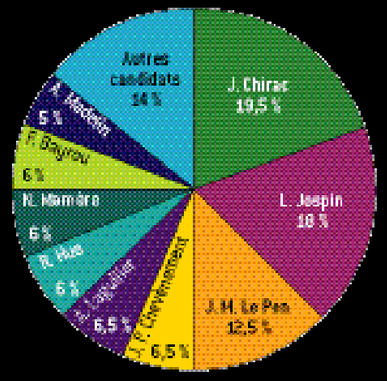
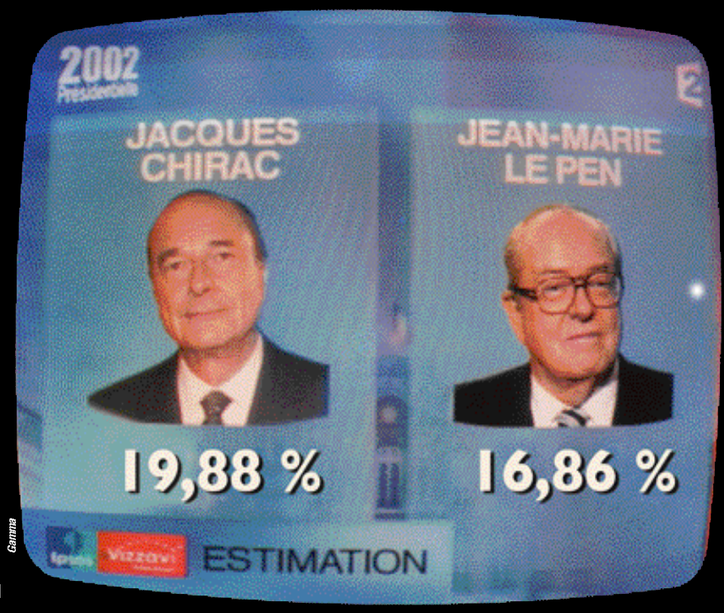
Peut-on croire aux sondages ?

Pour qu'une enquête statistique par sondage soit fiable, l'échantillon doit être choisi selon des règles probabilistes rigoureuses, et les biais doivent être maîtrisés. Les sondages commerciaux ou préélectoraux obéissent rarement à ces critères.

Jean-Claude Deville





Nous sommes quotidiennement abreuvés de statistiques et de résultats de sondages : indices boursiers, moral des ménages, évolution du chômage, popularité des hommes politiques, etc. Généralement, on n'explique pas au public comment ces chiffres sont construits ; tout au plus cite-t-on une source (l'institut Untel) ou précise-t-on qu'ils proviennent d'un sondage « auprès d'un échantillon représentatif ».

Dans quelle mesure ces évaluations statistiques sont-elles proches de la réalité ? On est par exemple souvent surpris que les résultats d'une élection politique diffèrent notablement de ce qu'indiquaient les sondages sur les intentions de vote, comme cela a été le cas avec les dernières élections législatives en Italie ou avec le premier tour des élections présidentielles françaises en 2002 *(voir la figure 2)*. En fait, la confiance dont sont dignes les sondages et autres enquêtes statistiques dépend beaucoup du sérieux et de la rigueur avec lesquels ils ont été élaborés. En quoi consiste un sondage ? À quelles conditions doit-il satisfaire pour que ses résultats soient pertinents ? Quels sont les biais dont il faut tenir compte ? Les sondages commerciaux sont-ils fiables ? C'est ce que nous examinerons successivement.

1. Les non-réponses aux enquêtes par sondage constituent un biais dont il faut savoir tenir compte.

Précisons d'abord la différence entre un sondage et un recensement. Le dénombrement de la population d'un pays et sa répartition par âge, par type d'habitat ou par activité économique se fondent sur des recensements périodiques. Ces gigantesques et coûteuses enquêtes mobilisent des dizaines de milliers d'enquêteurs qui dénombrent tous les logements et font remplir à leurs occupants un questionnaire nécessairement simple et court. Qui plus est, le dépouillement exige beaucoup de temps.

La technique du recensement a ainsi ses limites. D'où l'idée de recourir à des enquêtes portant sur une partie seulement de la population, idée qui remonte au moins à la fin du XIX^{e} siècle. En 1895, le Norvégien Anders Nicolai Kiaer présenta au congrès de l'Institut international de statistique les résultats obtenus par ce qu'il nommait la « méthode représentative ». Dans son enquête, il avait sélectionné un certain nombre de communes qu'il jugeait typiques, puis, à l'intérieur de chacune d'elles, des personnes choisies selon leur âge et l'initiale de leur nom. Or Kiaer montrait que les chiffres (des proportions) obtenus dans cette enquête portant sur 120 000 personnes étaient très similaires à ceux du dernier recensement. Encouragé par ce succès, il récidiva quatre ans plus tard avec une sélection de 10 000 personnes seulement, et il souligna l'intérêt de sa méthode pour obtenir des résultats qui seraient inaccessibles par un recensement, ainsi que l'économie associée en temps et en moyens.

La méthode représentative semblait efficace, mais on en ignorait les raisons précises. Aussi mit-elle du temps à convaincre les statisticiens, et c'est seulement en 1925 que le principe de l'utilisation des échantillons de population fut admis. Mais quelle était la bonne façon de procéder ? La réponse attendit encore quelques années. En 1928, Corado Gini et Luigi Galvani mettaient en évidence qu'un sondage portant sur un échantillon de 15 pour cent de la population italienne, tiré du recensement de 1921, aboutissait à des résultats douteux en dépit des précautions prises. Cela conduisit le statisticien d'origine polonaise Jerzy Neyman *(voir la figure 3)* à publier en 1934 un article fondamental, qui montrait que seul un échantillonnage aléatoire « contrôlé », en un sens que l'on précisera, offre une justification théorique rigoureuse aux enquêtes par sondage.

