JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

PIERRE THIONET

Développements récents de la théorie des sondages

Journal de la société statistique de Paris, tome 100 (1959), p. 279-296
<http://www.numdam.org/item?id=JSFS_1959__100__279_0>

© Société de statistique de Paris, 1959, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

VI

VARIÉTÉS

Développements récents de la théorie des sondages

(Exposé fait au Séminaire de M. le Professeur Georges DARMOIS, à Paris les 8 et 22 janvier 1959)

Le titre de cet exposé est ambigu : La *technique* des Sondages s'est développée plus que la *théorie*. L'auteur passera en revue la littérature des deux ou trois dernières années (à l'exception de sa propre contribution). Il n'est pas d'ailleurs parvenu à en faire le tour complet (1) : la production des États-Unis, des Indes et de la Grande Bretagne n'est plus monopolistique, on doit compter de nos jours avec des pays comme le Japon, la Tchécoslovaquie, le Canada, la Suède, le Liban, la Suisse, l'Allemagne, le Brésil, l'Iran, etc...

I — LES DÉVELOPPEMENTS DE LA TECHNIQUE

A) *La technique des calculs d'erreur*

1° Les calculs d'erreur ont perdu, dans bien des pays, leur caractère de travail très compliqué, très coûteux et artisanal. On sait à présent les faire en grande série; et l'heure est venue d'y employer le calculateur électronique. Bien entendu ce sera onéreux; et l'exécution de ces calculs suppose que les gens qui contrôlent les crédits ne se désintéressent pas de ce genre de production. KEYFITZ [1] a expliqué comment il faisait déjà au Canada; il s'agit d'appliquer le théorème bien connu :

— si X et X′ sont 2 aléatoires indépendantes suivant la même loi de probabilité $(X-X')^2/2$ a pour espérance mathématique $\mathcal{V}X = \mathcal{V}X'$.

(1) Au Congrès de Washington (*Mars* 1957) de l'*Institute of Math. Statistics*, Carl KOSSACK (Purdue University) a fait une communication intitulée *A new approach to general purpose sampling* d'après *Annals of Math. Stat.*, June 1957. Nous n'avons pas eu communication de ce texte; et ce n'est certes pas la seule lacune dans nos informations.

Lorsqu'on disposera du matériel le plus récent, on considérera X^1 X^2 ... X^{10} et la moyenne des 45 carrés de la forme $(X^j - X^j)^2$ jouit de la même propriété; cet estimateur sera beaucoup plus précis que le précédent (il existe un papier de DESABIE sur la question Bull. d'Inf. de l'INSEE, 1959, nº 2, p. 65-74).

2º Jusqu'ici on admettait plus ou moins que la variance d'échantillonnage d'un estimateur X de μ était une fonction de μ, souvent $A\sqrt{\mu}$ ou $A\,\mu_0$; on estimait par V_i la variance des estimations X_i de quelques μ_i sans faire aucune hypothèse de structure; mais on portait X_i et V_i (ou log X_i, log V_i plutôt) sur un graphique et on ajustait une *courbe* (une droite) sur les quelques points connus. Un premier progrès avait consisté à remplacer la notion de courbe par celle de familles de courbes; car on s'était aperçu (dès qu'on avait des points un peu nombreux) que le nuage de points était médiocrement étiré.

3º Le progrès consiste donc dans deux faits : (*a*) on produit les variances en grande série; (*b*) on s'affranchit des hypothèses de structure.

Le progrès suppose cependant que les X_1 X_2... sont indépendants; avec des échantillons à plusieurs degrés il est douteux qu'on tire jamais 10 unités de sondage distinctes par strate; le calcul de variance est un sous-produit, il ne doit pas obliger à dépenser plus d'argent pour maintenir la qualité du produit principal, c'est-à-dire des estimations; et il est probable qu'on trichera avec l'indépendance quand on aura la machine électronique. Si X et X′ ne sont pas indépendants, on a :

$$\mathcal{E}\,(X - X')^2 = \mathcal{V}\,X + \mathcal{V}X' - 2\,\mathrm{Cov}\,XX' = 2\,\mathcal{V}X\,(1 - \rho)$$

$$\mathcal{V}\left(\frac{X + X'}{2}\right) = \frac{1}{4}\,\mathcal{V}\,X + \mathcal{V}\,X' + 2\,\mathrm{Cov}\,XX') = \frac{1}{2}\,\mathcal{V}\,X\,(1 + \rho)$$

$$\mathcal{V}\left(\frac{X + X'}{2}\right) = \frac{1}{4}\,\frac{1 + \rho}{1 - \rho}\,\mathcal{E}\,(X - X')^2$$

Si ρ *est négatif*, l'erreur commise sera *dans le bon sens* quand on admettra $\rho = 0$

B) *La technique des estimations régionales*

1º Le temps est passé où l'on ne savait faire que de mauvaises ventilations par région des résultats d'un sondage. Les Allemands [2] venus tard aux sondages ont apporté leurs problèmes (dont celui d'avoir des évaluations pour des *Länder* de 1,2 millions d'habitants) et aussi leurs méthodes. On y trouve beaucoup d'esprit pratique; on ne s'est pas trop pressé et on a dépensé beaucoup pour la préparation (*Aufbereitung*) d'un seul grand sondage périodique très stable (le *Mikrocensus*).

a) Une méthode bien connue (sondage équilibré) a été employée pour le tirage des unités primaires de sondage (en pratique cela exige des travaux préparatoires massifs, méthodiques et longs).

b) Une méthode assez empirique a été imaginée, celle du transfert de renseignements statistiques hors sondage qu'on incorpore aux évaluations pour remplacer les estimations par sondage les plus douteuses; ceci suppose qu'on prend son temps pour préparer les publications des résultats [2 *bis*].

c) On a essayé par ailleurs de tenir compte des non-réponses en remplaçant les estimations par leurs valeurs asymptotiques (en appliquant une méthode d'échantillon « censuré ou tronqué » très empirique) (Pr. KELLERER).

2º Il y a également lieu de citer le sondage agricole néo-zélandais qui cherche à atteindre

des résultats par COMTÉ. Les problèmes des sondages agricoles sont partout plus difficiles que ceux des sondages démographiques (il subsiste inévitablement des catégories de culture où d'élevage très spécialisées et localisées qui sont réfractaires au sondage). Le sondage néo-zélandais est destiné à rendre quinquennal un recensement qui était annuel; les données disponibles pour préparer le plan d'échantillonnage ne font donc pas défaut; on n'a pas hésité à dépenser beaucoup d'argent pour utiliser les données élémentaires, autrement dit les *cartes-détail* elles-mêmes [3].

C) *La technique des questionnaires et des enquêtes*

Le temps n'est plus où l'on préparait le questionnaire et le plan d'enquête définitif après une enquête pilote portant sur quelques dizaines de personnes. De plus en plus l'enquête-pilote devient un vrai sondage, sur échantillon valable; et il arrive qu'on s'arrête là. Le « questionnaire » cède la place à un jeu de questionnaires, distincts par l'ordre des questions ou par la façon de les poser. On ne tranche plus entre 2 méthodes possibles d'enquête, on les essaie parallèlement. C'est la technique des expérimentations (agricoles) qui a pénétré celle des sondages (*replicated sampling*). Des résultats pour l'ensemble des échantillons peuvent toujours être obtenus, mais en outre on a des idées sur l'influence qu'exerce l'instrument de mesure (enquêteur, questionnaire, etc...) en comparant les 2 moitiés (ou les 4 quarts) d'échantillon. MAHALANOBIS avait imaginé cette méthode comme moyen de contrôle des enquêteurs indiens (moyen assez médiocre, semble-t-il). Les exemples récents abondent où on l'emploie bien plus largement (par exemple [4]) : sondage de Cambridge; Enquêtes Budget de famille en Israël. Voir aussi [4 *bis*].

