JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Théorie des sondages : quelques problèmes récents

Journal de la société statistique de Paris, tome 108 (1967), p. 9-30

<http://www.numdam.org/item?id=JSFS_1967__108__9_0>

II

THÉORIE DES SONDAGES : QUELQUES PROBLÈMES RÉCENTS

1. Préambule

Voici le vingt-cinquième anniversaire de mon entrée dans la statistique enseignante (c'est-à-dire du jour où je fus chargé des Travaux pratiques de M. le professeur Fréchet) : qui ne précéda que de quelques mois mon entrée dans la statistique militante (comme dit mon ami Schützenberger), c'est-à-dire mon entrée à la Statistique générale de la France où m'accueillit si cordialement M. Bunle — lequel m'enrôla bien vite à la Société de Statistique de Paris. Ma première communication eut lieu peu après la libération de Paris, sur l'École italienne de Statisticiens de Corrado Gini, thème qui m'avait été fourni par M. Michel

Huber, et sur lequel j'avais travaillé un ou deux ans (apprenant au passage à lire l'italien). J'évoque à cette occasion la mémoire de notre secrétaire général de l'époque, M. Barriol, et du président Leprince-Ringuet (père de l'académicien, qui lui ressemble fort), sans oublier celle de mon chef de bureau, plus tard mon directeur, M. Raymond Rivet; tous se dépensaient pour faciliter mes débuts dans la carrière statistique (reconversion d'un professeur agrégé de lycée et non d'un officier comme c'était plus souvent le cas alors). Malheureusement mon papier, rédigé trop vite, fit très mauvaise impression en Italie, alors que j'avais été dans l'ensemble fortement impressionné par tout ce que j'avais lu, comme le montrera l'anecdote suivante : M. Sauvy m'adressa un soir un dominicain, le révérend père Lebret, qui m'entretint, non du salut de mon âme, mais de représentation graphique des statistiques; il était alors passionné pour les profils, et j'avais justement appris à l'école du maître italien que la méthode des profils peut induire dangereusement en erreur par ses illusions d'optique; j'écoutai poliment le révérend et ne donnai aucune suite à son invitation de l'aller voir à Écully (près de ma bonne ville de Lyon). Il est mort récemment, vous l'avez probablement lu dans la presse. Nos routes ne se rencontrèrent jamais plus, bien que voisines : d'abord même passion pour les enquêtes statistiques, puis pour la planification économique.

C'est aussi par M. Sauvy que je fus amené à faire une étonnante rencontre, celle d'un biologiste, le professeur André Mayer qui rapportait des États-Unis les matériaux pour faire un livre sur les Sondages. Nous fêtons aussi le vingtième anniversaire de ce livre, mince plaquette publiée chez Hermann en 1946 aux frais du professeur Mayer [1]. Je l'avais composée à la hâte l'hiver précédent; la correction d'épreuves (pendant un séjour dans l'Administration militaire en Allemagne) a laissé bien des erreurs typographiques et a oublié la préface. Le livre est épuisé, c'est-à-dire que ses invendus sont allés au pilon; l'éditeur en refusa la traduction en espagnol, dans l'espoir assez vain de mieux vendre son stock. Quoi qu'il en soit, ce mince ouvrage (trop dépourvu de démonstrations mathématiques), en avance de deux ans sur les manuels en anglais, me fit sacrer sondeur et fut à l'origine des décisions de M. le directeur général Closon de me confier en 1947 l'enseignement et la préparation des sondages à l'I. N. S. E. E.

Ce long préambule historique n'est peut-être pas tout à fait inutile; car, si j'ai beaucoup écrit dans le Journal de la Société de Statistique de Paris, j'en ai peu fréquenté les réunions. Et même, depuis 1960, ayant des cours à Poitiers le mercredi, il me faut une autorisation d'absence du recteur. Bien entendu, c'est pour parler des sondages que je suis venu ce soir.

Il convient maintenant de préciser que je ne délivrerai aucun message, ne prêcherai aucune croisade en faveur de telle ou telle méthode de sondage; je ne jetterai aucune excommunication ni aucun anathème (M. Sauvy me conféra jadis la dignité de pape des sondages).

Je crois encore ne pas venir en demi-solde, avec l'intention délibérée de jeter des pierres dans le jardin de l'I. N. S. E. E. ou — en bon ours — quelques pavés sur la tête de ses amateurs de jardins. Quand je quittai les sondages de l'I. N. S. E. E. en 1954, j'ai dit : ouf ! après quoi j'ai toujours cherché à trouver loisir d'en approfondir les aspects théoriques. M. le professeur Darmois, en trouvant tout naturel que je fasse une thèse [2] sur les Sondages, me fit diverger et revenir dans l'Université, mais cette fois dans l'enseignement supérieur. Celle-ci (je l'avoue) se désintéresse encore, en province, de la formation des statisticiens; mais elle a commencé à prendre conscience que les professeurs de mathématiques qu'elle forme pour le 2e degré devront bientôt initier à la Statistique et aux Probabilités les élèves de Seconde, Première, Classe terminale de certaines sections. Mais la réforme de l'enseignement n'est pas à l'ordre du jour de la présente réunion.

Vous savez tous, au moins par la presse, la radio, la télévision, que l'enseignement supérieur n'a pas seulement une tâche enseignante, mais aussi une tâche de recherche scientifique. Mes collègues ont tendance à qualifier leurs recherches de *fondamentales*, ce qui m'a tout l'air d'un mauvais jeu de mots; la Direction de l'Enseignement supérieur considère qu'il s'agit de *recherche légère*, par opposition à la *recherche lourde*, dont le coût se chiffre par 10^n francs lourds ($n \to +\infty$). Les artilleurs ici présents ont reconnu la distinction entre artillerie légère et artillerie lourde.

C'est essentiellement de telles recherches scientifiques que je compte vous entretenir; elles ne sont ni fondamentales, ni lourdes; assez légères du fait que nous n'avons aucun moyen de les appliquer, elles ne sont pas théoriques *stricto sensu*, car la mathématique pure les vomit comme trop impures. Si on ne peut les qualifier de recherches appliquées, elles sont (espérons-le) *applicables*. Leur éventail en est largement ouvert, suivant la répulsion plus ou moins grande que suscitent chez le chercheur les possibilités d'application des mathématiques.

Parmi les recherches auxquelles je m'intéresse, du fait de mon obstination naturelle et aussi sans doute d'une certaine paresse intellectuelle, une partie est restée axée sur les Sondages. La recherche comprend bien entendu, d'abord un travail de documentation, qui ici est très long et n'est pas toujours passionnant : il s'agit d'essayer de se tenir au courant des travaux des collègues. En ce qui me concerne, ces travaux sont presque exclusivement écrits en anglais. Le travail consiste, outre la production d'études personnelles, à faire un cours de 3e cycle (à Paris) et à diriger (« séminaire ») la préparation de mémoires et thèses de quelques chercheurs.

On peut à peine parler d'équipe; en tous cas, nous ne faisons concurrence ni à l'I. N. S. E. E. ni à quiconque en France à notre connaissance (pas même au C. N. R. S. et c'est bien dommage).

