).

Un des arguments les plus forts contre l'appariement reste cependant la « perte » de (r − 1) N sujets imposée par cette méthode. Ce « coût » a été quantifié par Billewicz (2) et Mc Kinlay (13) mais de telle sorte que leurs résultats sont un majorant de la perte de sujets. Nous avons vu précédemment, avec la méthode d'appariement proposée par Rubin (15), que dans le cas d'une seule cova-

riable r = 3 ou 4 suffit. De plus, cette difficulté réelle s'estompe dans les enquêtes, fréquentes en épidémiologie, où le problème est de trouver suffisamment de sujets malades (ou exposés à un risque) mais où on dispose d'un grand nombre de témoins potentiels.

Conclusions

Les principales conclusions concernant les méthodes étudiées peuvent être résumées ainsi :

- La précision n'est pas un critère discriminant. Les variances des estimateurs de d sont en effet sensiblement égales.
- Le choix entre les différentes méthodes repose sur :
  - — la linéarité de la régression de Y sur X;
  - — le rapport des variances de X dans les deux groupes étudiés.
- L'appariement réduit mieux le biais que l'analyse de la covariance si la régression est non linéaire. En contrepartie, cette méthode pose plus de problèmes de mise en œuvre. Si elle est choisie, il est préférable de la combiner avec une analyse de la covariance.

Pour répondre aux faiblesses respectives des différentes méthodes, des améliorations peuvent être (et ont été) proposées. Ainsi, on peut envisager un appariement avec plusieurs témoins pour chaque « cas » (moins de perte de sujets, moins de risque de « veuvage » pour une paire). De même, une régression quadratique (ou avec un polynôme de degré supérieur) peut être préférable à une analyse de la covariance linéaire.

Le gain en performance est d'autant meilleur que de tels « raffinements » sont choisis en fonction de la nature du problème posé.

Cependant, la comparaison des méthodes générales présentées ici permet à elle seule d'affirmer que la prise en compte de covariables par un échantillonnage approprié (et donc adapté aux objectifs de l'enquête) est souvent aussi efficace — et parfois plus — que l'analyse statistique classique. C'est dire, pour revenir à notre problème initial, qu'à chaque enquête, il faut se demander si le tirage au sort est la meilleure méthode d'échantillonnage.

Les résultats donnés ci-dessus doivent permettre de répondre à cette question en fonction de chaque cas particulier.

L'ESTIMATION

De même que dans la partie précédente, à propos des tests, il est impossible pour l'estimation d'entrer dans le détail de toutes les méthodes d'échantillonnage envisageables suivant l'enquête à mener.

Les qualités du tirage au sort étant bien connues — et non remises en cause ici —, nous allons illustrer, dans cette partie, les difficultés qu'il entraîne et les possibilités de s'en passer par trois exemples : le plan de sondage des enquêtes de l'I.N.S.E.E., « référence » de l'échantillonnage aléatoire; la méthode des quota, exemple type d'échantillonnage empirique; un type d'échantillonnage fondé sur un point de vue différent et validé par les résultats obtenus.

Les enquêtes de l'I.N.S.E.E.

Les enquêtes de l'I.N.S.E.E. auprès des ménages sont réalisées par tirage au sort de logements au sein d'un échantillon maître. Ce dernier est lui-même obtenu par échantillonnage aléatoire (à un, deux ou trois degrés selon l'importance des communes) des fiches-logements du recensement. L'échan-

tillon maître est disponible environ deux ans après le recensement. Il permet que le tirage au sort d'un échantillon pour chaque enquête soit ensuite — heureusement ! — beaucoup plus rapide.

Ce système, qui est probablement le seul possible, a l'inconvénient essentiel de figer l'image de la population pour toute la période inter-censitaire (environ 7 ans), malgré la prise en compte des logements neufs à chaque enquête. En particulier, les caractéristiques des logements et de leurs habitants dont on dispose datent de l'année du recensement. La mobilité des populations interdit donc d'échantillonner sur les individus et impose de se satisfaire d'un échantillon de logements. Même si les caractéristiques du lieu d'habitat sont liées à nombre de variables concernant l'individu, il peut se créer des distorsions de l'échantillon maître, et par suite de l'échantillon d'enquête, au niveau individuel dont il faudra tenir compte en exploitant les résultats.