II — LES NOUVEAUX PROBLÈMES QUE LA THÉORIE VEUT TRAITER

A) *Il arrive que l'on traite de nouveaux problèmes. En voici 2 exemples* : [5-6]

1° Une enquête sur l'anthracose a été organisée pour 10 ans au Royaume-Uni dans 25 centres miniers choisis suivant un plan factoriel à 16 cases (2^4). Dans chaque centre, un sondage permanent étalé sur 6 mois environ permet de recueillir, avec un appareil spécial, un échantillon représentatif de poussières identiques à celles respirées par les mineurs.

L'échantillonnage porte sur les postes de travail et les équipes. Les hommes eux-mêmes changent d'équipe et de poste de travail (grâce aux feuilles de paie on a des données précises sur la durée de leur présence dans tel ou tel poste de travail).

On a *adapté* la théorie des sondages à ce problème assez neuf : on a calculé la *répartition optimum* de l'effort de sondage (c'est-à-dire des appareils de mesure) entre les équipes les postes de travail (le calcul a montré qu'on n'avait pas intérêt à adopter une nomenclature trop détaillée des postes de travail).

En même temps, il est prévu qu'on retrouvera toutes les observations correspondantes lorsque l'un des 35 000 mineurs de l'échantillon (des 25 centres) se présentera à la visite médicale. Aussi pourra être fait le rapprochement entre la quantité de poussières respirée et les observations d'ordre médical (ASHFORD [5]).

2° Des mathématiciens ont étudié les méthodes par lesquelles on évalue les effectifs des populations animales, notamment de *poissons*. On retire de l'eau, disons 1 000 poissons, on les marque et on les remet à l'eau; plus tard on retire de l'eau, disons 10 000 poissons,

on trouve que 25 sont marqués, et on en déduit que la population avait un effectif de $10\,000 \times 1\,000/25 = 400\,000$ poissons.

Le milieu sondé est tout naturellement partagé en zones ou strates; et comme une partie des poissons ont changé de zone dans l'intervalle, on considère la matrice obtenue en croisant les zones des 2 époques (en général matrice carrée) et les données numériques ventilées par strate; on estime ainsi à la fois la population et les mouvements migratoires.

Mais suivant les hypothèses faites sur la nature plus ou moins aléatoire (*a*) de la pêche, (*b*) des déplacements de poissons et (*c*) l'influence que l'une a sur les autres, on trouve bien entendu qu'il faut employer des estimateurs différents.

Il existe des moyens statistiques de tester ces hypothèses (CHAPMAN et JUNGE [6]).

B) On n'a malheureusement pas rencontré grand'chose pour traiter le problème qui, professionnellement, nous préoccupe le plus : *Comment utiliser correctement un échantillon qui n'est pas du tout tiré au sort?* (1)

A propos du cancer du poumon, signalons une note intéressante de BROWNLEE sur l'effet des non-réponses [7], qui donne un modèle plus général que celui de BERKSON sur le même sujet. Les formules publiées sont valables quel que soit le sujet de l'enquête mais ne concernent que les réponses par *Oui* ou *Non;* et on ne peut en faire grand'chose. Il faudrait sans doute que, comme pour les poissons du N° 2 ci-dessus, nous arrivions à tester quelque schéma de tirage d'échantillon. Pour l'instant la principale ressource que nous offre la technique est l'estimation par une formule de régression pondérée dont la mise en œuvre est très lourde. Les méthodes actuelles d'échantillons « censurés ou tronqués » sont insuffisantes ; elles ne correspondent guère aux structures réelles des populations sondées.

Le papier de BLACKWELL et HODGES sur la *stratégie* à adopter pour éliminer les biais dans le *choix* de l'échantillon présente peu d'intérêt pratique pour nous, mais prouve qu'on s'occupe du problème [8].

III — LES PROGRÈS DE LA THÉORIE CLASSIQUE

A) *La stratégie des sondages*

Le choix du plan d'échantillonnage et de la formule d'estimation correspondante relève de la Recherche Opérationnelle [10]; et il y a une tradition (*Neyman* 1934) qui définit le sondage optimum comme un extrêmum lié :

Variance minimum à coût constant ou coût minimum à variance donnée.

HAJEK [11] emploie le mot « stratégie » de la théorie des jeux pour désigner l'optimum neymanien. KHADJENOURI [12] emprunte au traité de SUKHATME (1954) une autre notion d'optimum : l'optimum rendrait minimum le produit « Coût par Variance ». Dans le cas précis, il y a équivalence avec la notion classique.

HAJEK s'est aperçu que, plus généralement, on avait $(C-C_o)$ $(V-V_o)$ minimum si C et V étaient « *canoniques* » (à son sens) (et que c'était une simple conséquence de l'inégalité de SCHWARTZ). Soit y_1 y_2 y_3... les paramètres, on entend par là que

$$C - C_o = y_1 C_1 + y_2 C_2 + y_3 C_3 \ldots$$

$$V - V_o = \frac{V_1}{y_1} + \frac{V_2}{y_2} + \frac{V_3}{y_3} + \ldots$$

(1) Cette question est examinée avec plus de détails dans un autre article publié dans le présent journal (Comptabilité Sociale et Méthodologie Statistique).

On constate que l'optimum au sens de NEYMANN, qui correspond à

$$d\,V + \mu\, d\,C = 0$$

(μ étant le multiplicateur de LAGRANGE) s'écrit alors

$$d\,V + \frac{C}{V}\, d\,C = 0$$

ou

$$d\,(V \cdot C) = 0$$

ou

$$(V - V_o)\,(C - C_o) \text{ minimum}$$

La courbe C (V) des optima est une hyperbole équilatère, enveloppe des familles d'hyperboles équilatères obtenues en faisant varier soit y_1 soit y_2, soit y_3, etc...

Pour HAJEK on a donc : $d\,C + \frac{V}{C}\, d\,V = 0$.

Avec BILLETER (14) on écrit au contraire $dC + b\,dV = 0$ (en se référant à certaines idées de WALD) b étant une constante qui représente le coût d'une perte unitaire dV. Cette façon de faire paraît satisfaisante si l'on pense à V, perte d'information (1). Mais elle ne convient pas du tout dans le cas de formes canoniques (hyperboles équilatères au lieu de droites comme lieu des optima). Elle ne convient pas plus dans le problème des 2 strates qu'étudie BILLETER, où 1 strate est sondée et l'autre enquêtée à 100 %; on a alors

$$C = a\,n + a^*\,(N - N_1)$$
$$V = f\,(N_1)/n$$

les paramètres étant n et N_1 (mis pour y_1 et y_2). L'optimum Neymanien (2) conduit à

$$\mu = a\,f'^2/a^{*2}f = f/a\,n^2$$

Si l'on veut que μ soit constant, il vient (en intégrant)

$$f = k\,\int \text{ (avec } k = \mu\,a^{*2}/4\,a)$$

Avec un sondage bernoullien dans la strate sondée, on a

$$V = \mathcal{V}\,(N_1\,\overline{X}_1 + (N - N_1)\,\bar{x}_2) = N_1^2\,\frac{\sigma^2}{n}$$

Donc :

$$\sigma^2 = k$$

Mais il est impossible que σ^2 soit constant si la frontière entre les 2 strates est variable (le problème étant justement de trouver la valeur optimum de N_1).

Conclusion : l'article de BILLETER (avec la collaboration de Van IJZEREN) contiendrait une erreur de logique (le multiplicateur de Lagrange ne peut être constant).

Ajoutons qu'il l'eût évitée s'il avait connu notre article de 1955 sur la même question, qui complète les travaux de DALENIUS en traitant le cas où est V donné et C minimum, cas que DALENIUS avait omis de traiter [15].

---

(1) Voir Étude théorique n° 7 de l'INSEE, 1959.