2. Nature de l'exposé qui suivra

On pouvait envisager, pour l'exposé de ce soir, que j'enchaînerais sur l'un ou l'autre de mes deux articles de 1959-1960.

[3] Développements récents de la théorie des Sondages, *J. S. S. P.*, octobre-décembre 1959, pp. 279-295.

[4] Quelques aspects de la théorie des Sondages, *J. S. S. P.*, avril-juin, 1960, pp. 99-111.

Dans le premier cas, nous serions gêné par les articles analogues publiés par Dalenius (1962) [5], par Murthy (1963 [6],— voir cette année. Il y a encore un papier de Godambe (1965) [7] qui, sans chercher à présenter tout ce qui s'est fait en sondages, couvre déjà pas mal d'articles. A moins que nous ne pillions ces « reviews » elles-mêmes, car nous n'avons pas pu tout lire.

Dans l'autre cas, ce serait que nous avons fait faire quelque progrès notable à la théorie des sondages ces derniers temps, et que nous venions vous l'exposer; or ce n'est absolument pas le cas. D'ailleurs, si nous avons publié quelques études (depuis 1960), nous n'entendons pas les démarquer ici, tout au plus les signaler.

Nous allons donc adopter une démarche intermédiaire, nous promenant à travers les grands chapitres d'un cours de sondage, — nous arrêtant çà et là au gré de l'actualité, quand nous croirons pouvoir dire autre chose que des banalités.

3. Les « mauvais échantillons »

On avait envisagé de vous entretenir exclusivement d'expériences faites sur l'utilisation d'un mauvais échantillon. Nous avons procédé à une telle expérience en 1963 et en avons parlé [8] au Congrès de l'Institut international de Statistique de Belgrade (1965); et actuellement nous dirigeons la préparation d'une thèse de l'I. S. U. P. sur ce même sujet, mais concernant un échantillon d'exploitations agricoles (et non plus des entreprises commerciales). On entend par « mauvais échantillon » un échantillon qui nous est donné tel quel — dont on peut toujours retrancher des éléments (quitte à perdre des informations) mais auquel on n'est en mesure d'ajouter aucun élément. On peut lui adjoindre seulement des informations sur l'univers sondé. Cette thèse n'est malheureusement pas encore soutenue; lorsqu'elle le sera (bientôt peut-être), son auteur viendra vous en parler lui-même. Il serait prématuré de le faire aujourd'hui — et nous nous en tiendrons aux sondages organisés comme tels.

4. Les sondages proprement dits : sondages par « quotas » — sondages « probabilistes »

Nous ne vous ferons pas l'injure de croire que vous pourriez ignorer la distinction essentielle entre échantillons dont les éléments sont choisis par les enquêteurs sous contrôle d'un plan de répartition (disons : sondages *par quotas*) et les échantillons dont les éléments sont désignés par tirage au sort dans le cadre d'un plan de sondage qui, lui, est choisi par l'organisateur (sondages « probabilistes »).

5. Sondages par quotas

1 — Les applications qu'on fait du Calcul des Probabilités aux sondages par quotas sont en réalité d'une interprétation délicate, en ce sens que le calcul fournit des conclusions, dans l'hypothèse où un tirage au sort aurait donné justement cet échantillon dont nous disposons et qui, en fait, a été obtenu autrement. A vrai dire la littérature en ce domaine est très réduite. Signalons pourtant l'article de Sudman (Seymour) 1966 [9] sur une théorie probabiliste du sondage par quotas, accompagnée de résultats numériques comparés entre sondages des 2 types (sur les mêmes sujets) effectués aux États-Unis.

2 — Voici à présent une étude théorique faite à Paris et concernant les sondages par quotas.

Un de nos étudiants de Paris, M. Chahine, a soutenu en janvier 1966 une thèse (de 3e cycle) (non publiée) sur un sujet qui lui a été fourni par sa profession de statisticien dans une entreprise importante de *marketing*. Vous prescrivez à un enquêteur de constituer un échantillon de n entreprises dont a de catégorie A, b de catégorie B, etc. (ce sont les *quotas*). On peut acheter à l'I. N. S. E. E. des listes d'entreprises : la catégorie A, B, n'est pas indiquée sur la liste. Il se trouve que les entreprises A sont rares. *Problème :* A supposer qu'on tire au sort N entreprises sur la liste et qu'on les visite (N étant juste assez grand pour renfermer a entreprises A, b entreprises B, etc.), N est bien entendu une variable *aléatoire*. Quelle est sa loi de probabilité? Quelle valeur N a-t-elle 1 chance sur 20, 1 chance sur 100 de dépasser? En fait N peut atteindre facilement plusieurs fois $n = a + b + \ldots$ (son minimum). D'où

on conclut sans peine qu'en pratique on ne tirera pas au sort l'échantillon quand on lui imposera des quotas; ce serait ruineux.

Pour traiter ce problème, on utilise une loi hypergéométrique inverse, qui est à la loi hypergéométrique ce qu'est la loi de Pascal (binomiale négative) à la loi de Bernoulli. Une partie des résultats de M. Chahine a été publiée (fin 1965) dans la Revue de Statistique appliquée [10]. Nous lui avions proposé de faire calculer quelques résultats numériques par notre petit ordinateur poitevin, mais il a fallu se rendre à l'évidence que ses formules (d'ailleurs fort élégantes) défiaient les capacités d'un aussi modeste engin, et qu'il valait mieux utiliser des méthodes approchées et simplifiées (qu'on trouve d'ailleurs dans sa thèse).

6. Sondages « probabilistes »

A présent nous parlerons des seuls échantillons tirés au sort, dans le cadre d'un plan plus ou mieux complexe. Le sujet est à la fois trop technique (qu'on veuille bien nous en excuser), et trop vaste pour qu'il soit possible ici de tout passer en revue.

7. Les sondages systématiques

Le sondage systématique est, rappelons-le, un sondage en progression arithmétique dont l'un des éléments est tiré au sort. Nous avons publié une note à son sujet dans la Revue de Statistique appliquée 1965 [11] et une autre est sous presse [12]. La matière fait l'objet encore d'une importante étude de Törnqvist (1963) [13] et d'un article d'un canadien français M. Zinger (1963) [14].

Nous n'avons pas jugé utile d'entrer ici dans un examen (même sommaire) de ces articles trop techniques. Le fait de tirer un échantillon par le procédé en question et non pas *strictement* au sort est susceptible de perturber les erreurs d'échantillonnage dans une mesure qu'il n'est pas toujours facile d'apprécier; et ceci rentre dans le cadre des calculs de variance, dont on va parler un peu plus.

8. Les calculs de variance

a) On sait que la caractéristique essentielle des sondages probabilistes (ce par quoi ils se distinguent notamment des sondages par quotas) est qu'on peut en principe calculer quelle est en moyenne leur erreur d'échantillonnage, c'est-à-dire l'erreur affectant les résultats du sondage du fait qu'on ait substitué un simple échantillon à la population concernée.