D'autre part, un certain nombre de gens ne répondent pas (refus ou absence). Dans les enquêtes de l'I.N.S.E.E., le taux de non-répondants est en général de l'ordre de 10 %.

Il faut donc redresser les résultats obtenus :

- au niveau de l'analyse statistique par une pondération des sujets (voir par exemple le redressement de l'« enquête santé 1970 » (12));
- au niveau des réponses individuelles en attribuant à un non-répondant les résultats d'un sujet de mêmes caractéristiques (5).

Ces résultats sont fondés sur une image de la population (fournie par l'« enquête emploi » bisannuelle) qui ne peut être qu'imparfaite. Ils s'appuient sur l'hypothèse — seule possible, bien qu'incontrôlable de façon certaine — d'une identité moyenne entre les répondants et les non-répondants.

Il reste donc inévitablement des biais dont l'ampleur n'est pas mesurable; même s'il est légitime de penser qu'elle est faible.

La méthode des quota

Apanage des instituts de sondages, cette méthode vise à pallier l'absence de base de sondage et à se débarrasser de la lourdeur des enquêtes par échantillon aléatoire ainsi que du problème des non-répondants.

Sur ce dernier point, il faut cependant noter que la difficulté n'est que masquée puisqu'un éventuel non-répondant est immédiatement remplacé par un autre sujet de mêmes quota; on retrouve donc l'hypothèse d'identité des répondants et des non-répondants.

On sait que l'inconvénient de cette méthode — majeur pour les calculs statistiques ultérieurs — est l'impossibilité de calculer la précision des estimations obtenues [8].

Celle-ci a été évaluée par Moser et Stuart [14] en répétant plusieurs fois la même enquête : les auteurs la situent de 1 à 3 fois moins bonne que celle des sondages aléatoires. Dans la même étude, Moser et Stuart montrent que la méthode des quota donne des estimations proches de celles des enquêtes de type aléatoires.

Ces auteurs insistent, à juste titre, sur le fait que ces résultats ne peuvent en aucun cas donner de fondements théoriques à la méthode des quota et ne doivent pas être généralisés sans précaution. Il n'empêche que les statisticiens ne devraient pas balayer trop facilement de la main cette méthode dont les vertus de rapidité et d'économie pourraient être exploitées dans certains domaines : étude de la morbidité, enquêtes préliminaires, ...

Échantillonnage validé par les résultats obtenus

Les deux méthodes qui viennent d'être examinées ont en commun le fait que la représentativité de l'échantillon est « acquise » par le procédé même d'échantillonnage. Une fois l'échantillon

constitué, il n'est pas besoin, en principe, de le contrôler. Ces méthodes n'incluent pas de moyens de vérifier les hypothèses nécessaires (validité des listes, pertinence des quota ([1]), identité répondants-non-répondants), et donc de déceler un biais éventuel.

Nous allons maintenant envisager une autre méthode, déplaçant ces difficultés. Nous nous appuierons pour cela sur l'exemple d'une enquête menée au niveau européen dont un des objectifs était d'estimer la plombémie moyenne des habitants de plusieurs grandes villes [1].

Supposons que l'on cherche à estimer la moyenne d'une variable X sur une population et que, pour diverses raisons, il est jugé impossible de constituer un échantillon aléatoire.

> *Dans l'enquête plombémie, les chercheurs français, contrairement à leurs collègues anglais et allemands, ont jugé non éthique d'aller frapper à la porte des gens tirés au sort pour leur demander un prélèvement sanguin. Ils attendaient de plus d'une telle procédure un fort taux de refus.*

Le principe est alors le suivant :

On cherche à définir plusieurs sous-populations particulières accessibles à l'échantillonnage aléatoire et supposées fournir des résultats sans biais. Si les estimations obtenues ne diffèrent effectivement pas, on conclut, conformément à l'hypothèse initiale, d'une part que chacune est non biaisée, d'autre part qu'on peut regrouper les échantillons pour avoir une estimation unique et plus précise.

Pour ce qui est des analyses statistiques ultérieures, tout se passe, concernant la variable Y étudiée, comme si l'échantillon était obtenu par tirage au sort.

Précisons comment choisir les sous-populations et ce qu'on entend par « les estimations ne diffèrent pas ».