De l'échantillon à l'estimation

Neyman a montré que l'on pouvait considérer les résultats d'une enquête par sondage comme ceux d'une expérience aléatoire, à laquelle sont associées une certaine espérance mathématique et une certaine dispersion (ou écart-type). Supposons que l'on veuille mesurer, dans une population donnée, une grandeur T (la proportion de chômeurs, celle des individus aux yeux bleus, le salaire médian, la proportion d'entreprises ayant des accords de réduction de temps de travail, etc.). Pour un échantillon s donné, le sondage fournira la valeur estimée $T(s)$. L'espérance mathématique T^* de T est alors, par définition, une moyenne sur les résultats que l'on obtiendrait en sondant tous les échantillons possibles *(voir l'encadré ??)*. Neyman a démontré que, sous certaines conditions que nous allons décrire, l'espérance T^* est précisément égale au résultat réel T_R que fournirait un recensement. On peut aussi estimer, d'après la dispersion statistique au sein même de l'échantillon choisi, la dispersion des résultats $T(s)$, ce qui permet d'exprimer le résultat sous la forme d'un intervalle de confiance : c'est un intervalle qui contient la vraie valeur T_R avec une probabilité fixée, en général 95 pour cent *(voir la figure 3)*.

Neyman a également examiné la notion de stratification, c'est-à-dire la possibilité de partager la population en groupes homogènes, géographiques par exemple. Il a montré que l'échantillonnage optimal, c'est-à-dire le plus précis à coût fixé, n'est pas, comme on le croyait, de constituer un simple modèle réduit de la population, mais d'adapter la taille de l'échantillon de chaque strate à son homogénéité. Plus une strate est homogène relativement au caractère étudié, moins on a besoin d'y collecter d'informations, et Neyman a établi une formule qui quantifie cela *(voir l'encadré ??)*.

Le point de vue de Neyman fut rapidement adopté par les statisticiens chargés des enquêtes officielles dans tous les pays et organismes internationaux, et complété par

2. Le 21 avril 2002, les résultats du premier tour des élections présidentielles françaises ont créé la surprise. Ils n'étaient pas conformes aux résultats des sondages effectués quelques jours à peine avant le vote. Ainsi, selon le sondage LCI/SOFRES du 17-18 avril 2002, le candidat Lionel Jospin talonnait Jacques Chirac, assez loin devant Jean-Marie Le Pen. De tels sondages ont une faible fiabilité, notamment en raison de la petite taille de l'échantillon (autour de 1 000 personnes interrogées), de la façon dont il est choisi et du traitement des non-réponses.

de nouveaux outils apparus dans les années 1940 et 1950 et sur lesquels je reviendrai : l'échantillonnage à plusieurs degrés, l'échantillonnage à probabilités variables, l'usage d'informations auxiliaires pour l'estimation.

La stratégie d'un sondage comporte deux étapes. La première est l'échantillonnage aléatoire, le tirage au sort étant effectué selon un « plan de sondage » destiné à contrôler les propriétés des échantillons engendrés. La seconde est la construction des estimations sur l'échantillon et des extrapolations à la population entière. Elle consiste à définir, pour chaque échantillon possible s, une estimation $T(s)$ de la grandeur T à laquelle on s'intéresse (taux de chômage, dispersion des revenus, etc.). L'estimation est dite sans biais si son espérance mathématique T^* est égale à la valeur réelle T_R (la valeur pour l'ensemble de la population, celle que donnerait un recensement) de T. En plus de cette estimation dite ponctuelle, il s'agit d'obtenir une estimation par intervalle, c'est-à-dire la « fourchette » $[T^-(s), T^+(s)]$ qui recouvre la valeur T_R avec une probabilité fixée, généralement 95 pour cent. Plus cet intervalle est étroit, plus le sondage est précis. En première approximation, la largeur de l'intervalle de confiance est proportionnelle à l'inverse de la racine carrée de la taille de l'échantillon ; ainsi, pour diminuer de moitié la largeur de l'intervalle de confiance, il faudra multiplier cette taille par quatre.

Supposons que l'on veuille estimer le total d'une variable y (si y désigne le volume d'eau consommé mensuellement par un foyer, son total est la consommation mensuelle de l'ensemble des foyers) ou l'effectif d'un groupe (qui est le total d'une variable y valant 1 ou 0 selon que l'individu appartient ou non au groupe). En 1952, les Américains Daniel Horvitz et David Thompson ont introduit et étudié un estimateur naturel dépourvu de biais, c'est-à-dire dont la moyenne sur tous les échantillons possibles est égale à la valeur réelle T_R.

L'estimateur de Horvitz-Thompson est la somme, sur tous les membres de l'échantillon considéré, des quantités y_k/π_k où π_k est la probabilité qu'avait le membre k d'être sélectionné dans l'échantillon. Autrement dit, les valeurs de y sont pondérées par les inverses de ces quantités : si l'individu k avait une chance sur 1 000 d'être sélectionné, son poids sera de 1 000, et, en quelque sorte, il représentera 1 000 personnes de la population. L'estimateur de Horvitz-Thompson est en réalité peu utilisé, car on connaît des moyens simples et efficaces de l'améliorer. Mais il constitue le point de départ de tous les estimateurs sans biais, ou à biais dit négligeable, utilisés en pratique.

La constitution d'un échantillon ne va pas de soi. Pour tirer au sort un échantillon, il est nécessaire de disposer d'une liste complète d'unités « tirables ». Cette liste constitue la « base de sondage » et permet d'effectuer un tirage véritablement aléatoire (à l'aide d'un algorithme utilisant des nombres pseudo-aléatoires engendrés par un ordinateur).