(2) $d\,V + \mu\,d\,C = \left(-\frac{f}{n^2} + \mu\,a\right) d\,n + \left(\frac{f'}{n} - \mu\,a^*\right) d\,N_1 = 0$ quels que soient $d\,n$ et $d\,N_1$; d'où $\mu = \frac{f}{a\,n^2} = \frac{f}{a^*\,n} = \left(\frac{f'}{a^*\,n}\right)^2 \cdot \left(\frac{a\,n^2}{f}\right) = \frac{f'^2}{f}\,\frac{a}{a^{*2}}$

B) *Stratification optimum*

Pour l'ensemble de cette question (assez neuve), nous renvoyons à la thèse de DALENIUS [16], ainsi qu'au Chap. III d'une édition passée de notre Cours de sondage [17].

1° Limitons-nous aux articles récents. Reprenons nous-même le problème à côté duquel BILLETER est passé. Retenons de son papier l'idée de prendre un écart-type de la strate sondée proportionnel à une certaine puissance de la *taille* de la strate : $\sigma^2 = k^2 N_1^g$; c'est l'une de ces hypothèses de structure qu'on retrouve, depuis le début des sondages (MAHALANOBIS, FAIRFIELD SMITH) pourquoi pas ici?

Nous trouvons immédiatement le résultat très simple :

$$\frac{n}{N_1} = \frac{a^*}{a\,(g+2)} = \text{constante}$$

Ainsi, il y aurait une *fraction de sondage optimum* pour la strate sondée, que l'on suppose le coût donné ou au contraire la variance donnée.

Bien entendu nous avons fait des hypothèses approchées; si la frontière entre les 2 strates se déplace beaucoup, on n'a plus le droit de supposer constants les coûts unitaires a et a^*; et si les ressources financières étaient assez grandes, il serait incorrect de supposer n petit à côté de N_1 (c'est-à-dire de supposer le sondage bernoullien).

2° DALENIUS (avec HODGES) a publié un nouveau mémoire en 1957 [18]. Il revient sur son résultat de 1950, qui consiste en un système d'équations non résolu (dont la simplicité n'est qu'apparente). Lorsqu'on sonde une population n'ayant pas une loi de distribution théorique connue, ce résultat n'est guère utilisable; et c'est justement là tout le problème réel des sondages. DALENIUS avait conjecturé en 1950 que, dans le doute, on avait intérêt à choisir des strates de taille N_h inversement proportionnelle à l'écart-type σ_h (c'est souvent une hypothèse constructive que de supposer qu'on connaît les ordres de grandeur des σ_h).

MAHALANOBIS (1952), HANSEN, HURWITZ et MADOW (1953) se sont déclarés partisans à priori de strates d'égal volume (c'est-à-dire N_h inversement proportionnel à μ_h). Un autre auteur, AOYAMA (1954, *Annals of Math. Stat. p. 1*) propose des strates d'égale longueur ($x_h - x_{h-1}$ = constante). D'autres préconisent des strates *d'égale dépense.*

Précisons qu'il ne s'agit pas du problème des 2 strates. Il s'agit du partage en L strates, entre lesquelles l'échantillon est réparti suivant l'optimum neymanien. Et DALENIUS va même s'intéresser au cas où L est très grand.

A priori on aurait pensé que la « stratification optimum » n'était un problème intéressant que si L reste petit; car si le nombre de strates est grand il paraît intuitif que l'emplacement exact de leurs frontières importe très peu.

En réalité cette impression serait pleinement justifiée si l'on répartissait l'échantillon entre les strates « à la BOWLEY », avec d'égales fractions de sondage; elle est totalement erronée avec une répartition « à la NEYMAN ».

On fait d'abord des hypothèses simplificatrices un peu gênantes (1) : la variable étudiée (qui est confondue avec la variable stratifiante) aurait un intervalle fini (a, b) de variation; et sa distribution serait très régulière (dans un intervalle $x - \varepsilon/2$, $x + \varepsilon/2$, sa densité resterait compromise entre $f(x) - \eta$, $f(x) + \eta$, η très petit avec ε).

(1) C'est l'absence d'une *queue* de distribution qui est gênante en théorie sinon en pratique.

On introduit une variable auxiliaire

$$y = \int_a^x \sqrt{f(u)\,du} = G(u), \qquad \text{avec} \qquad G(a) = 0,\ G(b) = K;$$

on prouve que K est fini (par l'inégalité de SCHWARTZ).

On découpe alors (a,b) en L intervalles de frontières x_h^* correspondant à d'égales variations de G :

$$G(x_h^*) = h\ K/L$$

Alors si l'on considère les poids

$$w_h = F(x_h) - F(x_{h-1})$$
$$\text{avec } F = \int f(u)\,du$$

et l'écart-type de l'estimateur Neymanien $\sigma = \Sigma\, w_h\, \sigma_h$, en supposant L fixe très grand, on constate que σ est minimum lorsque les x_h coïncident avec les x_h^*.

Le calcul consiste (en gros) à remplacer $f(u)$ par une constante dans chaque petit intervalle de longueur ε_h; alors σ_h^2 est équivalent à $\varepsilon_h^2/12$; et σ est de la forme $\Sigma\, \xi_h^2$ avec $\Sigma\, \xi_h =$ constante; le minimum de σ correspond à $\xi_1 = \xi_2 = \xi_h\ (= h\ K/L)$.

On tire de ce calcul la conséquence annoncée :

Asymptotiquement $[w_h\, \sigma_h =$ constante$]$ entraîne : w_h (ou N_h) proportionnel à $1/\sigma_h$.

On n'en tire pas du tout une solution asymptotique de l'équation de DALENIUS (en x_h) qui est :

$$\frac{\sigma_h^2 + (\bar{x}_h - x_h)^2}{\sigma_h} = \frac{\sigma_{h+1}^2 + (\bar{x}_{h+1} - x_h)^2}{\sigma_{h+1}}$$

3° (1) Les mêmes auteurs ont publié (début 1959), avec une revue de l'ensemble des travaux antérieurs, des résultats qui complètent substantiellement ceux de 1957. Il s'agit : d'une part d'une méthode pour améliorer par approximations successives le découpage des x_h^* ci-dessus; d'autre part, pour 2, 3, 4, et 5 strates, des valeurs numériques des (x_h) et des variances pour divers types de loi de distribution $[f(x) = e^{-x}$, ou $x\, e^{-x}$, ou $2\,(1 - x)]$ dont l'intérêt pratique est certain et auxquels DALENIUS n'a cessé d'appliquer ses méthodes depuis le début; enfin d'une illustration du gain de variance réalisé en employant sa méthode au lieu de celle de MAHALANOBIS [18 *bis*].

4° Un travail un peu moins récent sur le même sujet est dû à KITAGAWA [19] (Le mémoire de KITAGAWA est énorme, date de 1953 et est réparti sur plusieurs fascicules du *Sankhyà* dont nous n'avons vu que le dernier; pour autant qu'on sache, la loi de GAUSS y joue un rôle excessif). KITAGAWA étudie le cas d'*un nombre fini ou infini de strates.* Il ne contredit pas MAHALANOBIS mais il semble souhaiter qu'on introduise des coûts différents dans chaque strate quand on parle de stratification optimum; et il a, à ce propos, une idée intéressante dont nous allons tirer (à sa place) les conséquences :

A coût égal $(C_1 = C_2 \ldots)$ l'optimum Neymanien consiste à minimiser $y = \Sigma\, N_h\, \sigma_h$, sachant que $\Sigma\, N_h$ et $\Sigma\, N_h\, \mu_h$ sont constants l'un de l'autre.

1° Si par exemple on suppose $\sigma_h^2 = k^2\, N_h^g$, on a $\Sigma\, N_h\, \sigma_h = k\, \Sigma\, N_h^{1+g/2}$ d'où : Optimum $= N_h$ *égaux entre eux.*

(1) Alinéa ajouté après le séminaire.