Les sous-populations doivent être choisies de manière à asseoir solidement l'hypothèse d'absence de biais. Cela nécessite deux conditions :

- L'absence de biais doit être probable et étayée sur chacun des échantillons (éventuellement après prise en compte d'une covariable).
- Un biais éventuel doit être différent d'une sous-population à l'autre. De sorte que des estimations identiques le soient plus probablement par absence de biais que par existence d'un même biais partout.

> *Les échantillons des villes françaises ont ainsi été constitués à partir de trois sous-populations subissant de toute façon un prélèvement sanguin : les femmes enceintes, les donneurs de sang, et les consultants d'un centre de santé.*
>
> *Ces groupes diffèrent par l'âge et le sex-ratio, variables liées à la plombémie, dont il faudra tenir compte. Il se peut de plus que les consultants en centre de santé soient, en moyenne, en moins bonne santé que les sujets des deux autres groupes. Mais, en dehors de cela, on ne voit pas de raisons précises pour que l'un d'eux ait une plombémie moyenne différente de celle de la population générale.* Surtout, *on ne voit aucun argument pouvant expliquer un même biais si les moyennes observées ne sont pas différentes.*

Pour vérifier que les estimations calculées sur les divers échantillons ne diffèrent pas, le premier moyen qui vient à l'esprit est un test de comparaison des moyennes :

- s'il est « non significatif », cela peut vouloir dire deux choses : ou bien il n'y a pas de biais, ou

1. Bien que les variables habituellement utilisées dans la méthode des quota — âge, sexe, C.S.P. — soient des cofacteurs d'une importance assez universelle.

bien il est le même pour tous les échantillons ([1]). Or ceux-ci ont été choisis pour rendre ce dernier cas très improbable;

• si le test est significatif, il se peut que les hypothèses faites soient fausses. Mais il se peut aussi que la prise en compte de certaines covariables (âge, sexe, ...) fasse disparaître la signification. Il est enfin possible que, moyennant d'autres hypothèses ou des exigences de précision moins strictes, comme on le verra plus bas, on puisse continuer les calculs.

Les résultats de l'enquête plombémie sont les suivants :

Plombémie chez les hommes (en µg/dl)

Centre de prélèvement			Classe d'âge 16-30	31-45	46-60	61-75	Degré de signification
PARIS	CTS	m	18,69	24,27	23,36		p = 0,56
		s^2	99,16	62,35	63,94		
		n	16	15	14	0	
	CS	m	20,64	21,80	25,33	28,00	
		s^2	27,66	30,89	87,00	48,67	
		n	11	15	9	4	
MARSEILLE	CTS	m	15,07	14,89	14,71		p = 0,022
		s^2	40,90	30,68	43,13		
		n	14	26	29	0	
	CS	m		18,92	16,57		
		s^2		147,75	55,72		
		n	0	80	22	0	

Plombémie chez les femmes (en µg/dl)

Centre de prélèvement			Classe d'âge 16-30	31-45	46-60	61-75	Degré de signification
PARIS	CTS	m	13,60	14,86	17,46		p = 0,29
		s^2	20,36	52,75	28,87		
		n	20	14	11	0	
	MAT	m	14,04	13,42			
		s^2	14,50	17,72			
		n	27	12	0	0	
	CS	m	14,40	16,50	19,08	21,00	
		s^2	17,16	28,67	49,74	0	
		n	10	16	13	1	
MARSEILLE	CTS	m	8,72	9,65	13,84		p = 0,67
		s^2	7,08	17,90	31,47		
		n	17	21	10	0	
	CS	m		10,54	11,00		
		s^2		18,99	8,82		
		n	0	57	14	0	
TOULOUSE	MAT	m	7,51	9,17			p = 0,24
		s^2	16,35	53,9			
		n	63	20	0	0	
	CS	m	10,30	8,17	10,81	9,08	
		s^2	14,67	19,86	16,56	11,26	
		n	11	13	24	7	

• Dans chaque case, sont indiqués de haut en bas : la moyenne de la plombémie, sa variance, le nombre de sujets.
• Le degré de signification est celui de la comparaison par analyse de la covariance entre les deux ou trois types d'enquête.
CTS : Centre de Transfusion Sanguine;
MAT : Maternité;
CS : Centre de Santé.