Supposons que l'on veuille obtenir un échantillon de la liste où chaque unité sera sélectionnée avec, par exemple, la probabilité 0,01. La méthode la plus simple consiste à générer un nombre aléatoire r_k compris entre 0 et 1 pour chaque unité k de la population et à la sélectionner si r_k est inférieur à 0,01. Mais on produit ainsi un échantillon dont la taille est aléatoire, ce qui est gênant pour la planification du budget d'une enquête.

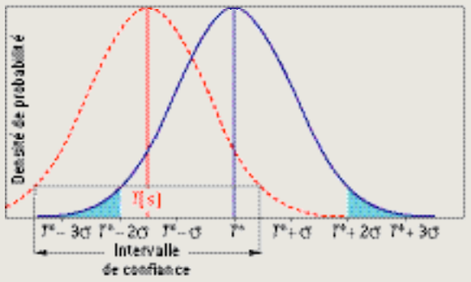
Des tirages au sort contrôlés

On préfère tirer des échantillons de taille n fixée et, si toutes les unités doivent être tirées avec la même probabilité, celle-ci vaut n/N où N est la taille de la population. On parle de sondage aléatoire simple si, de plus, tous les échantillons sont équiprobables. On connaît de nombreux algorithmes efficaces pour effectuer un tel tirage au sort de l'échantillon. On en connaît également dans le cas plus compliqué des probabilités variables, c'est-à-dire lorsque le plan de sondage n'affecte pas aux différents individus (ou unités) de la population les mêmes probabilités d'être sélectionnés dans l'échantillon.

Ces procédures d'échantillonnage ne sont cependant pas parfaites. Supposons que l'on applique l'estimateur de Horvitz-Thompson à des variables vérifiables, c'est-à-dire dont on connaît les valeurs pour chaque unité de la population ; par exemple, dans le cas où le sondage porte sur l'ensemble des communes, on peut considérer le nombre d'habitants âgés de moins de 20 ans : grâce aux enregistre-

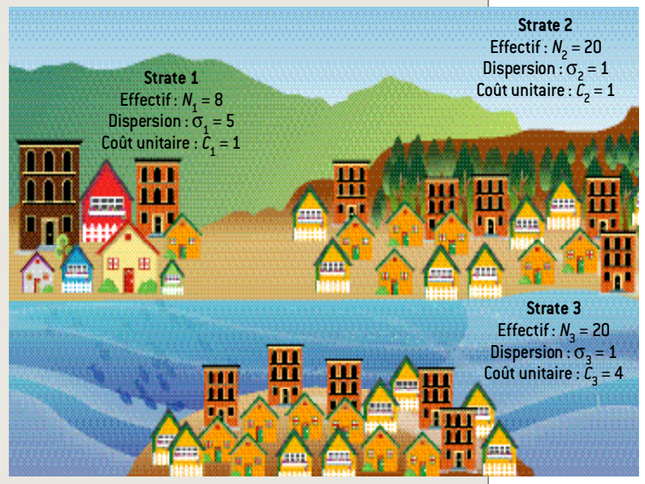
Espérance mathématique et intervalle de confiance

Lorsqu'on mesure par sondage sur une population donnée une grandeur T (la proportion de chômeurs par exemple), on sélectionne un échantillon s et on détermine une valeur $T(s)$. Comme il y a de nombreux échantillons possibles, le statisticien attribue à chaque échantillon s une certaine probabilité $p(s)$ de le sélectionner, ces probabilités dépendant des critères adoptés pour sélectionner les échantillons. L'espérance mathématique T^* est la moyenne, sur tous les échantillons possibles, de $T(s)$ pondéré par la probabilité $p(s)$ de sélectionner l'échantillon s (l'espérance T^* est donc la somme, sur tous les échantillons possibles, des produits $p(s)T(s)$). Si l'estimation est sans biais, ce que l'on suppose, T^* est égal à la valeur réelle T_R de T pour l'ensemble de la population. Les fluctuations de l'écart entre $T(s)$ et T^* sont caractérisées par l'écart-type noté σ, quantité que l'on peut estimer à partir de l'échantillon. L'estimation $T(s)$ a alors 95 pour cent de chances d'être comprise entre $T^* - 2\sigma$ et $T^* + 2\sigma$. De même, la fourchette ou intervalle de confiance $[T(s) - 2\sigma, T(s) + 2\sigma]$ a 95 pour cent de chances de contenir la bonne valeur T^*.



L'échantillonnage par strates

Comment échantillonner dans le cas où la population étudiée se répartit en strates plus ou moins homogènes ? Neyman a établi que pour un échantillonnage optimal, la taille n_h de l'échantillon dans une strate donnée h doit être proportionnelle à l'effectif N_h de la strate et à la dispersion (ou l'écart-type) dans cette strate du caractère étudié, mais inversement proportionnelle à la racine carrée du coût de la collecte de chaque donnée dans la strate. L'exemple illustré ici porte sur une population répartie en trois strates géographiques, qui comptent respectivement 8, 20 et 20 unités (des maisons individuelles). Chaque strate est caractérisée par son effectif, la dispersion ou écart-type du caractère considéré (ici la taille du logement) et le coût du recueil d'information par unité. Dans notre exemple, la strate 1 contient des logements de tailles très variables, et sa dispersion σ_1 est cinq fois plus élevée que celle des deux autres strates. La strate 3, d'accès difficile puisqu'elle est située sur une île, a un coût unitaire C_3 quatre fois plus élevé que dans les strates 1 et 2. L'optimum de Neyman consiste à répartir l'échantillon entre les strates proportionnellement aux quantités $N_h\sigma_h/\sqrt{C_h}$, qui valent ici 40, 20 et 10 respectivement. Par conséquent, si n est la taille totale de l'échantillon, les quatre septièmes seront alloués à la strate 1, deux septièmes à la strate 2 et un septième à la strate 3. Le coût total du sondage sera égal à $1 \times 4n/7 + 1 \times 2n/7 + 4 \times n/7 = 10n/7$. Si le budget de l'enquête ne doit pas dépasser dix unités de coût, on trouve que $n = 7$ et que l'échantillon comportera quatre unités de la strate 1, deux de la strate 2 et une seulement de la strate 3.