2° Si en revanche, on suppose $\sigma_h = k' \mu_h^{1+e} N_h^e$
c'est-à-dire $\gamma_h = k' (N_h \mu_h)^e$ on a $\Sigma N_h \sigma_h = k' \Sigma N_h \mu_h \gamma_h$
d'où : Optimum $= N_h \mu_h$ *égaux entre eux.*

On peut en inférer que MAHALANOBIS avait postulé pareille hypothèse de structure quand il a préconisé une répartition en strates d'égal volume.

Cette hypothèse de structure (coefficient de variation proportionnel à une certaine puissance du *volume* de la strate) est d'ailleurs fort plausible.

Remarque : Comment se concilient le découpage optimum de telle ou telle distribution particulière et le découpage préconisé par DALENIUS et HODGES, lorsque les strates sont très nombreuses? Ces auteurs précisent qu'il n'y a pas convergence des découpages les uns vers les autres, mais *équivalence des expressions* $\Sigma N_h \mu_h$.

Il ne faudrait pas croire que, lorsque L est grand, si $\sum_1^L y_i =$ Constante L. $\sum_1^L y_i^2$ tend vers une limite indépendante du découpage. C'est pourtant l'erreur de ceux qui croient que la stratification est toujours bonne si le nombre de strates est assez grand.

C) *Un type peu connu de Sondage avec probabilités égales et remise*

L'article de DES RAJ et KHAMIS [20] concerne ce que nous avons appelé ailleurs le sondage *bernoullien avec identification;* il suppose en effet que bien qu'on remette les boules tirées une à une, on soit capable de s'apercevoir qu'on a tiré 2 fois la même boule et non pas une autre boule portant la même marque x. Ceci revient à dire que les boules portent un numéro d'identification.

Les résultats des 2 professeurs de Beyrouth recouvrent en partie les nôtres. Soit une urne renfermant 10 boules, faisons y 3 tirages avec identification; 3 cas sont possibles :

1er *cas* (probabilité 72 %) : on tire 3 boules distinctes; l'estimateur est $(x_1 + x_2 + x_3)/3$;

2e *cas* (probabilité 27 %) : on tire 2 fois x_1 et une fois x_2; l'estimateur usuel est $(2x_1 + x_2)/3$; et on s'est aperçu que $(x_1 + x_2)/2$ était bien meilleur;

3e *cas* (probabilité 1 %) : on retire 3 fois x_1; et l'estimateur est x_1 de toute façon.

Bref, c'est seulement grâce au 2e cas qu'on peut réduire la variance; avec l'estimateur usuel, la variance est $\sigma^2/3$. Si l'on sait identifier les boules, la variance tombe à :

$$0{,}72 \cdot \frac{\sigma^2}{3} \frac{10-3}{10-1} + 0{,}27 \cdot \frac{\sigma^2}{2} \frac{10-2}{10-1} + 0{,}01 \cdot \sigma^2$$

On peut calculer que, dans le 2e cas, l'emploi de l'estimateur défectueux élève la variance de 1 à $1 + 5/36$.

Plus généralement si l'urne a N boules et qu'on y fait m tirages donnant u boules distinctes, D. R. et K. ont comparé les variances σ^2/m et V (avec identification) et trouvé :

$$\frac{\sigma^2}{m} - V = \frac{N\sigma^2}{N-1}\left[\frac{1}{m} + \frac{1}{N} - \frac{1}{mN} - \xi\left(\frac{1}{u}\right)\right]$$

Alors avec l'aide de la loi de u qu'on trouve dans le traité de probabilités de FELLER, ils établissent que cette expression est positive (nulle si $m = 1$). Ainsi *en moyenne* l'estimateur « avec identification » est meilleur que l'estimateur bernoullien.

Ce beau calcul nous renseigne peu sur ce qui se passe réellement; c'est-à-dire que nous aurions aimé savoir s'il y a des valeurs de u pour lesquelles on n'aurait pas intérêt à employer l'estimateur avec identification.

Une erreur à ne pas commettre est de comparer σ^2/m et σ^2 $(N - u)/(N - 1)$, c'est-à-dire $1/u$ et $Q = 1/m + 1/N - 1/mN$; par exemple pour $N = 100$, $m = 25$, $(Q = 1/20.2)$ il serait erronné de croire qu'on n'a plus intérêt à employer l'estimateur « avec identification » lorsque u ne dépasse pas 20 (cas d'ailleurs exceptionnels). Nous disons que ce serait une erreur car, pour chaque valeur donnée de u, (σ^2/m) n'est pas la *variance liée par u*.

Nous sommes bien tenté en revanche de dire que la variance d'une expression comme

$$(x_1 + x_2 + x_3)/3$$

est *toujours* plus petite que celle (disons) de : $(2x_1 + x_2 + x_3)/4$

« *Toujours* » n'est d'ailleurs pas le mot qui convient : nous avons étudié les estimateurs pondérés et constaté que *certains* pouvaient présenter des variances plus faibles que l'estimateur non pondéré; mais les cas visés ne sont pas de même nature qu'ici. Ici c'est la valeur de u (et non pas les xi) qu'on prend en considération. Il existe un théorème très général de Basu (Sankhya, décembre 1952 page 46) d'après lequel le minimum de variance est atteint par les estimateurs *symétriques;* et nous croyons bien que ce théorème s'applique ici.

En résumé nous croyons que l'inégalité démontrée par DR et K n'est pas seulement vraie *en moyenne* mais vraie pour toute valeur de u, et que (si l'on sait identifier les boules tirées) l'estimateur avec identification est le meilleur *quel que soit u.*

D) *Sur l'Estimation par Ratio*

1° Hajek [11; 21] a eu la bonne idée de remplacer l'estimation par ratio par une estimation linéaire asymptotiquement équivalente dont la vraie variance égale la variance approchée de l'estimation par ratio (autrement dit sa perte d'information). En 1957 il s'agissait du ratio banal avec 1 seule variable auxiliaire — en 1958 d'une estimation beaucoup plus générale, qui comprend comme cas particulier l'estimation par *stratification a posteriori* de l'échantillon, autrement dit la « repondération » des données échantillons (procédé très courant).

La parenté entre ces procédés était déjà bien connue, mais la théorie faisait défaut. Lorsqu'on remplace $\frac{a+x}{b+y} - \frac{a}{b}$ par $\frac{a}{b}\left(\frac{x}{a} - \frac{y}{b}\right)$ par exemple, on fait finalement une *hypothèse de structure* sur la population de base (on admet que la droite de régression passe par l'origine — ou tout près de l'origine). Rien n'empêche de faire une hypothèse analogue avec plusieurs variables auxiliaires y z t (et un hyperplan de régression). Allant plus loin, on peut supposer gaussien l'écart (à cette droite ou cet hyperplan), ce qui permet de construire des intervalles de confiance et des khi-deux.

2° (1) On a parlé à Hajek à Bruxelles de l'article de Goodman et Hartley sur les estimateurs du type-ratio sans biais. Il ne le connaissait pas encore, nous non plus [23]; il fait suite à un article de Hartley et Ross dans la revue *Nature* (1954), à un rapport non publié de Mickey et à un article de Robson [22] sur les « *polykays* ».

En réalité il existe (2) deux estimateurs simples par ratio, biaisés l'un et l'autre; le plus courant est établi à l'aide du ratio $\bar{x}'/\bar{y}'$ ou Sx_i/Sy_i; mais on peut aussi bien former le ratio $r_i = x_i/y_i$ de l'unité échantillon (i) et prendre la moyenne des n ratios r_i. On voit immédiatement que ces deux estimateurs ont même variance (en sondage bernoullien) mais que le *biais* du second est fini, alors que le biais du premier est de l'ordre de $(1/n)$, de sorte

(1) Alinéa ajouté après le Séminaire.
(2) Nous l'avions signalé (1953) dans : *Théorie des Sondages* p. 222-223.

qu'on employait toujours jusqu'ici le premier estimateur. HARTLEY et ROSS se sont aperçus que le biais du second estimateur *s'éliminait* totalement par simple adjonction d'un terme correctif adéquat.