1. Il se peut également que, par manque de puissance, on ne mette pas en évidence une différence qui existe vraiment.

En dehors des hommes de Marseille, les différences entre type d'enquête sont non significatives.
Cette « exception » ne doit pas conduire à rejeter l'hypothèse d'absence de biais. En effet :
— à Marseille, la différence n'est significative que chez les hommes;
— les autres degrés de signification sont très loin de la limite 5 %. Ce qui tend à écarter un éventuel manque de puissance;
— l'ensemble des résultats de l'enquête (non reproduits ici) ne montre pas de tendance systématique à ce que la plombémie corrigée par l'âge soit plus faible en CTS qu'en CS. Les « hommes de Marseille » apparaissent donc plus comme une exception que comme le signe d'un phénomène général.

Même s'il est le plus couramment utilisé, le test précédent n'est pas toujours le mieux adapté. En effet, d'une part on cherche à montrer l'égalité entre les moyennes et non pas l'existence d'une différence, d'autre part la stricte absence de biais peut ne pas être nécessaire ou être une condition trop forte. On utilisera dans ces cas un test d'équivalence ou on testera si la différence entre les moyennes est inférieure à un seuil fixé [17].

La mise en œuvre d'enquêtes dans ces conditions n'exige pas les hypothèses fortes énoncées plus haut. On peut se contenter d'hypothèses plus souples d'échantillonnage « presque non biaisé ». L'analyse, elle, se trouve sensiblement limitée. On ne peut pas en effet regrouper sans précaution des échantillons dont on sait qu'ils diffèrent, même si c'est de peu. Pour cela, il faut faire l'hypothèse supplémentaire que le biais moyen d'un échantillon à l'autre est nul. Tout se conclut alors par une augmentation de la variance de l'estimateur.

Une telle hypothèse peut paraître rarement justifiable de manière convaincante. Il faut, cependant, souligner qu'elle est présente implicitement jusque dans les enquêtes menées sur des échantillons aléatoires. Toute enquête contient en effet des sources de biais inévitables; on suppose, sans toujours le dire, que ces biais sont faibles et en moyenne nuls.

On ne fait ici que déplacer les sources possibles de biais, de façon à pouvoir déceler leur existence, en choisissant des populations facilement accessibles (donc sans problèmes de non-réponse ni d'exhaustivité de liste).

POUR CONCLURE

Les examens successifs des problèmes d'échantillonnage liés aux tests et à l'estimation ont permis d'aborder les deux grands types d'« ennuis » propres aux enquêtes d'observation :

- ceux qu'on peut appeler « confondants ». Ils sont apparus à propos des tests et sont, d'une certaine manière, inhérents au problème étudié : le fait de constituer deux groupes à comparer peut entraîner des différences entre eux autres que le seul facteur étudié;
- ceux qui sont à l'origine de biais au niveau de l'échantillonnage. Ils ont été examinés dans la partie consacrée à l'estimation. Ils sont introduits par le chercheur lui-même au moment de la constitution de l'échantillon.

On a vu que le tirage au sort permet, en général, de résoudre ces problèmes sur le plan théorique (avec quelques réserves toutefois quand il s'agit de tests). Mais, en pratique, il ne permet pas toujours d'éviter certains biais propres aux situations d'observation. Il peut être alors préférable d'utiliser d'autres méthodes d'échantillonnage, choisies en fonction des objectifs que l'on s'est fixé et des caractéristiques de la population étudiée.

Ce sont les contraintes propres à l'enquête qui doivent guider le choix d'une méthode d'échantillonnage. L'assise théorique et le caractère universel du tirage au sort constituent à la fois sa puissance et ses limites. Puissance que l'on retrouve par sa présence (même partielle) jusque dans les méthodes alternatives présentées ici. Limites car sa généralité même l'empêche parfois d'être l'outil le mieux adapté et est à l'origine de biais évitables par d'autres méthodes, plus particulières.

Bibliographie

[1] Awad L., Huel G., Lazar P., Boudène C. — Facteurs de variation interindividuelle de la plombémie - *Rev. Épidém. et Santé Publ. 29,* 113-124, 1981.

[2] Billewicz W.Z. — The efficiency of matched samples : an empirical investigation - Biometrics 21, 623-644, 1965.