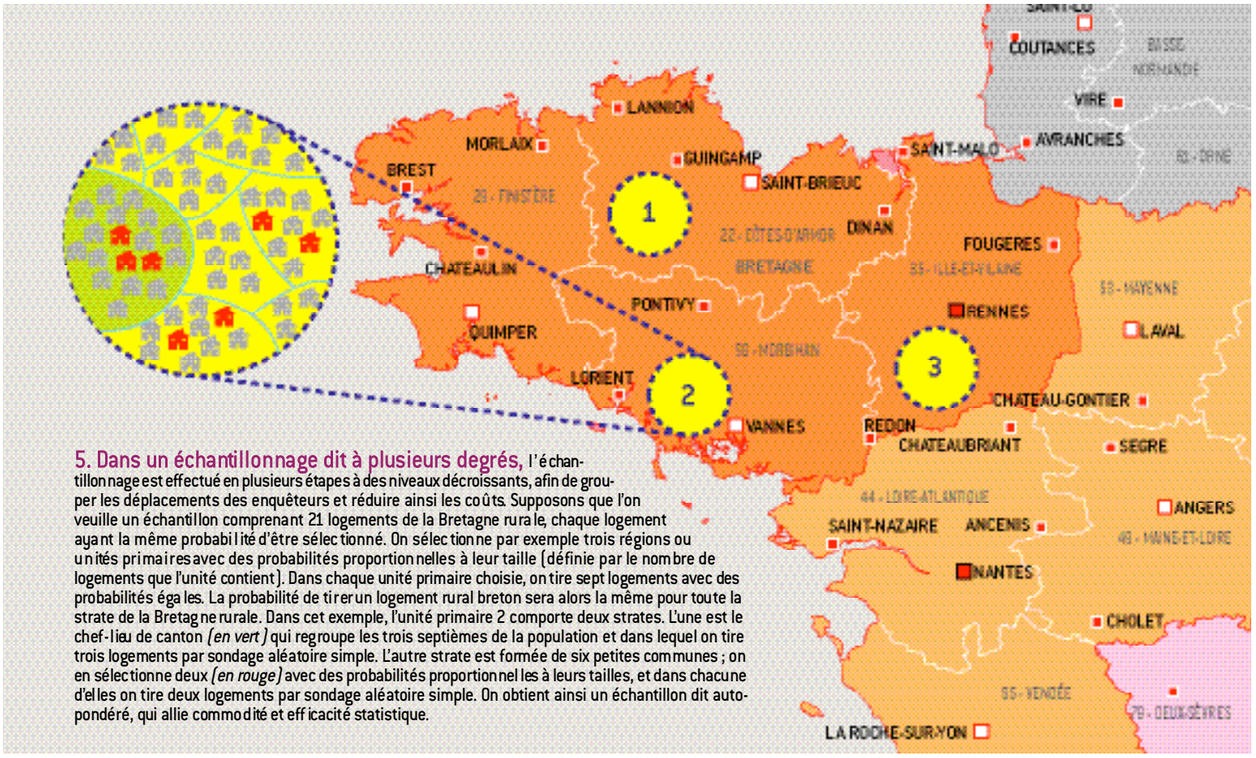
ments administratifs, il est connu pour chaque commune, et le total pour l'ensemble des communes sélectionnées dans l'échantillon est donc également connu. On s'aperçoit alors que les valeurs obtenues en appliquant l'estimateur de Horvitz-Thompson sur l'échantillon ne correspondent pas avec précision aux vraies valeurs, même si celles-ci sont contenues dans les intervalles de confiance.

Comment y remédier ? Avec Yves Tillé, de l'Université de Neuchâtel, nous avons découvert vers 2000 un procédé d'échantillonnage aléatoire dit équilibré qui permet de respecter non seulement une taille d'échantillon fixée, mais aussi la valeur du total de n'importe quel ensemble de variables vérifiables *(voir l'encadré ci-dessus)*. Il a l'intérêt d'augmenter considérablement la qualité des échantillonnages : il ramène à zéro (aux arrondis près) la largeur de l'intervalle de confiance des variables vérifiables et réduit, parfois radicalement, celle de toutes les variables auxquelles on peut s'intéresser. Cette méthode a été utilisée en France par l'INSEE, l'Institut national de la statistique et des études économiques, pour construire les échantillons du nouveau « recensement » annuel par sondage démarré en 2004, ainsi que l'échantillon de communes où se déroulent ses enquêtes depuis 2001.

Cette panoplie d'algorithmes – sondage aléatoire simple, probabilités variables, échantillonnage équilibré – n'épuise pas toutes les subtilités de l'échantillonnage. Les bases de sondage contiennent aussi des informations qui stratifient la population, c'est-à-dire qui la partagent en sous-populations relativement homogènes, les strates. Dans les enquêtes auprès des personnes, la stratification est surtout fondée sur des critères géographiques et d'habitat. Pour les enquêtes auprès d'entreprises, on utilise souvent le type d'activité économique et l'effectif salarié. Dans chacune des strates (il peut y en avoir une centaine ou plus), on tire un échantillon dont la taille est déterminée par le plan de sondage de façon à optimiser la précision des résultats.