Si r' désigne la moyenne des n (r_i) échantillon, leur estimateur sans biais est (avec nos notations)

$$\bar{x}'' = r'\,\bar{y} + \frac{(N-1)\,n}{N\,(n-1)}\,(\bar{x}' - r'\,\bar{y}')$$

où $\bar{x}'/\bar{y}'$ serait justement le ratio banal (celui de HAJEK).

La variance de cet estimateur est d'obtention ardue, et ne demande rien moins que l'énorme machinerie des « *polykays* » (1); toutefois pour N infini, sa limite est relativement simple.

Cette variance est-elle plus ou moins grande que celle d'autres estimateurs? C'est là toute la question. Or il n'est pas apporté de résultat général vraiment clair.

Un exemple *numérique* montre $\bar{x}''$ meilleur que $\bar{y}\,\bar{x}'/\bar{y}'$ (ratio banal), mais moins bon que $\bar{x}'$ (estimateur direct); dans un autre, $\bar{x}''$ est le meilleur des trois : le tout est de choisir des cas où la régression de x en y est très différente d'une régression linéaire sans terme constant (pour être sûr que le ratio banal soit mauvais).

Ceci nous montre bien la *portée limitée* de la théorie de HAJEK.

3° *Un problème d'évaluation de variance*

DURBIN [24] dans sa communication au congrès de Stockholm démontre et étend considérablement une formule donnée par YATES (dans son manuel de sondages). Cette formule permet d'évaluer la variance d'une estimation obtenue sur un échantillon constitué par un segment de l'échantillon complet (autrement dit une estimation concernant un segment de la population sondée) lorsque ledit segment est découpé à travers les strates. Jusqu'ici, on n'avait guère de variance commode que si l'on prenait un segment formé de *strates entières*.

Comme il s'agit d'évaluation en 1re approximation, nous préférons de refaire le calcul de Durbin dans un univers qui ne comprenne ni strates, ni unités du 2e degré. Considérons un échantillon bernoullien de n ménages sur N; soit x le nombre d'enfants par ménage et (u) une variable égale à 1 si le ménage a une machine à laver (et à 0 s'il n'en a pas). Le problème est d'évaluer la variance de l'estimation du nombre moyen $\bar{z}$ d'enfants dans les ménages *qui ont une machine à laver* (tel est le « segment » dont on parlait plus haut, défini par $u = 1$).

Soit $\mathcal{E}\,u = p$ (p est la proportion de ménages ayant une machine à laver).

Considérons la variable $y = xu$, c'est-à-dire :

$$y = x \text{ si } u = 1 \text{ ou } y = 0 \text{ si } u = 0; \text{ soit } \bar{y} \text{ sa moyenne.}$$

Le sondage nous donne des estimateurs *sans biais* $\bar{y}'$ de $\bar{y}$ et p' de p, et l'estimateur-ratio $\bar{y}'/p'$ de l'inconnue $\bar{z}$.

Ce ratio a une variance bien connue.

Ici DURBIN fait *une approximation*. Il confond p' et p; d'où :

$$\mathcal{V}\left(\frac{\bar{y}'}{p'}\right) \# \frac{1}{n}\,\frac{\mathcal{V}y}{p^2} = \frac{1}{np^2}\left(\frac{\Sigma\,y_i^2}{N} - \bar{y}^2\right)$$

(1) Théorie récente de Tukey, Hooke, Robson, etc... qui sort du cadre de cet exposé.

Il oppose à cette approximation la valeur erronée qu'on obtient si on écrit

$$\frac{1}{np'}\sigma^2 \# \frac{1}{np^2}\left(\frac{\Sigma\, y_i^2}{N} - \frac{\overline{y}^2}{p}\right)$$

np' désignant le nombre n' de ménages échantillon ayant une machine à laver; z désignant une variable égale à x si $u = 1$, et *non définie* si $u = 0$.

La comparaison est facile; plus p est petit, plus $\overline{y}^2/p$ est supérieur à $\overline{y}^2$, plus la variance erronée est faible à côté de la variance correcte, plus l'approximation faite se justifie.

E) *Méthode de la superpopulation à structure particulière*

L'idée de considérer les populations soumises au sondage comme de grands échantillons de superpopulation (de taille infinie ayant une certaine loi de structure) est une idée qui commence à s'imposer à notre esprit. (Cochran en aurait la paternité).

Des Raj [25] la reprend avec l'hypothèse de structure assez banale :

— régression linéaire de la variable étudiée x par rapport à une variable auxiliaire comme y;
— variance de la loi liée proportionnelle à une certaine puissance g de la variable auxiliaire y.

Il calcule alors les espérances mathématiques des variances d'échantillonnage; et ses résultats, valables seulement *en moyenne* pour les populations finies, sont du genre de ce qui suit :

Si $g \geqslant 1$ le sondage avec probabilités proportionnelles à la taille y est meilleur que le sondage avec probabilités égales, même avec stratification (fractions sondées égales);
si $g < 1$ le sondage stratifié à fractions égales l'emporte.

De toute façon le sondage stratifié « à la Neyman » est meilleur que le sondage avec probabilités inégales.

En pratique, on pourrait appliquer ces résultats, sous réserve de s'être d'abord fait une idée de *l'ordre de grandeur de g* (et de ne pas avoir affaire à des populations qui refusent ce mode de traitement). Ils auraient leur place entre les comportements extrêmes consistant l'un à choisir à l'aveuglette probabilités égales ou inégales, — l'autre à faire des calculs numériques *massifs* sur des données antérieures ou voisines avant de se décider.

F) *Sondages jumelés*

Tirages avec remise et probabilités inégales

Lahiri a traité un problème pratique nouveau, apparemment insoluble : étant donné qu'on doit tirer un échantillon de villages avec probabilités proportionnelles à la population humaine (pour un sondage démographique) et un échantillon de villages avec probabilités proportionnelles à la superficie cultivée (pour un sondage agricole), s'arranger pour que ces 2 échantillons coïncident presque totalement pour employer les mêmes enquêteurs. Nous n'avons pas lu le papier de Lahiri; mais Des Raj dans Sankhya [26] améliore ses résultats. En outre il retrouve un problème traité dès 1951 par Keyfitz (et qui n'avait guère attiré notre attention), et aussi un problème de Goodman et Kish (1950) sur lequel nous avions travaillé [voir 27]. Au Canada Keyfitz, disposant d'un échantillon de comtés tiré avec des probabilités proportionnelles à leurs populations passées se proposait de lui substituer un échantillon quasi identique mais tiré avec les populations du nouveau recensement; (de manière à avoir le moins possible besoin de modifier le réseau d'enquêteurs) : c'était le même problème.

Quant au problème de GOODMAN et KISH (qui est lié à l'amélioration des estimations régionales, voir IB ci-dessus), DES RAJ ne le fait pas avancer; mais nous ne serions pas surpris qu'il reçoive bientôt ainsi une solution.

La question revient en définitive à garnir les cases d'un tableau (qui n'est plus carré dans le dernier cas) dont on se donne les *marges* et qu'on soumet à la condition d'en minimiser une certaine combinaison linéaire (comparable à un coût). C'est un programme linéaire de transport classique (c'est d'ailleurs aussi un problème traité par M. FRECHET par d'autres moyens). DANTZIG a donné une méthode de résolution qui consiste à partir d'un programme concentré au plus près de la 1re diagonale du tableau (coût nul) et à procéder ensuite par itération; cette méthode vaut quel que soit le tableau des coûts de déplacement entre le village i et le village j; mais on démontre que le programme de départ est ainsi le meilleur quand le tableau des coûts est garni de coûts proportionnels à $i - j$.