[3] Bouyer J. — Contribution à l'étude des procédures d'échantillonnage dans les enquêtes d'observation - Diplôme d'Études et de Recherches en Biologie Humaine 1982 - Université de Paris-Sud.

[4] Breslow N.E., Day N.E. — Statistical methods in cancer research. Volume 1 : the analysis of case-control studies. IARC Scient. Pub. n° 32, W. Davis Ed. Lyon 1980.

[5] Chapman D.W. — A survey of non response imputation procedures - Proceedings of the American Statistical Association, Social Statistics Section 245-251, 1976.

[6] Cochran W.G. — The planning of observational studies of human populations - J.R.S.S. A 128, 234-266, 1965.

[7] Cochran W.G. — The effectivness of adjustment by subclassification in removing bias in observational studies - Biometrics 24, 295-313, 1968.

[8] Deroo M., Dussaix A.M. — Pratique et analyse des enquêtes par sondage. P.U.F. 1980.

[9] Draper and Smith — Applied regression analysis - Wiley 1967.

[10] Kruskal W., Mosteller F. — Representative sampling. III : The current statistical litterature - *Int. Stat. Review 47 (3),* 245-265, 1979.

[11] Lellouch J., Lazar P. — Méthodes statistiques en expérimentation biologique - Flammarion 1974.

[12] Lemel Y. — Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondage (ex. : redressement de l'enquête santé) - *Annales de l'I.N.S.E.E.,* n° 22-23, 273-281, 1976.

[13] Mc Kinlay S.M. — The expected number of matches and its variance for matched pair design - Applied Statistics 23, 372-383, 1974.

[14] Moser C.A., Stuart A. — An experimental study of quota sampling - J.R.S.S. A 116, 349-405, 1953.

[15] Rubin D.B. — Matching to remove bias in observational studies - Biometrics 29, 159-183, 1973.

[16] Tuyns A.J., Jensen A.M., Pequigniot G. — Le choix difficile d'un bon groupe témoin dans une enquête rétrospective - *Rev. Épidém. et Santé Publ. 25,* 67-84, 1977.

[17] Westlake W.J. — Symmetrical confidence intervals for bioequivalence trials - Biometrics 32, 741-744, 1976.

Annexe

A — *Calcul du pourcentage de réduction du biais* θ *par analyse de la covariance*

On se place donc ici dans le cas où la régression de Y sur X est non linéaire et a pour équation : $u(x) = e^{ax}$.

On a donc $B_I = E_1(e^{aX}) - E_2(e^{aX})$ où E_i est l'espérance dans la population P_i.

La différence d entre les populations est estimée par : $\delta = \bar{y}_1 - b\bar{x}_1 - (\bar{y}_2 - b\bar{x}_2)$

Or $$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{1}{n}\sum_{j=1}^{n} e_{ij}^{ax} + \alpha_i + \bar{\varepsilon}_i$$

d'où $$\delta = \alpha_1 - \alpha_2 + \frac{1}{n}\sum_{j=1}^{n}(e_{1j}^{ax} - e_{2j}^{ax}) - \frac{b}{n}\sum_{j=1}^{n}(x_{1j} - x_{2j}) + \bar{\varepsilon}_1 - \bar{\varepsilon}_2$$

et donc $$B_F = E_1(e^{aX}) - E_2(e^{aX}) - b\,[E_1(X) - E_2(X)]$$

X suivant, dans la population P_i, une loi normale $N(m_i, \sigma_i^2)$, on obtient :

- $E_i(X) = m_i$
- $E_i(e^{aX}) = \exp\left(\frac{2am_i + a^2\sigma_i^2}{2}\right)$
- $E_i(XY) = (m_i + a\sigma_i^2)\exp\left(\frac{2am_i + a^2\sigma_i^2}{2}\right)$

Par suite, l'estimation commune de la pente $b = \frac{cov_1(X, Y) + cov_2(X, Y)}{var_1(X) + var_2(X)}$ a pour expression :

$$b = \frac{a\sigma_1^2 \exp(2am_1 - a^2\sigma_1^2) + a\sigma_2^2 \exp(2am_2 - a^2.\sigma_2^2)}{\sigma_1^2 + \sigma_2^2}$$