Du moins est-ce notre point de vue d'occidental, car sur ce point, un expert officiel soviétique pour les Sondages, rencontré en 1963 à l'O. N. U., ne semblait guère convaincu.

b) Pour notre part, nous avions publié dans la R. S. A. (1965) une note sur les calculs de variance [15]. Il y est entre autres question de la méthode de calcul qui consiste à tirer d'un échantillon 10 sous-échantillons en tous points comparables et à comparer les 10 résultats pour juger de leur dispersion. Lisant un compte rendu du dernier livre de notre ami Deming [16] nous avons appris que cette méthode était due à Tukey. Rendons à César ce qui

est à César. De toute façon nous ne critiquons pas le principe de la méthode, qui consiste à réunir en un échantillon global 10 sous-échantillons fournis par un même plan de sondage; nous en critiquons un certain emploi pratique.

c) Que Deming nous excuse de ne pas parler ici des échantillons qui s'interpénètrent. Nous avons lu qu'il avait beaucoup fait (comme consultant en sondages) pour simplifier les méthodes de sondage (sans les mutiler) et ainsi les faire pénétrer dans le monde des affaires. Ce n'est pas facile, croyez-moi.

d) Nous avons (dans [15]) exprimé encore des craintes quant à l'incertitude des calculs de variance. Ceci mérite une certaine mise au point, et nous profitons des circonstances pour la faire ici.

d 1) Tout d'abord, nous raisonnons comme si les distributions dans lesquelles a lieu le sondage n'étaient pas trop différentes des distributions de Laplace-Gauss, N (μ; σ^2). C'est ce qui nous permet de dire qu'en première approximation une expression de la forme $\Sigma\ (x_i - \bar{x})^2$, somme étendue à n termes, est voisine de $\sigma^2 x^2$, où x^2a ($n - 1$) degrés de liberté. On pourrait certainement nous fournir des distributions (notamment celles qui ne possèdent pas de moment d'ordre 2) — disons une distribution de Pareto banale, où cette assimilation à $\sigma^2 x^2$ n'est pas justifiée. En principe la méthode des sondages devrait être *distribution-free*, ne devrait pas être tributaire d'une hypothèse précise sur les distributions sondées; en fait on suppose toujours implicitement que la distribution sondée admet des moments d'ordre 1, 2, 3, etc.

d 2) Nous avons retrouvé depuis, dans un article de Tate et Klett (1959) [17] — étranger aux sondages et où (par conséquent) l'hypothèse Laplace-Gaussionne est explicite — nos propres résultats sur la dispersion des variances estimées. Il s'agit des intervalles de confiance *optimals*, c'est-à-dire les plus *courts* possibles, pour l'estimation de variance par intervalles. Nous y lisons par exemple (n étant la taille de l'échantillon) :

$n = 20$ { $1 - \alpha = 0{,}90$; — Intervalle : 11,2586 — 32,3478 (Rapport 1 à 3)
{ $1 - \alpha = 0{,}95$; — Intervalle : 9,9579 — 35,2267 (Rapport 1 à 3,5)

($1 - \alpha$) est la probabilité que l'estimation appartienne à l'intervalle indiqué.

Les propos que nous tenons dans [15] sont en accord parfait avec [17], dont l'antériorité est évidente. Mais nous nous sommes aperçu que nous avions « dramatisé » la situation à l'excès; car si nous substituons aux variances les *erreurs-types* (leurs racines carrées) nous obtenons

$$\sqrt{9{,}9579} = 3{,}15; \qquad \sqrt{35{,}2267} = 5{,}93$$

L'écart entre ces deux limites n'est plus aussi scandaleux.

Après tout, le bon utilisateur du sondage ne va pas jusqu'à croire qu'un résultat affecté d'une erreur type de 3,1 est forcément meilleur qu'un autre affecté d'une erreur type de 5,9. Il est conscient que le calcul d'erreurs n'est jamais très précis, il est surtout conscient que ces erreurs d'échantillonnage sont souvent peu de chose à côté des erreurs d'observation et de dépouillement qui, elles, échappent au calcul.

En revanche il serait bon qu'on se souvienne que, dans la variance d'un sondage à deux degrés (échantillon de communes, échantillon de logements de ces communes) la variance entre communes est déterminée avec un nombre de degrés de liberté bien inférieur

à celui de la variance entre ménages à l'intérieur des communes. D'où il suit que ces deux composantes de variance sont évaluées avec d'inégales précisions. Il existe une théorie de la répartition optimale des ressources : faut-il beaucoup de communes, et quelques ménages par commune-échantillon? ou l'inverse? Dans la mesure où l'on emploie vraiment cette théorie (ce dont je ne suis pas tout à fait convaincu), on s'expose à des déboires, du fait qu'on oublie les erreurs affectant les paramètres du calcul, et les affectant de manière très inégale.

d 3) Dans le cas d'un échantillon stratifié, la variance n'est plus en général un *Ki-carré* mais une combinaison linéaire de *Ki-carrés* indépendants. Il ne semble pas, d'ailleurs, que cela change grand-chose (qualitativement) aux propos tenus sur l'imprécision de la variance estimée. Malheureusement les combinaisons linéaires de *Ki-carrés* n'ont pas encore mérité qu'on s'intéresse à elles au point d'en construire des tables numériques.

Il n'est peut-être pas trop tard pour que nous ajoutions à [15] le petit résultat mathématique que voici :

Le cas (en somme favorable) où l'on peut assimiler (sans trop se tromper) une variance estimée et une variable *Ki-carré* de Pearson, est (à peu de choses près) celui où l'échantillon stratifié est réparti entre les strates dans les proportions optimales telles que Neyman les a définies en 1934 (voir par exemple [18].

En effet

$$\mathscr{V}\bar{x} = \Sigma_h \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}$$

avec : estimation de $\sigma_h^2 = \Sigma_i (x_{hi} - \bar{x}_h)^2/(n_h - 1)$

avec : $\Sigma_i (x_{hi} - \bar{x})^2 = \sigma_h^2 \chi_h^2$ (en 1re approximation)

Confondant $(n_h - 1)$ et n_h et supposant (avec Neyman) n_h proportionnel à $N_h \sigma_h$, on voit que $V\bar{x}$ est proportionnel à $\Sigma_h \chi^2_h$ c'est-à-dire est un χ^2 à $\Sigma_h(n_h - 1)$ degrés de liberté cqfd.

9. Les sondages avec probabilités inégales

Nous étant ainsi expliqué dans la Revue de Statistique appliquée sur une philosophie en somme assez libérale des calculs d'erreur (sous réserve qu'on calcule délibérément des valeurs approchées par excès de la variance), nous sommes conscient que le praticien ne se croira plus guère tenu de tirer 2 communes de chaque strate, ce qui n'a d'intérêt réel que pour permettre le calcul de la variance entre communes. Dès lors, le praticien se désintéressera (c'est à craindre) du fameux problème dont on va parler et qui fait couler tant d'encre (en anglais) : Comment tirer un échantillon sans remise avec des probabilités inégales?

Il y a bien longtemps que nous connaissons ce problème. Pour la banlieue de Paris, les communes ayant été stratifiées suivant un critère sociologique (leur caractère bourgeois ou prolétarien, d'après les élections) par R. Lévy-Bruhl (alors notre adjoint à l'I. N. S. E. E.), nous avions mis dans la même strate des communes d'importance très inégale. Pour ne pas tirer 2 fois la même grande ville (disons Colombes) nous finissions par donner aux petites communes une représentation tout à fait excessive (comme dans un Conseil général).