Or ce n'est pas là une hypothèse absurde si l'on a soin de numéroter les N villages indiens de 1 à N en serpentant toujours d'un village à un village *contigu*. Finalement la méthode empirique de LAHIRI (un peu simplifiée) est (pour DES RAJ) la méthode *optimum*.

G) *Tirages sans remise avec probabilités inégales.*

1° *L'article de* MURTHY *dans un tout récent fascicule de Sankhya* [*30*]

Les inventeurs du tirage au sort avec probabilités inégales (HANSEN, HURWITZ, 1943) s'en étaient tenus prudemment au tirage d'une seule unité par strate. Si non il faut accepter ou bien que la même unité puisse être tirée plusieurs fois, ou bien que les probabilités de tirage soient modifiées après chaque tirage (et différemment suivant les tirages) ce qui complique notablement les calculs; en pratique on aurait intérêt à savoir tirer, disons, *2 unités* différentes (encore qu'on n'ait pas à craindre d'erreur grave à procéder de façon un peu incorrecte, ce qui est usuel).

Nous ne reviendrons pas sur les travaux de HORVITZ et THOMPSON (J. A. S.A. 1952) de YATES et GRUNDY (J. R. S. S. 1953 B) et de DURBIN (*ibidem*) provoqués par un article de MIDZUNO (1950). Voir par exemple [17].

En fait la littérature est plus riche encore, car SINGH et SAXENA [28], dans leur communication à l'IIS, citent également une thèse de SEN (1952) et un article de SINGH (1954).

Nous venons de lire un article de MURTHY [29] qui apporte peut-être le mot de la fin dans cette controverse. Il distingue les estimateurs qui dépendent de *l'ordre* de tirage des unités (chaque ordre ayant une certaine probabilité) et l'estimateur, espérance mathématique des précédents, ne dépendant plus de l'ordre de tirage et de variance minimum. Il retrouve ainsi l'estimateur d'HORVITZ et THOMPSON et aussi celui de MIDZUNO (le schéma de MIDZUNO consiste à prendre au 2e tirage d'égales probabilités).

Les critiques de YATES n'ont donc de portée réelle que contre *l'estimateur de variance* de HORVITZ et THOMPSON, sans biais mais si mauvais qu'il lui arrivait d'être négatif. MURTHY apporte un estimateur de variance, qui écrase à son tour celui de YATES sur le propre exemple fabriqué par YATES.

2° *La solution de* STEVENS [30]

STEVENS (1) a tourné la difficulté en admettant qu'il est possible de regrouper les unités de sondage constituant l'univers, en groupes d'unités ayant même probabilité (à l'intérieur du groupe); alors si une unité est tirée 2 fois, il lui adjoint une unité *du même*

(1) Stevens enseigne à Sao Paûlo (Brésil) où il a fait aussi un projet de sondage sur la récolte du café (*Journ. Amer. Stat. Assoc.* sept 1955, p. 775).

groupe. L'estimateur correspondant est sans biais. La variance conserve son aspect habituel mais il faut lui retrancher un terme correctif $\Sigma N_i \sigma_i^2 (n - 1)/n$ où σ_i désigne l'écart type de la variable à l'intérieur *du groupe* (de même probabilité). L'estimateur de cette variance est lui aussi (tous calculs faits) très élégant.

Malheureusement cette solution ne présente guère d'intérêt pratique.

S'il s'agit de tirer au sort dans une strate 2 ou 3 communes rurales ou petites communes urbaines, la constitution de groupes de communes de même taille (donc de même probabilité) est fort aisée; mais c'est justement le cas où l'on ne tire presque jamais 2 fois la même commune (événement très improbable).

En revanche une difficulté pratique réelle est constituée, disons, par les communes de banlieue de Paris qui, classées en strates suivant des caractères sociologiques (pourcentage de population ouvrière etc...) sont de tailles très différentes; ceci tient au caractère tout à fait artificiel du découpage en communes d'une banlieue formant un ensemble continu. Si lorsqu'on a tiré 2 fois une grosse commune on lui substitue une commune plus petite, on introduit une erreur systématique qui peut être sensible; mais pour appliquer la méthode de STEVENS, encore faudrait-il que la strate de 4 communes, dont on doit tirer 2 communes au sort, en comprît bien 2 grosses et 2 petites. Ce n'est pas forcément le cas.

Exemple réel :	Levallois 27 000 ménages Courbevoie 21 400 — Colombes 21 400 — La Garenne 9 300 —	Si La Garenne est sur-représentée dans l'échantillon, et si l'on combine les résultats de nombreux sondages successifs, on aura une déformation systématique dans l'image de la population active.

H) *Sondages à 2 degrés — Les unités du 1er degré étant tirées avec des probabilités inégales.*

1° *Introduction*

Il y a quelque chose de très *artificiel* dans le problème agité en G ci-dessus : car les tirages avec probabilités inégales ne sont le plus souvent que le 1er degré d'un sondage à plusieurs degrés; alors on n'emploie pas d'estimateur tels que Sy_i/P_i, car on a d'autres moyens (grâce au 2e degré de sondage) d'éviter les biais.

Soit à tirer 2 communes au sort et 30 ménages par commune; supposons qu'on ait tiré A deux fois de suite; il n'y a qu'à faire un 3e tirage qui donnera (disons) la commune B. On n'aura *pas de biais* si l'on tire alors au 2e degré 40 ménages de A et 20 de B. Les difficultés n'apparaissent qu'au moment du calcul *de variance* (si l'on se permet de modifier ainsi le plan de sondage compte tenu des résultats des premiers tirages au sort, la variance ne présente plus les caractères d'une mesure des pertes d'information; la variance n'a plus de signification précise).

Ne peut-on s'arranger pour ne jamais tirer 2 fois la même commune? la même unité primaire?

2° *La méthode de* HAJEK

HAJEK [11] a une façon originale de tourner la difficulté : il ne s'impose pas à l'avance le nombre (exact) d'unités du 1er degré à tirer, il se donne seulement l'espérance mathématique de ce nombre (1). Soit P_i la probabilité de tirer l'unité i; $\Sigma_i P_i$ est égal à la dite espérance.

Considérons sur un axe des temps t les unités de sondage alignés en $t = 1,2, ...i ...N$. L'inclusion de l'unité (i) dans l'échantillon résulte de N tirages indépendants avec des proba-

(1) On était déjà habitué à substituer *au coût* d'un sondage son espérance mathématique pour certains calculs d'échantillons « optimum » (cas du sondage à 2 degrés avec probabilités égales, notamment).

bilités P_i variant avec $t = i$. C'est un processus *stochastique*, que HAJEK appelle processus de Poisson. Son intérêt est de conduire à des formules d'estimation et de variance d'une extrême simplicité.

Ainsi l'estimateur

$$\hat{Y} = S\ \hat{y}_i/P_i$$

a pour variance

$$\hat{Y} = \Sigma\ y_i^2 \left(\frac{1}{P_i} - 1\right)$$

Passons au sondage à 2 degrés. Soit Q la probabilité pour chaque unité du 2e degré d'être tirée; on a

$$\frac{n_i}{N_i} - \frac{Q}{P_i}$$

Soit $\hat{Y}_i$ l'estimation de y_i dans l'unité tirée i (du 1er degré); on a un estimateur (à 2 degrés)

$$\hat{Y} = S\ \hat{Y}_i/P_i$$

avec (cette fois) :

$$\mathcal{V}\ \hat{Y} = \Sigma\ y_i^2 \left(\frac{1}{P_i}\right) + N_i\ \sigma_i^2 \left(\frac{1}{R} - \frac{1}{P_i}\right)$$

A présent on peut se demander si l'on ne perd pas d'un côté ce qu'on gagne de l'autre.