Pour les enquêtes impliquant plusieurs petites villes et la campagne, on recourt à l'échantillonnage à plusieurs degrés pour décomposer le tirage *(voir la figure 6)*. Qu'est-ce que cela signifie ? Supposons qu'une strate soit constituée par les 1 005 communes rurales de Bretagne, où il faudra collecter, mettons, 160 questionnaires. Si l'on sélectionnait les 160 logements correspondants par sondage aléatoire simple, on obligerait les enquêteurs à suivre un coûteux itinéraire dans la campagne, d'autant plus que la réalisation à domicile d'un questionnaire nécessite en général plusieurs déplacements. On cherche donc à regrouper géographiquement les observations dans des zones relativement petites attribuées chacune à un enquêteur. La méthode consistera à tirer au hasard huit des 162 cantons ruraux de la région, par échantillonnage équilibré utilisant des probabilités proportionnelles aux nombres de logements de chaque canton. Chacun d'eux, quelle que soit sa taille, fournira 20 logements à l'échantillon. Cette répartition simple a une efficacité presque magique, car tous les logements de la strate ont la même probabilité de figurer dans l'échantillon. Elle assure la même charge de travail à chaque enquêteur, et elle optimise la précision des résultats à coût donné en supposant celui-ci composé d'un coût fixe par enquêteur augmenté d'un coût fixe par questionnaire.

Ce n'est pas fini : dans certains cantons, le chef-lieu peut regrouper 40 pour cent de la population, abrite la coopérative agricole, la banque, le collège et presque tous les commerces. Le canton est donc lui-même stratifié, et le chef-lieu se verra attribuer 8 questionnaires (40 pour cent des 20), les 12 restants devant être répartis entre les autres (petites)



5. Dans un échantillonnage dit à plusieurs degrés, l'échantillonnage est effectué en plusieurs étapes à des niveaux décroissants, afin de grouper les déplacements des enquêteurs et réduire ainsi les coûts. Supposons que l'on veuille un échantillon comprenant 21 logements de la Bretagne rurale, chaque logement ayant la même probabilité d'être sélectionné. On sélectionne par exemple trois régions ou unités primaires avec des probabilités proportionnelles à leur taille (définie par le nombre de logements que l'unité contient). Dans chaque unité primaire choisie, on tire sept logements avec des probabilités égales. La probabilité de tirer un logement rural breton sera alors la même pour toute la strate de la Bretagne rurale. Dans cet exemple, l'unité primaire 2 comporte deux strates. L'une est le chef-lieu de canton *(en vert)* qui regroupe les trois septièmes de la population et dans lequel on tire trois logements par sondage aléatoire simple. L'autre strate est formée de six petites communes ; on en sélectionne deux *(en rouge)* avec des probabilités proportionnelles à leurs tailles, et dans chacune d'elles on tire deux logements par sondage aléatoire simple. On obtient ainsi un échantillon dit autopondéré, qui allie commodité et efficacité statistique.

communes. Ici encore, on ne va pas disperser l'échantillon dans toutes ces communes, mais en tirer disons trois parmi la dizaine possible, en utilisant, de nouveau, un échantillonnage fondé sur des probabilités proportionnelles à la taille des communes. Enfin, dans chacun des trois villages, on tirera quatre logements par tirage aléatoire simple.

Ainsi, un échantillon comptant 15 000 ménages, taille fréquente pour les enquêtes officielles, est obtenu par de nombreux petits tirages au sort successifs, qui fournissent chacun une vingtaine d'adresses au plus. L'usage des ordinateurs a fortement allégé ces tâches et, surtout, autorisé l'emploi d'algorithmes efficaces.

Dans cet enchaînement de tirages aléatoires, chaque probabilité élémentaire est rigoureusement respectée : celle de tirer une unité primaire (un canton), puis une unité secondaire (un village dans un canton) et ainsi de suite jusqu'aux unités finales (ici les logements). La probabilité de tirer un logement est le produit de ces probabilités élémentaires. Elle est donc parfaitement connue et contrôlée, et son inverse donne le poids correspondant dans l'estimateur sans biais de Horvitz-Thompson. Cet ensemble de probabilités permet aussi de calculer l'intervalle de confiance. Tout semble parfait, mais deux éléments conduisent à nuancer la situation. Le premier est la disponibilité d'informations (par exemple la structure par âge de la population) non utilisées dans l'échantillonnage. Leur exploitation, faite après l'échantillonnage, permet d'améliorer sensiblement la précision d'un sondage *(voir la figure 7)*.

Dans les enquêtes sur l'emploi, par exemple, le nombre de chômeurs en France est estimé à plus ou moins 100 000 près si l'on utilise l'estimateur de Horvitz-Thompson. En modifiant les poids de cet estimateur par de simples règles de trois pour chaque tranche d'âge, de façon à en retrouver les effectifs exacts – cette technique est nommée poststratification –, la fourchette se réduit à plus ou moins 50 000. Cet intervalle de confiance correspond à celui d'un échantillon quatre fois plus grand auquel serait appliqué l'estimateur de Horvitz-Thompson.