Que faire? Avouons que la méthode du Japonais Midzuno, consistant à tirer une commune avec probabilités inégales (convenablement modifiées) et la seconde avec probabilités

égales, nous avait fort séduit (encore qu'elle soit inapplicable quand l'inégalité est trop grande). Seulement c'est reculer pour mieux sauter; car si l'on s'arrange pour conserver l'espérance mathématique on change la variance de ce fait, on l'accroît même beaucoup.

Ce qui est étonnant, c'est que personne, jusqu'à présent, n'ait pu traiter tout le problème. Il faudrait : 1° un moyen vraiment pratique de tirer les échantillons de 2 ou 3 communes ou davantage; 2° un estimateur aussi simple que possible, c'est-à-dire un dépouillement des données sans intervention de coefficients de pondération; 3° une variance minimale pour cet estimateur; 4° un calcul commode, sinon du meilleur estimateur sans biais de cette variance, au moins d'un estimateur de variance, par excès mais très peu.

C'est du moins ce qu'il ressort de plus clair de l'article de Des Raj paru en 1966 [19]. Il existe notamment un long article de Hartley et Rao [20] paru en 1962 dans les Annales, avec un appareil mathématique inquiétant, et complété dès 1963 par un article du même Rao (J. N. K.) [21] dans le Journal de l'Association américaine de Statistique, suivi de bien d'autres dont un de Hartley en 1966 : bien qu'une solution du problème (satisfaisante en apparence) ait été donnée par Horvitz et Thompson dès 1952 [22].

Si la mathématique de Hartley-Rao manque de simplicité, c'est que le sondage systématique est substitué au tirage exhaustif pour assurer la non-duplication des unités tirées (soit p_i la probabilité de tirer (i) en n tirages; le sondage systématique convient si tous les p_i sont inférieurs à $1/n$, c'est presque évident).

La méthode est simple, mais le calcul correct de variance est difficile, comme nous l'avons dit plus haut, d'une manière générale, pour tous les sondages systématiques.

Si nous nous attardons, c'est pour dire qu'en fait les problèmes concrets sont ceux où le tirage avec probabilités inégales est le premier degré d'un sondage à 2 degrés et plus. Au lieu du problème restreint à 1 degré, envisageons de l'élargir à 2 degrés : ceci ouvre de nouvelles possibilités. Reprenons par exemple le cas de la ville de Colombes sous-représentée dans l'échantillon.

Il aurait été facile de supprimer le *biais* (l'erreur systématique) en convenant de tirer un plus grand nombre de ménages de cette ville que des autres, lorsqu'elle est tirée au sort. Plus généralement soit $p_1\ p_2\ p_3$ les probabilités de tirer les villes n^{os} 1, 2, 3 ... au 1er tirage; elles sont au 2^{e} tirage ($q_1\ q_2\ q_3\ \ldots$) avec

$$q_1 = p_1 \left(\frac{p_2}{1 - p_2} + \frac{p_3}{1 - p_3} + \ldots \right) \qquad \text{avec } q_1 + q_2 + q_3 \ldots = 1$$

Supposons par exemple $p_1 = 3/6$, $p_2 = p_3 = p_4 = 1/6$; on aura alors $q_1/p_1 = 3/5$; $q_i/p_i = 7/5$ ($i = 2, 3, 4$); $q_1 = 9/30$, $q_i = 7/30$

Les espérances (avec 2 tirages) sont $p_1 + q_1 = 2/5$; $p_i + q_i = 1/5$.

Finalement la représentation de la ville 1 devient correcte si l'on majore de 50 % le nombre de ses ménages échantillon. La première complication réside dans le fait que, pour représenter cette strate par 24 ménages, il faudrait tirer 15 ménages de la ville n° 1, 10 ménages des villes n^{os} 2, 3, 4, le nombre moyen de 24 étant l'espérance mathématique de 25 ou 20 ménages échantillon suivant les résultats du tirage au sort des villes.

Nous voici donc tout à la fois débarrassés du biais et disposant d'un dépouillement sans complication aucune.

Mais ce plan de sondage serait certainement rejeté par l'utilisateur, qui lui reprocherait d'abord de porter sur un nombre aléatoire de ménages, donc d'avoir un coût aléatoire au départ. Il serait rejeté aussi par l'amateur de calculs de variances, qui lui reprocherait la difficulté d'estimation de la variance entre communes : il faudrait comparer la moyenne d'une donnée sur 15 ménages (ville 1) à la moyenne sur 10 ménages (ville 2, 3 ou 4), alors qu'il est si commode de comparer les totaux de 12 ménages dans chaque ville échantillon. Enfin, la variance du sondage aurait une expression *théorique* mal commode, faisant intervenir l'espérance de $1/n$, $n = 15$ ou 10.

La covariance entre les 2 moyennes échantillon a certainement sur la variance un effet favorable, mais qu'il ne sera pas forcément commode de calculer et d'estimer sur échantillon.

Rien ne permet de penser, enfin, qu'on ait obtenu ainsi la variance minimum à attendre d'un échantillon de 2 villes et 24 ménages.

A présent, nous commençons à entrevoir la profondeur des difficultés du sujet, au point de nous demander s'il serait prudent d'attendre d'un bon étudiant auquel on demanderait d'en faire sa thèse de 3e cycle une solution nouvelle du problème (tout en y consacrant en 1966-67 une partie de notre cours).

10. Estimateurs sans biais du type ratio

10 1. Un autre problème a suscité autant de curiosité, mais moins longtemps. Il s'agit ici d'obtenir une estimation sans biais de type *ratio* (alors que l'estimation par ratio habituelle est affectée d'une erreur systématique ou biais). On pourra se reporter à [4] où nous indiquions (1960) une façon d'obtenir un estimateur sans biais en éliminant le biais entre deux estimateurs biaisés ayant la forme de *ratio*.

Tout d'abord il est étrange qu'on s'intéresse tant au *rapport* de deux aléatoires, alors que le *produit* de deux aléatoires présente la même difficulté : par exemple récolte estimée en multipliant la superficie estimée par le rendement (à l'hectare) estimé. Lorsque les deux aléatoires ne sont pas indépendantes, la formule de la variance de leur produit était (semble-t-il) ignorée, sous sa forme exacte donnée par Leo Goodman en décembre 1960 [23]. Le biais est bien connu. La question est ainsi réglée.

La difficulté pour obtenir l'expression exacte de la variance d'un estimateur de type ratio (et pour estimer cette variance sur échantillon) est toujours exceptionnelle.

Mais qu'il s'agisse de probabilités inégales, ou de probabilités égales avec estimateurs se présentant comme un ratio, ou comme un produit, — le plus remarquable (à notre avis) est la place privilégiée qu'occupe (dans l'esprit des spécialistes) l'estimateur sans biais au sens strict, puisqu'on en désire obtenir à tout prix; alors qu'il est si simple d'employer un estimateur biaisé, dont le biais devient négligeable si l'échantillon grandit, — autrement dit d'employer un estimateur asymptotiquement sans biais.