Lorsqu'on a découpé l'univers en strates très nombreuses, on ne tirait plus que 2 unités par strate. Si l'on adoptait le point de vue de HAJEK, c'est à-dire si l'on se contentait de choisir les P_i de façon à tirer *en moyenne* 2 unités par strate, il y aurait beaucoup de strates qui ne fourniraient à l'échantillon qu'une unité, ou même pas d'unité du tout. Il ne resterait plus qu'à recommencer les N tirages; non seulement ce serait malcommode mais on peut douter que la formule de variance soit encore correcte, qu'elle ait encore un sens. (C'est là d'ailleurs un problème très général).

En outre si l'on veut estimer les variances, c'est *au minimum* 2 unités du 1er degré qu'il nous faut pour chaque strate. Le procédé de HAJEK devient encore moins praticable.

3o *La communication de* WILKS [31]

Les difficultés du sondage à 2 degrés avec probabilités inégales ne nous avaient pas échappé en 1947. Nous étions d'avis de pondérer par 2 les questionnaires (c'est-à-dire doubler les cartes perforées à la reproduction) si une commune était tirée 2 fois. M. HENRY a trouvé qu'il était plus satisfaisant de doubler le nombre de questionnaires à y collecter [32].

C'est son schéma que SUKHATME a retrouvé (cf son traité, [13] et que WILKS s'est proposé de perfectionner; c'est encore un travail de mathématicien et non de statisticien, les cas qu'il vise étant rares en pratique.

En effet le problème courant est par exemple de tirer au sort 2 communes et 25 ménages par commune; par hypothèse aucun village n'a moins de 25 ménages ni même en fait moins de 50. La méthode de HENRY et SUKHATME consiste à prendre 25 ménages d'une commune tirée 1 fois, ou 50 ménages d'une commune tirée 2 fois. Il pourrait arriver qu'on tire 2 fois une commune de 30 ménages; mais ou bien de telles communes sont rares et on a alors bien peu de chances de la tirer, ne fût-ce qu'une fois; — ou bien la strate comprend un nombre appréciable de communes aussi petites, mais alors il est invraisemblable qu'on court le risque d'avoir à interroger presque tous les gens de la commune. La solution normale est, dans la première hypothèse, de rattacher les communes trop petites aux communes

les plus proches, — et dans la seconde hypothèse de choisir un plan de sondage tel que chaque commune tirée ne fournisse (disons) que 10 ménages-échantillon.

Or la méthode de WILKS consiste justement à supposer que chaque commune contient un nombre entier de fois 25 ménages, soit M_i fois, à placer dans un chapeau $M_1 + M_2 + \ldots + M_i + \ldots$ billets et à en tirer *n sans remise*. Chaque commune est tirée A_i fois (en général 0 fois) et le nombre de ménages à en extraire par tirage au sort est 25 A_i.

La méthode de SUKHATME (avec cette même hypothèse) revenait à faire *n* tirages avec *remise*.

On risquait d'avoir parfois A_i supérieur à M_i; avec le schéma de WILKS c'est devenu impossible. Ce serait intéressant si le risque était grand; par exemple avec un échantillon (disons) de 25 % de la population, le risque existe, — ou encore avec des strates peu nombreuses et un nombre de tirages *n* très élevé; mais ce n'est pas là le genre de sondage qu'on fait souvent en pratique.

Donc WILKS substitue une distribution *hypergéométrique* à une distribution *multinomiale*; les calculs sont un peu plus longs. Le tirage effectif, sans billets de loterie mais sur une liste de totaux cumulés serait facile. Le résultat de WILKS est élégant : la variance a l'aspect habituel, mais avec un 3e terme soustractif, — à un facteur près :

$$\sigma_w^2 + m\,\sigma_b^2 - \delta$$

avec

$$\delta = (m - 1)\,\Sigma\,\sigma_i^2/U - 1)$$

où m est le nombre de ménages au 2e degré, soit 25 dans l'exemple précédent;
U le nombre de ménages que comprend au total la population;
σ_b^2 la variance entre communes (b = *between*);
σ_w^2 la variance à l'intérieur des communes (w = *within*);
σ_i^2 la variance à l'intérieur de la commune i.

En revanche les estimateurs de ces 3 composantes (surtout de la 2e) sont compliqués.

Il ressort des papiers de WILKS et HAJEK qu'on n'est pas parvenu jusqu'ici à éviter complètement de tirer 2 ou 3 fois la même commune.

I) *Détermination des probabilités optimum dans un sondage avec probabilités inégales*

La recherche du jeu optimum de probabilités à affecter aux unités de sondage est un problème *académique*, que HANSEN et HURWITZ n'avaient traité qu'en 1949 (*Ann. Math. Stat.* p. 426) et seulement dans le cas particulier des tirages avec remise. Et d'abord *de quel optimum* s'agit-il? HAJEK [11] et DES RAJ [33] ont traité en fait des problèmes bien différents quand ils se sont proposé de compléter le travail de 1949.

1° DES RAJ se propose de tirer un couple (i j) d'unités primaires avec une probabilité P_{ij}, satisfaisant à des conditions qui sont celles d'un *programme linéaire* classique :

a) La probabilité P_i de tirer (i) quel que soit (j) est donnée (proportionnelle à la taille y de l'unité i).

b) La variance de l'estimateur de HORVITZ et THOMPSON (voir ci-dessus G) pour tout caractère étudié X ne saurait bien entendu être rendue minimum; mais on cherchera à le faire dans le cas où X est fonction linéaire de y :

$$X = a + b\,y.$$

On s'aperçoit que, quels que soient a et b, la dite variance est minimum en même temps que

$$\Sigma \Sigma P_{ij}/P_i P_j$$

expression qui joue le rôle du *coût* linéaire à minimiser (dans la théorie des programmes linéaires).

Des Raj a repris en particulier l'exemple de Yates et Grundy (voir G) et montré que ceux-ci étaient très éloignés de l'optimum (ils ne prétendaient d'ailleurs pas être parvenus à l'optimum).

2° Hajek a des soucis beaucoup plus pratiques et plus proches de ceux de Hansen et Hurwitz. C'est moins le sondage à 1 degré avec probabilités inégales qui l'intéresse que le sondage à 2 degrés dont le 1er degré est avec probabilités inégales; et pour lui l'optimum ne consiste pas à réduire la variance pour un échantillon de taille 2, mais à réduire *la variance du sondage à 2 degrés* pour un coût de sondage donné.

D'ailleurs c'est *l'espérance du coût* qu'il utilise. Son théorème sur la stratégie optimum s'applique, vu que sa variance et son coût ont l'expression canonique (ci-dessus III A) vis-à-vis des paramètres qui sont ici :

les P_i, probabilités de tirer l'unité (i) de sondage (par un processus de Poisson);

et Q, probabilité de tirer une unité du 2e degré par les 2 degrés de sondage réunis.

C'est donc comme une conséquence immédiate de sa grande théorie, que Hajek trouve les P_i optimum.

J) *Le Sondage Systématique*

Sur la théorie du sondage systématique (c'est-à-dire le fait extrêmement banal de constituer l'échantillon avec des unités de sondage dont les numéros d'ordre forment une progression arithmétique), nous avons rencontré des contributions de Hajek [11] et de Gautschi [34].

1° Pour Hajek (1), le sondage systématique est la stratégie optimale du sondeur lorsque les valeurs $y_1\, y_2 \ldots\, y_N$ (que la variable étudiée prend sur les unités de sondage de la population) sont supposées former une suite stationnaire dont la fonction de corrélation est convexe (concavité du corrélogramme tournée vers le haut). De même, le sondage stratifié « à la Neyman » est la stratégie optimale lorsque les valeurs $y_1\, y_2 \ldots\, y_N$ sont présumées indépendantes.

Hajek généralise ainsi un théorème de Cochran (*Ann. Math. Stat.* 1946, p. 164).