Tenir compte des informations auxiliaires

Des travaux effectués dans les années 1990 avec Carl Särndal, de l'Université de Montréal, et Olivier Sautory, de l'INSEE, nous ont permis de comprendre toute l'importance de l'incorporation des informations auxiliaires dans les pondérations des estimateurs. Nous avons alors conçu des techniques généralisant la poststratification, aujourd'hui banalisées dans tous les instituts nationaux de statistique, qui améliorent l'estimateur de Horvitz-Thompson : on maintient presque parfaitement son caractère sans biais, et on diminue la largeur de l'intervalle de confiance en fonction de l'information fournie par des variables (sexe, âge, répartition géographique, catégorie sociale, niveau d'études, revenus, etc.) dont la répartition dans la population est connue.

Le second élément à prendre en compte est plus délicat : quelle que soit la qualité de l'échantillonnage et de la collecte des questionnaires, l'échantillon sera déformé par le fait que certaines unités sélectionnées ne répondront pas (unités injoignables, ou refus explicites de répondre, ou encore non-réponses dues à une incapacité physique, intellectuelle ou linguistique). S'y ajoutent des questionnaires inexploitables, incompréhensibles ou incohérents. La proportion de non-réponses varie selon la complexité de l'enquête : de 2 à 3 pour cent dans les (bons) recensements, 5 à 7 pour cent dans les enquêtes sur l'emploi, environ 15 pour cent dans des enquêtes relativement légères. Elle dépasse parfois les 40 pour cent pour des enquêtes complexes portant par exemple sur les déplacements ou sur la consommation des ménages, où l'on demande un relevé précis de toutes les dépenses pendant 14 jours consécutifs. C'est le maximum tolérable pour une enquête de la statistique officielle.

La correction pour ces non-réponses s'appuie sur ce que l'on sait des facteurs qui la font varier. Ainsi, pour les enquêtes auprès des personnes, le taux de réponse diminue avec la taille de l'agglomération et il est plus élevé dans les banlieues et les périphéries des villes qu'au centre. Il augmente avec la taille des ménages et il est meilleur dans les catégories populaires (ouvriers, agriculteurs) que parmi les cadres supérieurs et les entrepreneurs. Les probabilités individuelles de réponse dépendent donc de ces facteurs et chaque poids dans l'estimateur devra être corrigé en conséquence. Malheureusement, leurs valeurs exactes sont inconnues et varient d'une enquête à l'autre. On doit donc estimer ces probabilités à partir d'un modèle statistique. Cela présente d'importantes difficultés lorsque les variables sondées sont une cause directe de non-réponse – cas des revenus par exemple.

Beaucoup de non-réponses sont dues au refus ou à l'impossibilité de répondre à certaines parties du questionnaire. Cette même raison peut aussi engendrer des réponses fausses, volontairement (malgré la promesse d'une exploitation anonyme des données) ou, très souvent, involontairement. Il arrive que le questionnaire soit imprécis, par défaut de conception ou par nécessité (ainsi, la détermination du revenu exige de consulter plusieurs pièces justificatives, de sorte qu'on se contente souvent d'une approximation). De nombreuses enquêtes font intervenir la mémoire des personnes sondées, qui est parfois sélective. Il arrive aussi que le concept que l'on croit mesurer rigoureusement n'ait pas de réalité objective (une opinion par exemple) ou recouvre des images différentes pour différentes parties de la population. Ainsi, la notion de résidence principale perd sa pertinence pour des groupes aussi différents que les sans-logis, les retraités multirésidents, les étudiants ou une partie du milieu artistique ou sportif.

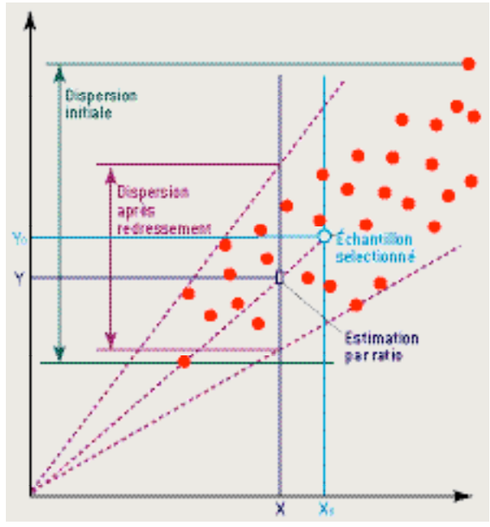
Dans les instituts nationaux de statistique, la réalisation d'une enquête par sondage s'étale sur plusieurs années. On met au point une problématique, un ensemble de buts à atteindre en termes de précision statistique et de budget. Si tout cela semble compatible, on choisit un mode de collecte de l'information (interviews face à face, interviews téléphoniques, questionnaires postés, etc.), puis on élabore un questionnaire, des variables statistiques et un système de codification. On recherche puis valide une base de sondage et un plan d'échantillonnage, des stratégies de collecte, de correction pour non-réponse et de mise au point de pondérations définitives. Puis viennent les phases d'exploitation, d'analyse des données et de calcul des précisions effectivement obtenues. Chaque phase est étudiée et testée avec soin, de façon à garantir la rigueur, l'objectivité et la transparence de la méthodologie.

Qu'en est-il avec les sondages commerciaux ? Ce type de sondages s'est développé aux États-Unis au milieu des années 1930. En France, on compte aujourd'hui plusieurs centaines de sociétés d'études qui réalisent des sondages pour répondre aux besoins d'information des entreprises, de la presse, des organismes politiques et même, parfois, des administrations. Leur chiffre d'affaires provient à plus de 90 pour cent des études de marchés. Moins rentables financièrement, les études politiques confèrent cependant de la notoriété.

La rentabilité, ennemie de la fiabilité

Les impératifs de rentabilité dans un milieu concurrentiel conduisent ces sociétés d'études à travailler avec des coûts faibles et des délais réduits. D'où l'abandon plus ou moins définitif de nombreux principes de la théorie. Les sondages qui utilisent une base de sondage et un échantillonnage aléatoire rigoureux, tels les sondages en « sortie des urnes », sont des exceptions. Il est vrai que les bases tirées des recensements ou des fichiers administratifs sont soumises à des règles de confidentialité et que leurs utilisations sont rigoureusement contrôlées, en France, par la Commission nationale informatique et liberté (CNIL).