10 2. Ce souci est à rapprocher des travaux exclusivement théoriques des mathématiciens statisticiens de l'Inde : Roy, Chakravarti, Bahadur, Basu, Godambe, Murthy, Joshi, Aggarwal, Pathak [24] qui essaient notamment de reconstruire la théorie des sondages de façon à l'intégrer dans la statistique mathématique générale (Bayesienne notamment [25]).

Considérons par exemple le concept de *suffisance*, qui concerne l'estimation des paramètres d'une distribution de probabilité. Les lois à résumés exhaustifs de Darmois (1935)

possèdent des estimations « suffisantes » de leurs paramètres; et dans le cas régulier, ce sont même les seules lois qui en possèdent. La suffisance (découverte par Fisher et transposée par Savage et l'école néo-bayesienne au cas des estimations bayesiennes) est un concept paramétrique par excellence. Au contraire, la théorie des sondages est non-paramétrique : la population finie formée de N unités x_i dépend finalement des N paramètres x_i sans exception. Cela n'empêche pas Godambe d'une part, Pathak de l'autre, d'étendre la suffisance de Savage aux estimations des sondages [26].

Pour Godambe, l'estimateur doit rendre minimum la perte, espérance mathématique du risque; mais le risque lui-même est quadratique, la perte ressemble fort à la variance, et seule l'estimation sans biais la rend minimum. Finalement on ne veut connaître d'autres estimations que celles sans biais.

Comme les estimations par ratio sont trop commodes et trop connues pour qu'on les oublie, Godambe dit à peu près ceci : si vous croyez savoir que x_i et y_i sont fortement correllés, que vous connaissiez $\Sigma\, y_i$ et vouliez estimer Σx_i, *je vous conseille* de tirer les unités échantillon avec des probabilités proportionnelles aux y_i. Alors $\frac{1}{n}\sum_{(n)} x_i/y_i$ sera estimation sans biais de

$$\sum_{(N)} x_i \Big/ \sum_{(N)} y_i$$

Cette façon de ne pas traiter du sujet ne satisfera pas tout le monde.

a) On objectera tout d'abord qu'on peut connaître Σy_i sans pour autant connaître individuellement les y_i; et il est alors bien clair que le tirage au sort de l'échantillon prescrit est impraticable (faute de base de sondage).

Les Indiens répondront qu'on ne les prend pas au dépourvu et que, depuis 1951 au moins, ils attendent l'objection avec la riposte prête : une technique de Lahiri permettant de tirer les unités au sort proportionnellement aux y_i sans connaître les y_i.

b) Je me suis reporté au papier de Lahiri [27], un statisticien très modeste et sympathique, depuis longtemps collaborateur du professeur Mahalanobis. Il est effectivement possible de procéder comme suit :

1) tirer un grand échantillon d'unités avec d'égales probabilités;

2) aller sur le terrain mesurer la variable y_i de chacune de ces unités;

3) armé de nombres aléatoires, décider si l'on conserve ou rejette chaque unité tirée, par comparaison entre son y_i et un certain nombre aléatoire;

4) quand on conserve l'unité (i), procéder à la mesure des x_i sur le terrain.

Le procédé est donc *coûteux* puisqu'il suppose y_i mesuré sur des unités de sondage beaucoup plus nombreuses que celles qu'on retiendra finalement. De plus, nous sommes des plus sceptiques en ce qui concerne la comparaison sur le terrain entre un nombre aléatoire et y_i; nous pensons que l'enquêteur français est en général trop intelligent pour ne pas profiter de cette occasion de « diriger le hasard » dans tel ou tel sens qui (à tort ou à raison) lui paraît s'imposer (par insuffisance ou excès de zèle).

c) D'ailleurs, c'est plutôt en confrontant, pas à pas, x_i à y_i qu'on juge si la corrélation est ou n'est pas notable, et par suite si l'estimation par ratio — voire par régression — est préférable à l'estimation sans biais n'utilisant pas les (y_i). A notre avis, c'est le fait de pouvoir

admettre l'existence d'une régression linéaire forte de Bernstein de x en y qui justifie l'estimation par ratio, ou par régression.

Ceci ne signifie pas que la régression de y en x soit supposée elle aussi linéaire — ce qui impliquerait (on le sait) pour les (x_i, y_i) une distribution gaussienne, hypothèse beaucoup trop forte et non réaliste.

10 3. Nous préférerions (au lieu de ces propos désabusés) apporter notre propre théorie de l'estimation par ratio — ou d'un ratio. Orienté par M. Georges Darmois vers la notion d'information (voici une dizaine d'années), nous avons constaté [2] que la variance est une perte d'information (au sens général que Schützenberger donne à l'information) et que c'est l'information perdue quand on estime *sans biais* la *moyenne* de population par la moyenne échantillon (le tirage au sort de l'échantillon ayant lieu dans une urne, avec ou sans remise).

Pour un estimateur biaisé, la variance n'est plus perte d'information, et l'expression (Variance + Carré du Biais) inspirée par l'inégalité de Fréchet-Cramer n'est pas, elle non plus, une perte d'information. Tout au moins s'agit-il d'échantillons finis; car avec des échantillons très grands, on en arrive au cas où $(\bar{x}, \bar{y})$ suit une loi-limite de Gauss, quelle que soit la distribution (x, y).

Alors l'estimation par régression est la meilleure des estimations par ratio (engendrées en changeant l'origine), mais les biais sont négligeables.

Il est bien facile de construire des pertes d'information; en revanche nous ne savons pas en général leur donner une interprétation simple comme c'est le cas pour la variance. Et nous ne savons pas plus trouver une perte scalaire associée à un estimateur simple, comme l'estimateur ratio. L'expression approchée courante du cœfficient de variation du ratio

$$\gamma^2 - 2\rho\ \gamma\gamma^1 + \gamma^{12}$$

(γ, γ^1 étant les coefficients de variation de 2 termes du ratio, ρ étant leur corrélation) n'est qu'approximativement une perte d'information. Il en est de même, pour un produit, de l'expression : $\gamma^2 + 2\,\rho\,\gamma\gamma^1 + \gamma^{12}$

Durbin [28] a donné (1960) le concept de l'estimation par une équation sans biais (qui concerne le ratio $\Sigma(n)\ x_i/\Sigma(n)y_i$) se substituant à l'estimation sans biais.

Nous n'avons pas réussi à tirer de là des pertes d'information scalaires intéressantes [29].

Comme les chercheurs désirent comparer la précision de plusieurs estimateurs de la même grandeur ($\Sigma_{(N)}\ x_i/N$) dont au moins une estimation sans biais est connue ($\Sigma(n)x_i/n$), il est bien naturel qu'ils s'acharnent à calculer les variances (ou, comme succédané, variance + Biais carré) pour les comparer à la variance de l'estimateur sans biais; l'estimateur est jugé meilleur s'il y a baisse de variance, même si celle-ci n'a pas les caractères d'un gain d'information.