En posant, comme précédemment : $m_1 = -m_2$; $\sigma_1^2 + \sigma_2^2 = 2$; $B = 2m_1$, on trouve :

$$B_I = e^{\frac{aB + a^2\sigma_1^2}{2}} - e^{\frac{-aB + a^2\sigma_2^2}{2}} \text{ et } B_F = B_I - dB$$

d'où :

$$\theta = 100\left[1 - \frac{|e^v(1 - e^{a^2 - 2v}) - bB|}{|e^v(1 - e^{a^2 - 2v})|}\right]$$

$$v = \frac{1}{2}\left[aB + 2a^2\frac{\sigma_1^2/\sigma_2^2}{1 + \sigma_1^2/\sigma_2^2}\right]$$

où

$$b = \left[a\frac{\sigma_1^2}{\sigma_2^2}e^v + a\,e^{aB + v}\right]\frac{1}{1 + \sigma_1^2/\sigma_2^2}$$

B_1 — *Calcul du pourcentage de réduction du biais* θ *par ajustement*

L'expression de δ est ici : $$\delta = \frac{1}{\sum\omega_i}\sum_{i=1}^{c}\omega_i(\bar{y}_{2i} - \bar{y}_{1i})$$

où c est le nombre de classes de X constituées. Si x_i désignent les bornes de ces classes, on obtient, avec les notations précédentes :

$$B_I = \int u(x)\,[\Phi_2(x) - \Phi_1(x)]\,dx;\quad B_F = \frac{1}{\sum\omega_i}\sum_{i=1}^{c}\omega_i\left[\frac{\int_{x_i}^{x_{i'}} u(x)\,\Phi_2(x)\,dx}{\int_{i}^{x_{i'}}\Phi_2(x)\,dx} - \frac{\int_{x_i}^{x_{i'}} u(x)\,\Phi_1(x)\,dx}{\int_{i}^{x_{i'}}\Phi_1(x)\,dx}\right]$$

Pour une forme de régression et des densités de probabilités Φ_1 et Φ_2 données, la valeur de θ dépend du choix des x_i et des w_i.

Le plus souvent, on prend les x_i de sorte que les classes de X soient des intervalles de longueurs égales et on prend $\omega_i = \frac{p_{1i}\,p_{2i}}{p_{1i} + p_{2i}}$ où p_{ij} est le pourcentage de sujets dans la classe i du groupe j.

C'est ce qui sera fait dans le calcul de la variance de δ.

Pour le calcul de θ, Cochran [7] indique qu'on obtient des valeurs proches en prenant des poids w_i égaux et des classes de X de même probabilités attendues dans le groupe 1. C'est ainsi

qu'ont été construits les tableaux IV et V. Dans ce cas, l'expression analytique de θ est :

$$\theta = 100\left[1 - \frac{1}{B_i}\left|\frac{1}{c}\sum_{i=1}^{c}\frac{A_i}{B_i} - M\right|\right] \text{ où } M = m_i \text{ si } u(x) = kx,\ M = e^{\frac{2am_i + a^2\delta_i^2}{2}} \text{ si } u(x) = e^{ax}$$

$$A_i = \int_{x_{i-1}}^{x_i} u(x)\,\Phi_2(x)\,dx; \quad B_i = \int_{x_{i-1}}^{x_i} \Phi_2(x)\,dx$$

c = nombre de classes et x_i = bornes des classes

B_2 — *Calcul de la variance de* δ *par ajustement*

Comme indiqué plus haut, on prend : $\omega_i = \dfrac{p_{1i}\,p_{2i}}{p_{1i} + p_{2i}} = \dfrac{1}{n}\dfrac{n_{1i}\,n_{2i}}{n_{1i} + n_{2i}}$

Par suite : $$\text{var}\,\delta = \sigma_y^2(1-\rho^2)\frac{1}{(\Sigma\omega_i)^2}\sum\omega_i\left(\frac{1}{n_{1i}} + \frac{1}{n_{2i}}\right) = \frac{2}{n}\sigma_y^2(1-\rho^2)\ \frac{1}{2\sum\frac{p_{1i}\,p_{2i}}{p_{1i} + p_{2i}}}$$

c'est-à-dire que $$K = \frac{1}{2\sum\frac{p_{1i}\,p_{2i}}{p_{1i} + p_{2i}}}.$$