2° Ce même théorème de Cochran est étendu par Gautschi à ce qu'on peut appeler le sondage *semi-systématique* (méthode due à Tukey) : l'échantillon est cette fois constitué de plusieurs progressions arithmétiques (ayant même raison mais partant d'éléments aléatoires distincts). Là encore il s'agit de savoir quel mystère a présidé à l'attribution des numéros d'ordre 1, 2... N aux unités de la population; si ces numéros avaient été tirés au sort, les sondages systématiques et semi-systématiques équivaudraient au sondage aléatoire. Le *corrélogramme*, habituellement destiné à l'analyse des séries temporelles, est d'usage ici, comme si (1,2...N) étaient des dates t (interférence avec les processus stochastiques). Et on est conduit à s'intéresser aux cas où ce corrélogramme jouirait de propriétés assez particulières.

(1) Comme en H 2 ci-dessus, un temps fictif $t = 1, 2. \ldots N$ permet de décrire la population dont on suppose cette fois que $y(t)$ est fourni par tirage au sort dans une urne V (t).

Ainsi, s'il est convexe (concavité par le haut), GAUTSCHI trouve que le sondage systématique est meilleur que le semi-systématique (ce qui confirme l'énoncé de HAJEK). Mais moins le sondage est systématique et moins il est bon; l'accroissement de variance dû à l'emploi de 2 progressions arithmétiques est minime, mais avec 10 progressions arithmétiques il devient colossal.

GAUTSCHI signale que JONES avait étudié le même problème (1955-1956); mais nous n'avons pas lu ses travaux.

CONCLUSION

La méthode des Sondages semble s'enrichir présentement d'apports de la théorie des plans d'expérience, de celle des programmes linéaires, de celle des processus stochastiques.

La principale difficulté réside dans le fait que les chercheurs du début semblent avoir de nouvelles préoccupations et que les chercheurs actuels sont dispersés à présent sur toute la surface du globe.

Pierre THIONET.

BIBLIOGRAPHIE

[1] KEYFITZ. — *Calculation of variances in a monthly population survey.* Bull. Inst. Int. Stat. XXXV-2, p. 181 (1955 1958).

KEYFITZ. — *Estimates of sampling variance where two units are selected from each stratum.* Journ. Amer. Stat. Assoc. Dec. 1957 p. 503.

[2] KOLLER. — *On the problemes of replicated sampling in German Governmental Statistics.* — Congrès I. I. S. Stockholm 1957, (papier 102).

DALENIUS (Tore). — *Possibilities and limits of sampling in regional inquiries* (1955).

[2*bis*] KOLLER. — *The use of prior statistical information in problemes of estimation* — Congrès I. I. S., Bruxelles 1958 (papier 59).

[3] TAYLOR et CLEMENT. — *The New Zealand agricultural sample survey,* Journ. Royal Stat. Soc. A. 1956. 4 p. 409.

[4] D. COLE et UTTING. — *Estimating expenditure, saving and income from household budgets.* J. Roy. Stat. Soc. A 1956, p. 371.

PRAIS. — *Some problemes in the measurement of price changes with special reference to the cost of living.* J. Roy. Stat. Soc. A 1958, p. 312.

KEYFITZ. — *The design of surveys to provide experimental contrasts* (mimeograph. 1958).

[4*bis*] DEMING. — *On sampling by a system of replicated drawings with equal probabilities,* etc. J. Amer. Stat. Assoc. Mars 1956.

Communications au Congrès de l'Institut International de Statistique Stockholm, 1957. KOLLER. — *On the problems of replicated* etc... (papier 102).

Señorita FLORES. — *The theory of duplicated samples and its use in Mexico* (papier 113).

LAHIRI. — *Recent developments in the use of techniques...* in India (papier 75).

[5] ASHFORD. — *The design of a long term sampling programme to measure the hazard associated with an industrial environment* J. Roy. Soc. Stat. A 1958, 3, p. 333.

[6] CHAPMAN et JUNGE. — *The estimation of the size of a stratified animal population,* Ann. Math. Statis. June 1956, p. 375.

[7] BROWNLEE. — *A note on the effects of non response on surveys,* Journ. Amer. Stat. Ass. March 1957, p. 29.

[9] BLACKWELL et HODGES. — *Design for the control of selection bias,* The Annals of Math. Stat. June 1957, p. 449.

[10] THIONET. — *Décisions à propos de Sondages* — Revue de Statistique appliquée 1955.

[11] HAJEK. — *Some Contributions to the theory of probability sampling* Congrès de l'I. I. S. Stockholm, 1957 (papier n° 60).

[12] KHADJENOURI. — *Thèse de l'Université de Paris,* 1956.

[13] SUKHATME. — *Sampling theory of surveys with applications,* 1954.

[14] BILLETER. — *Optimum design in mixed sampling plans,* (Revue de l'I. I. S. 1956, p. 73).

[15] THIONET. — *Un problème de sondage parmi des éléments dont la distribution est très dissymétrique.* — Journal de la Société de Statistique de Paris, Juillet Sept. 1955, p. 192.
Paru aussi dans les Cahiers de l'ADETEM sous une forme simplifiée.

[16] DALENIUS. — *Sampling in Sweden* (1957) (thèse de doctorat).

[17] THIONET. — *Théorie des sondages* (Cours I. S. U. P.) (édition de novembre 1955).

[18] DALENIUS et HODGES. — *The choice of stratification points* (Skandinavisk Aktuarietidskrift 1957, 3-4, p. 198.)

[18*bis*] DALENIUS et HODGES. — *Minimum variance stratification,* Journ. Amer. Stat. Assoc. March 1959, p. 88.)

[19] KITAGAWA. — *Some contributions to the design of sample surveys Part VI*, Sankhya 17, 1, 1956, p. 27.
[20] DES RAJ et KHAMIS. — *Some remarks on sampling with replacement.* Annals of Math Stat. June 1958, p. 550.
THIONET. — *Comparaison à effectifs égaux entre échantillon bernoullien et échantillon exhaustif* (mimeog) (20 mars 1956). Repris dans l'Étude théorique n° 7 de l'INSEE (1959).
[21] HAJEK. — *On the theory of ratio estimates*, congrès de l'I. I. S. 1958, Bruxelles. Papier n° 33.
[22] ROBSON. — *Application of multivariate polykays to the theory of unbiased ratio type estimation.* Journ. Amer. Stat. Assoc. Dec. 1957, p. 511.
[23] GOODMAN et HARTLEY. — *The precision of unbiased ratio type estimators*, J. Amer. Stat. Assoc. June 1958. page 491.
[24] DURBIN. — *Sampling theory for estimates based on fewer individuals than the number selected*, Congrès de l'I. I. S. 1957, Stockholm, papier n° 106.
[25] DES RAJ. — *On the relative accuracy of some sampling techniques*, Jour. Amer. Stat. Assoc. March 1958, p. 98.
[26] DES RAJ. — *On the method of overlapping maps in sample surveys*, Sankhya 17. 1, 1956 page 89.
[27] THIONET. — *Étude théorique n° 6 de l'INSEE* 1953 page 84 à 93.
[28] SINGH et SAXENA. — *Congrès de l'I. I. S.*, 1955, Rio de Janeiro 2e (p. 163).
[29] MURTHY. — *Ordered and unordered estimators in sampling without replacement*, Sankhya 18,3 4, (1957).
[30] STEVENS. — *Sampling without replacement with probability proportional to size*, J. Royal Stat. Soc. B 20, 2 (1958), p. 393.
[31] WILKS. — *A two stage scheme for sampling without replacement*, Congrès de l'I. I. S. 1958, Bruxelles, papier n° 25.
[32] HENRY. — *Journal de la Société de Statistique de Paris*, Oct.-Déc. 1948.
[33] DES RAJ. — *A note on the determination of optimum probability in sampling without replacement*, Sankhya, 17, 2 p. 197, (1956).
[34] GAUTSCHI. — *Some remarks on systematic sampling*, Annals of Math. Stat. June 1957 p. 385. (Jones : J. ASA 1955, p. 763 et 1956, p. 54)