11. Les polykays

Dans un article de Robson [30] déjà un peu « vieux » (1961) relatif à un estimateur du type *ratio*, on emploie la théorie des *polykays* pour venir à bout du calcul de variance qui est très difficile si l'on ne se contente pas d'une expression approchée. Ceci nous a conduit

à faire une exploration dans l'énorme littérature relative aux *polykays*, pour notre cours (de 3e cycle) de 1965-1966. Il s'agit au départ d'une théorie d'algèbre, concernant certains invariants algébriques, qui remonte au XIXe siècle. Mais c'est Fisher (sir Ronald) qui en 1928 [31] a commencé à utiliser ces invariants pour faciliter les calculs relatifs à l'estimation des variances de variances. Auparavant, divers auteurs, notamment Tchouproff, s'étaient lancés dans des calculs inextricables pour traiter directement ce dernier problème; les *polykays* constituent une voie indirecte pour parvenir sans trop de mal au résultat. Au début, il s'agissait des *cumulants*, et les échantillons étaient tirés avec remise. Le passage aux tirages sans remise s'est fait avec Tukey (1950) [32] et Wishart (1952) [33].

Le sujet n'est pas épuisé, puisqu'on trouve en 1966 un article de Dayhoff [34] sur le polykays; et l'exposé d'ensemble le plus lisible est encore celui d'Esther Schaeffer et Dwyer, paru en 1963 dans le Journal de l'Association américaine de Statistique [35].

Si cette théorie témoigne au départ de l'intérêt qu'on portait jadis pour la précision des calculs de variance (1) il est remarquable qu'on n'en a fait presque aucun usage de cette sorte. Le premier article de Tukey [36] de 1956 s'arrête aux sondages stratifiés; les suivants concernent l'analyse de variance; et les chercheurs comme Dayhoff sont orientés vers les plans d'expérience et non vers les sondages.

Nous avions signalé la question à notre ami V. Fonsagrive; mais notre ami nous a quittés pour toujours sans avoir terminé son article sur les *polykays*.

12 Mode-médiane — Statistiques d'ordre

a) L'estimation du *mode* sur échantillon a été abordée (1966) par Dalenius [37] le spécialiste suédois des sondages. L'estimation de la *médiane* d'un échantillon provenant de 2 strates a été abordée par Mac Carthy (1965) [38]. Ce sont des sujets neufs. Je doute qu'on soit bien inspiré de les aborder avec le calcul du biais et de la variance.

b) Il y a environ dix ans, j'ai trouvé [2] que lorsqu'on estime la *médiane* (d'une seule strate) par la médiane échantillon, la perte d'information était un écart absolu moyen convenablement pondéré. Je n'ai pas su trouver la perte associée à l'estimation *du mode*. Bien entendu (sauf pour une distribution symétrique) l'estimation de la médiane par la médiane est biaisée, sa variance n'est pas perte d'information.

Plus généralement (nous le signalons à notre jeune camarade Dumas de Railly) les statistiques d'ordre sont des estimations biaisées des quantiles correspondants de la distribution sondée. Le biais a d'ailleurs une expression connue, pour une distribution continue de loi connue (voir en annexe).

c) A ce propos, disons quelques mots d'un problème de sondage apparemment inédit que nous a apporté l'un de mes étudiants de 3e cycle de Paris, M. Lerman; il l'a rencontré en faisant sa thèse, mais ne l'a pas traité.

Il s'agissait d'estimer la moyenne $\bar{\bar{\chi}}$ des éléments supérieurs à la médiane et plus généralement à un quantile de la distribution sondée. Par exemple, si l'on savait la vraie valeur de la médiane, μ, il n'y aurait aucun problème : la moyenne $\bar{\bar{x}}$ des éléments échantillons

1. Variance; Variance de la variance; variance de la variance de variance..., il n'était plus question de s'arrêter sur cette pente; il fallait estimer ces paramètres et les estimations devenaient de plus en plus mauvaises. Tout cela ne menait à rien en pratique.

plus grands que μ estime sans biais $\bar{\bar{\chi}}$. Mais en pratique μ est inconnue et on commence par l'estimer par la médiane-échantillon m. Alors $\bar{\bar{x}}$ est biaisée; et ce biais est défini par la loi de la distribution sondée; loi généralement inconnue. Il ne semble pas possible de traiter un tel problème sans faire sur cette distribution des hypothèses de structure (1); il conviendrait bien entendu d'en faire le moins possible.

Le problème peut être présenté d'une autre façon; il s'agirait d'estimer sur échantillon la courbe de concentration de Lorenz d'une distribution. Si le total général (ou la moyenne générale) était exactement connu(e) les divers points de cette courbe seraient estimés sans biais ni en abcisses ni en ordonnées.

Comme la moyenne générale est (en fait) estimée sur échantillon les ordonnées sont estimées par des ratios, et par suite sont biaisées. Par ailleurs on coupe cette courbe par des droites d'abcisse donnée (abcisse 1/2 par exemple pour la médiane) et on constate que l'ordonnée est estimée avec un biais, lequel n'est pas aussi petit qu'on le souhaiterait.)

Peut-être conviendrait-il encore, dans ce type de problème, d'accepter une estimation biaisée, mais de la choisir de façon à rendre minimum quelque écart (convenablement défini) entre la grandeur à estimer et son estimation (comme en *b* ci-dessus)?

d) Montrons quelle peut être l'ampleur du biais (d'après COX, 1964) [39]. Soit une distribution exponentielle $F(x) = 1 - e^{-x}$. Sa médiane est $\mu = \text{Log } 2$ et les vraies moyennes sont :

$1 - \text{Log } 2$ si x est inférieur à μ (x positif)

$1 + \text{Log } 2$ si x est supérieur à μ

L'estimation de $(1 - \text{Log } 2)$ au moyen des $\frac{n}{2}$ plus petites parmi n valeurs échantillons de x, a pour espérance mathématique :

$$E = \left[\frac{1}{n} + \left(\frac{1}{n} + \frac{1}{n-1}\right) + \left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2}\right) + \ldots + \left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{\frac{n}{2}+1}\right)\right]\frac{2}{n}$$

Quand on tend vers l'infini, cette expression E a pour limite $1 - \text{Log } 2$.

$$\text{Log } 2 = 0{,}693;\ 1 - \text{Log } 2 = 0{,}307$$

Lorsque n n'est pas très grand, le calcul direct donne

$n =$	2	4	6	8	10	20	30	40	50	100
E =	0,500	0,417	0,383	0,365	0,354	0,331	0,323	0,319	0,317	0,312

0,312 est encore éloigné de 0,307 de plus de 1 pour cent.

On est donc en présence d'un phénomène assez rare en sondage : le biais est très lent à s'éliminer.

Qu'en est-il pour d'autres distribution? Nous l'ignorons.

D'après un résultat de Ali et Chang (1965) [40] pour une distribution symétrique unimodale, le biais serait par excès; pour une distribution en U, ce serait l'inverse (*cf.* annexe ci-après).

1. A la réflexion, on peut trouver quelque procédé non paramétrique, un peu grossier, au moins dans des cas simples; on ne va pas très loin (1967, janvier).