Mais la vraie raison de cet abandon réside dans le coût de la collecte sur un échantillon aléatoire. Lorsque le hasard a désigné une adresse précise, on doit s'y tenir et, en cas



6. L'utilisation d'informations auxiliaires permet d'augmenter la précision des résultats d'un sondage. Il en est ainsi de la méthode dite d'estimation par ratio. Chaque point dans ce schéma théorique représente l'un des échantillons possibles, et l'on cherche à estimer le total d'une certaine variable y. L'échantillon sélectionné fournit la valeur Y_0 et, pour une autre variable x dont le total X est connu, la valeur totale X_0. L'estimation par ratio revient à corriger à l'aide d'une règle de trois : on remplace l'estimation initiale Y_0 par la valeur $Y = (X/X_0)Y_0$. On réduit ainsi la dispersion des estimations : quel que soit l'échantillon, la valeur estimée Y sera nécessairement comprise dans l'intervalle indiqué en violet, qui est plus petit que l'intervalle initial *(en vert)*. La méthode se généralise au cas de plusieurs informations auxiliaires. Notons qu'un échantillonnage équilibré aurait consisté à sélectionner un échantillon uniquement parmi ceux situés à proximité de la verticale d'abscisse X.

d'absence, ne pas frapper à la porte d'à côté pour éviter un déplacement supplémentaire ! L'expérience prouve que l'abandon du hasard contrôlé conduit toujours à des biais de sélection imprévisibles. Un ami, statisticien dans une compagnie d'assurance, avait demandé à sa secrétaire de « choisir au hasard » dans les archives quelques dossiers pour dégrossir un problème. Le nombre de sinistres qu'ils contenaient dépassait tout bon sens : en toute bonne foi, la secrétaire avait prélevé des dossiers au hasard, là où sa main se posait – c'est-à-dire, en général, sur les plus épais…

Dans les sondages commerciaux, l'échantillonnage n'est généralement pas réalisé selon une méthode probabiliste rigoureuse. Diverses pratiques ont cours. Dans la vieille méthode dite du choix raisonné, un « expert » supposé infaillible désigne les unités à interroger. Dans la méthode des quotas, très utilisée en France mais critiquée et peu utilisée dans le monde anglo-saxon, chaque enquêteur doit réaliser un certain nombre d'interviews vérifiant certaines contraintes, les quotas.

S'il doit faire 16 interviews, il devra par exemple partager son échantillon en 8 hommes et 8 femmes, en 4 personnes âgées de 15 à 29 ans, 3 de 30 à 44, 4 de 45 à 59, et 5 de plus de 60 ans, etc. Son choix est libre à l'intérieur de ces contraintes, dont l'addition donne les quotas globaux, proportionnels aux effectifs connus dans la population (le nombre d'hommes de l'échantillon global est proportionnel au nombre d'hommes dans la population, etc.). Chaque unité sondée est affectée du même poids et l'hypothèse d'extrapolation est qu'en contrôlant les effectifs des variables faisant l'objet de quotas, les répartitions de toutes les autres variables seront correctes. Ce n'est hélas vrai que sous des conditions très restrictives. Par exemple, il est beaucoup plus facile d'entrer en contact avec des personnes vivant en famille avec enfants que de tomber sur des célibataires ou des couples, ce qui faussera une enquête sur des pratiques culturelles ou sur la détention de contrats d'assurance-vie.

Des échantillons trop petits

Une autre caractéristique des enquêtes commerciales est la taille très réduite des échantillons, souvent moins de 1 000 individus. Il est ainsi étonnant de voir, en période électorale, une demi-douzaine d'instituts de sondage réaliser, chacun avec sa propre méthodologie, deux enquêtes par semaine sur 1 000 personnes – des enquêtes aux résultats peu précis alors qu'une seule enquête sur 12 000 personnes atteindrait une fiabilité raisonnable.

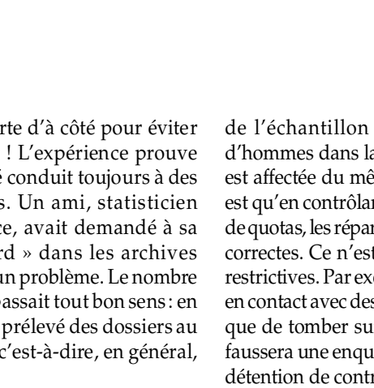
Pour un échantillon probabiliste de 900 personnes, l'intervalle de confiance est de plus ou moins trois pour cent pour un caractère présent dans environ 50 pour cent de la population. Autrement dit, si le sondage donne la valeur 50 pour cent, l'intervalle de confiance va de 47 à 53 pour cent. L'indication est certes intéressante, car elle fournit l'ordre de grandeur pour le caractère sondé, mais les analyses qu'on en fait généralement sont imprudentes.

Que dire des analyses des reports de voix dans les sondages préélectoraux ? Supposons que selon le sondage, le candidat X ait recueilli 11 pour cent des voix, soit 100 sondés, dont 40 pour cent iront au candidat A au second tour, chiffre en augmentation puisque le précédent sondage donnait 35 pour cent. Or l'intervalle de confiance associé aux 100 sondés, si on l'évalue comme si le sondage était rigoureusement probabiliste, va de 20 à 60 pour cent : la comparaison entre les 35 et les 40 pour cent n'a donc aucun sens.