13. Méthodes non paramétriques ou paramétriques?

Il semble bien qu'actuellement, il soit difficile de progresser autrement que nous le faisons. Les sondages meurent d'inanition s'ils sont soumis à un régime trop strictement non paramétrique. On notera d'ailleurs qu'il a toujours fallu supposer que les distributions sondées n'étaient pas totalement inconnues; dès 1934, Neyman suppose connus les écarts types de chaque strate, pour pouvoir répartir l'échantillon entre les strates de façon à minimer la variance d'ensemble.

Prenons encore l'exemple de la théorie de la *stratification optimale*, due à Dalenius, sur laquelle nous avons travaillé en 1954 ainsi que MM. Desabie et Chartier [4]. On suppose la loi de distribution de x connue; on traite le problème du découpage optimal de cette distribution; on s'aperçoit qu'en définitive on n'a pas besoin de renseignement sur les queues de distribution (toujours mal connues) mais sur la distribution au voisinage des points où se font les coupures.

Pour découper la population en 2 strates, quelques données statistiques suffisent pour appliquer une méthode pratique. Pour découper en 10 ou 25 strates, il faudrait connaître très bien la distribution; mais cela ne servirait plus à rien, car alors la variance n'est plus guère sensible au choix de telle ou telle frontière entre strates.

14. La stratification optimale

L'application des équations de Dalenius semble avoir posé des problèmes; elles ne sont simples qu'en apparence; on ne les peut résoudre que par itérations, et il n'y a aucune raison qu'elles admettent une solution unique (si la distribution sondée est un peu bizarre). Les recherches d'Ekman en Suède à ce sujet ont abouti (1959-1963) à trois articles, dont un nous a paru particulièrement difficile [42].

Notre attention a été davantage retenue par l'article de Ghosh (1963) [43] qui désire obtenir la stratification optimale par rapport à 2 variables (x, y) connaissant la loi de distribution des (x, y).

Cet essai n'est guère heureux par le choix qui est fait du critère à minimer : une certaine variance généralisée (qu'on attribue à Wilks et qu'on retrouve dans d'autres travaux [44] déterminant de la matrice des variances et covariances d'échantillonnage. Alors que cette matrice est une vraie perte d'information (ou encore sa forme quadratique associée), le déterminant ne jouit pas de cette qualité. Qu'arriverait-il d'ailleurs si ce déterminant était nul? Faudrait-il en conclure que la stratification est optimale. Pas du tout; la corrélation entre x et y serait égale à 1, il conviendrait de rendre minimum $\sqrt{\bar{x}}$ seul.

L'interprétation économique du critère à minimiser est aussi convaincante. Faute de pouvoir minimer simultanément $\sqrt{\bar{x}}$ et $\sqrt{\bar{y}}$, on songera à minimer $a\sqrt{\bar{x}} + c\sqrt{\bar{y}}$; le choix de a/c dépend des utilités comparées d'une bonne connaissance de $\bar{x}$ et de $\bar{y}$. Minimer $a\sqrt{\bar{x}} + b\mathrm{Cov}\,\bar{x}\,\bar{y} + c\sqrt{\bar{y}}$ signifie qu'en outre la connaissance de la corrélation existant entre x et y présente pour nous une utilité comparable aux précédentes.

En résumé, Ghosh a commis ici ce qu'on appelle une « erreur de 3e espèce », il a traité brillamment un problème qui n'était pas le bon.

15. La répartition optimale de l'échantillon

Le problème de *la répartition* (*allocation*) *optimale* de l'échantillon dans les strates, résolu (1934) par Neyman avec une seule variable privilégiée, a été traité par divers auteurs pour plusieurs variables, soit qu'on minime une forme quadratique comme ci-dessus, soit qu'on impose aux variances des bornes fixées d'avance et qu'on minime le coût de sondage.

Il s'agit ainsi de programmation non linéaire, *dépourvue* de programmes tout faits [45].

Par exemple, nous voyons Jagannathan [46] s'orienter vers les problèmes de programmation (1965) ce qui n'est pas pour nous surprendre (voir [4], 1960).

En revanche, Folks et Antle [47] font un exposé théorique plus raffiné que leurs prédécesseurs en s'inspirant du livre de Karlin (tome 1) *Mathematical methods and theory, in games, programming and economics* (1).

16. Théories nouvelles

Nous voyons présentement naître ou renaître des théories, encore embryonnaires, dont l'intérêt pratique pourrait devenir considérable.

Sondages à buts multiples. Il s'agit par exemple d'obtenir un échantillon de ménages ruraux qui (en majeure partie) puisse servir aussi bien à une enquête agricole qu'à une enquête démographique.

Sondages analytiques [48]. Il s'agit de choisir un plan de sondage qui mette en relief les contrastes entre des sous-populations, et non plus (comme dans la théorie classique) un plan destiné à informer le mieux possible sur une population sondée.

17. Le sondage a plusieurs phases

Le sondage à plusieurs phases, dont la première théorie (comportant une stratification entre les deux phases) remonte à 1938 et encore est due à Neyman, semble reprendre de nos jours une place de premier plan; car c'est lui qu'on utilise :

a) dans les sondages analytiques de Sedransk [49];

b) dans la théorie du « *post cluster sampling* » (c'est-à-dire le sondage par grappes constituées entre la première et la seconde phase) méthode réellement employée dans certains pays (Suède notamment) [50].

18. Théories a deux faces

Nous dirons un mot enfin d'une catégorie assez particulière de travaux qui ne sont pas le fait des spécialistes du sondage, qui peuvent être très mathématiques, et qui sont en réalité étroitement liés aux problèmes de sondage.

1. Un de nos collaborateurs, M. Raffin, poursuit des recherches chroniques sur la programmation convexe.

a) Le groupage [51]

Certains auteurs étudient le « *grouping* » c'est-à-dire la constitution de classes, sous-classes, etc., dans une population, tendant à réduire une fonction objective; si celle-ci est quadratique et analogue à une variance, ce problème n'est autre qu'une stratification optimale, le nombre de strates restant à choisir.

La différence essentielle entre les deux théories est que le « grouping » n'étudie qu'une population comprenant un millier d'éléments, et qu'il la stratifie à l'aide d'un calculateur électronique auquel il a fallu donner un programme assez court. La technique de Dalenius n'a bien entendu rien à voir avec celle-ci.

b) *Le « double sondage » de Stein*

Depuis 1945, époque à laquelle la terminologie des sondages n'était pas encore normalisée par l'O. N. U., Stein a appelé sondage à 2 degrés ce qu'on appelle sondage à 2 phases (en sondages) ou double sondage (en contrôle de qualité).

Le problème de Stein consiste à déterminer une « procédure » optimale pour estimer la moyenne d'une population gaussienne avec une précision donnée. Ce problème a intéressé un nombre croissant de chercheurs et s'est beaucoup étendu.

Nous nous contenterons de signaler l'article de Goldman et Zeigler des *A. M. S.* (août 1966) à partir duquel les personnes intéressées pourront remonter la filière [52].

Il est étrange que les techniciens des sondages se désintéressent du problème de Stein, même s'ils rencontrent fort rarement des distributions gaussiennes.

c) *Nombre de succès dans* n *épreuves indépendantes*

Ce titre désigna des articles de Wassily Hoeffding (1956), Darroch (1964) et Samuels (1965) [53]. Il s'agit en d'autres termes du schéma connu des urnes de Poisson, c'est-à-dire du sondage stratifié avec variable ne prenant que les valeurs 1 ou 0; on ne tire qu'une boule de chaque urne.