Par ailleurs, les enquêtes empiriques ne tiennent pas compte de la non-réponse. Même quand elles sont réalisées par téléphone, on ne se préoccupe pas des appels infructueux ou des refus. Or dans les enquêtes politiques, en particulier, on peut penser que les personnes ayant des opinions extrêmes répondent peu aux enquêtes ou, ce qui est encore plus délicat, ne répondent pas toujours sincèrement. Les spécialistes des sondages politiques le savent, bien entendu, et utilisent des méthodes de redressement dont le détail est tenu secret, mais qui sont souvent davantage fondées sur l'intuition et la connaissance de la vie politique que sur des techniques statistiques rigoureuses, formalisables et reproductibles.

Alors, peut-on croire aux sondages ? La théorie mathématique des sondages offre un cadre rigoureux et solide ; ses résultats sont vérifiés de façon parfaite par des simulations numériques où l'on tire des milliers d'échantillons dans une population virtuelle constituée par un fichier informatique. Comme la mécanique, qui ne sait tenir compte que de façon approximative des phénomènes de frottement et de turbulence, la théorie des sondages ne traite les erreurs et les absences de réponse qu'approximativement. La fiabilité d'un sondage résulte essentiellement du respect de règles méthodologiques fondées sur une théorie mathématique éprouvée. Cet argument ne fait pas partie de la communication des sociétés d'études, et d'ailleurs le public comme les clients y sont assez indifférents. Est-ce à dire que les sondages commerciaux et politiques sont sans valeur ? Bien sûr que non. Le chansonnier Robert Rocca disait que, comme la minijupe, ils cachent l'essentiel mais donnent des idées. Et les clients semblent en être satisfaits puisqu'ils en redemandent. Cependant, si on appliquait un principe de précaution analogue à celui en vigueur dans les transports par exemple, on en publierait bien peu. Mais la statistique n'a jamais tué personne...

L'algorithme du cube



L'algorithme du cube désigne un procédé que nous avons récemment mis au point avec Yves Tillé pour effectuer un échantillonnage aléatoire dit équilibré. L'idée de départ est géométrique. Si la population étudiée est de taille N, on peut assimiler tout échantillon à un point ayant N coordonnées, dont la k-ième est égale à 1 si l'unité k de la population fait partie de l'échantillon et 0 sinon. Par conséquent, tout échantillon d'une population d'effectif N correspond à l'un des 2^N sommets du cube (ou plutôt hypercube) de côté un dans l'espace à N dimensions. Ainsi, à trois dimensions (population composée de $N = 3$ unités seulement, cas illustré dans le schéma), le sommet de coordonnées (1, 0, 1) représente l'échantillon constitué des unités numéros 1 et 3 de la population.

L'algorithme d'échantillonnage doit respecter les probabilités d'inclusion, c'est-à-dire les probabilités π_k ($k = 1, 2, \ldots, N$) que l'unité k soit incluse dans l'échantillon. Le point P de coordonnées $(\pi_1, \pi_2, \ldots, \pi_N)$ se trouve à l'intérieur du cube unité à N dimensions, puisque chaque π_k est un nombre compris entre 0 et 1.

On désire par ailleurs que pour une variable x (ou plusieurs) dont le total $X = x_1 + x_2 + \ldots + x_N$ sur la population est connu, l'estimateur de Horvitz-Thompson fournisse la valeur exacte X. Autrement dit, il faut que la somme des x_k/π_k, pour tous les membres k appartenant à l'échantillon, soit égale à X. On peut voir que les sommets s vérifiant cette égalité appartiennent à l'hyperplan L d'équation $(x_1/\pi_1)a_1 + (x_2/\pi_2)a_2 + \ldots + (x_N/\pi_N)a_N = X$, où les a_k sont les coordonnées d'un point arbitraire de l'espace à N dimensions.

Le point P appartient à cet hyperplan, et l'algorithme simule dans cet hyperplan une marche aléatoire *(trajectoire verte ou bleue dans le schéma)* partant de P et arrivant sur une des faces à $N - 1$ dimensions de l'hypercube. Cette face est caractérisée par une valeur fixée (0 ou 1) de l'une de ses coordonnées ; si cette coordonnée k_1 est égale à 0, on élimine l'unité k_1 de l'échantillon, si elle est égale à 1 on l'inclut. Ainsi, dans le schéma, la marche aléatoire verte aboutit à la face où la deuxième coordonnée est nulle : l'unité numéro 2 de la population est alors écartée.

La marche aléatoire continue ensuite en se restreignant à cette face, qui est un hypercube à $N - 1$ dimensions, et on répète le processus jusqu'à ce qu'on se retrouve en un sommet du polyèdre formé par l'intersection entre le cube initial et l'hyperplan L. S'il s'agit d'un sommet du cube, l'échantillon est déterminé et le processus est achevé ; sinon, parmi les sommets du cube les plus proches, on en choisit un au hasard avec des probabilités qui conservent le caractère sans biais de l'échantillonnage.

Auteur & Bibliographie

Jean-Claude DEVILLE dirige le Laboratoire de statistique d'enquête au Centre de recherche en économie et statistiques de l'ENSAI (École nationale de la statistique et de l'analyse de l'information), à Bruz, près de Rennes.

P. ARDILLY, *Techniques de sondage* (2e édition), *Technip*, 2006

A.-M. DUSSAIX et J.-M. GROSBRAS, *Les sondages : principes et méthodes*, P. U. F. (collection *Que sais-je ?*), 1993.

C.-E. SÄRNDAL *et al.*, *Model assisted survey sampling*, Springer-Verlag, 1992.

W. COCHRAN, *Sampling techniques* (3e édition), *Wiley*, 1977.