Il existe par ailleurs un mathématicien tchèque, Jaroslav Hajek, qui a fort étudié ce même schéma sous le nom de sondage de Poisson puis de sondage avec rejet (*rejective sampling*) [54]. Les travaux de ces auteurs ont des orientations distinctes.

Les premiers ont par exemple étudié la forme de l'histogramme des probabilités du nombre B de boules blanches tirées.

Hajek au contraire utilise ce modèle pour tirer au sort un échantillon avec d'inégales probabilités et sans remise (voir [19 à 22] ci-dessus). A chaque élément de sa population il affecte une urne, et il imagine N urnes de Poisson distinctes; si le tirage de l'urne (i) donne une boule blanche, l'unité (i) fait partie de l'échantillon.

Il est clair que si l'on a $p_i = 1/N$, le sondage de Poisson-Hajek équivaut au tirage exhaustif de $n = B$ unités sur N; seulement B n'est pas connu d'avance, c'est une variable aléatoire ayant une distribution Binomiale entre 0 et N.

La variance du sondage de Poisson pour B donné est celle du sondage exhaustif; la variance quel que soit B est l'espérance mathématique de la précédente; pour avoir son expression on a besoin du moment d'ordre (— 1) de la loi binomiale. Nous en avons indiqué une méthode de calcul (1963) [55].

A présent on va supposer que les p_i ne sont pas tous égaux entre eux. Soit alors n le nombre de boules blanches qu'*on veut tirer* (nombre arrêté à l'avance). On convient de rejeter l'échantillon et de recommencer tant que B diffère de n. C'est le *sondage avec rejet.* Il lui correspond une technique de calcul. On a résolu le problème [19-22] par une voie extrêmement détournée et intelligente, bien qu'en fait peu praticable (à tous points de vue).

Une préoccupation de Hajek est de savoir si l'on a bien une loi de Laplace-Gauss limite. C'est un problème théorique général : il s'agit de moyennes de variables non indépendantes. C'est aussi d'un grand intérêt pratique.

Un autre problème général serait de construire des tests d'hypothèse utilisables pour les sondages : ce ne doit pas être commode puisque presque personne ne s'y attaque.

19. Conclusions

Parvenu au terme de cet exposé — qui est bien loin d'être exhaustif — mais qui est épuisant pour l'auditoire, nous pensons au bon mot de l'académicien Pierre Gaxotte : « ce mal qui répand la terreur, la conférence ».

Nous avons vu certains théoriciens à la recherche d'une théorie unifiée des sondages. Nous avons aussi eu bien l'impression qu'elle n'était pas du tout en voie d'unification par ailleurs.

Nous avons vu les pionniers partir à la découverte des terres vierges; et d'autres y parvenir par des voies fort détournées.

Nous avons remis en doute les dogmes non paramétriques — sans rien savoir mettre à la place des idoles jetées bas.

Qu'on aime la théorie — ou plutôt les problèmes pratiques; qu'on aime les calculs très compliqués ou qu'on les déteste — il semble qu'on trouve ici encore bien du travail à faire, difficile sans aucun doute, mais capable de passionner.

Espérons maintenant qu'on ne se hâtera pas, en haut lieu, de planifier nos recherches et d'abord qu'on ne nous interdira pas de les faire, à nous-mêmes et aux trop rares « disciples » (si l'on peut dire) qui choisissent (librement) de s'y engager. Espérons qu'on tolérera ainsi longtemps que chaque membre de l'Enseignement supérieur s'occupe de ce qui lui fait plaisir — et aussi, bien sûr, de ce qui intéresse ses propres chercheurs — ses assistants et maîtres assistants (qui, en mathématiques ne l'assistent généralement pas en matière de recherches, mais seulement en matière d'enseignement).

Il n'y a pas toujours concordance; et pour paraphraser un mot qu'on prêtait jadis au directeur de l'E. N. S. (Célestin Bouglé) dans la Revue de l'École 1939 : « Je suis leur directeur, il faut bien que je les suive », « Eh bien ! je suis leur directeur de thèse, il faut bien que je m'intéresse à leurs recherches, même si elles sont assez éloignées des miennes. » Or leurs recherches ne s'orientent guère vers les sondages probabilistes qu'ils ignorent et qu'en définitive, en France, malheureusement on ne connaît en dehors de l'I. N. S. E. E. que par ouï-dire.

P. Thionet.

ANNEXE

Statistique d'ordre

Soit $x\ (t)$ la statistique d'ordre, par laquelle on estime le quantile $\theta\ (t)$ d'une distribution (sondage bernoullien, échantillon de taille n).

Soit $p = t\ /\ (n + 1)$, $q = 1 - p$, $f\ (y)$ la densité de probabilité de la distribution.

Le biais est donné par la formule :

$$\mathscr{E}\, x\ (t) - \theta\ (t) = - pq\, f'\ [\theta\ (t)]\ /\ 2\ (n + 2)\ f^2\ [\theta\ (t)] + O\ (n^{-2})$$

qu'on trouve dans Walsh (John) *A. M. S.* 29 (June 1958), pp. 601-604, qui l'a lui-même prise dans David et Johnson :

David (F. N.) & Johnson (N. L.) : *Statistical treatment of censored data.* Part. I, Biometrika 41 (1954), 228-240.

On voit donc qu'en première approximation le biais est positif si f' est négatif : pour une distribution unimodale, à droite du mode, — pour une distribution en U, à gauche du mode (convexité).

Ali et Chang établissent un résultat peu différent; le biais est positif à droite de la *médiane échantillon* (distribution unimodale) ou à gauche (distribution en U) (1965), (distributions *symétriques*).

Passons de là au problème de Lerman, qui consiste à estimer l'espérance mathématique de la variable en dessous de $\theta\ (t_0)$ par la moyenne des statistiques d'ordre inférieures à $x\ (t_0)$. Le biais d'une telle estimation est peu différent de la moyenne des biais ci-dessus. Du moins est-ce vrai si la moyenne porte sur assez de données pour qu'on puisse valablement approcher une aire par la méthode des trapèzes, confondant

$$\sum_{t=i}^{t_0} \theta\ (t)/t_0 \qquad \text{avec :} \int y\ d\ \mathrm{F}\ (y)\ (\text{sur } y < \theta\ (t_0))$$

Dans le cas contraire, une telle approximation entraîne des erreurs, de sens prévisible (de chaque côté d'un point d'inflexion). Lorsque les deux erreurs sont de même sens (ce qui est fréquent), on peut prédire le signe du Biais (ce n'est pas d'ailleurs le cas pour la distribution exponentielle).

BIBLIOGRAPHIE

[1] Thionet (P.). — Méthodes statistiques modernes des administrations fédérales aux États-Unis. Hermann, 1946.

[2] Thionet (P.). — La perte d'information par sondage. *P. I. S. U. P.*, 1958.

[3] Thionet (P.). — Développements récents de la théorie des sondages. *J. S. S.P.*, octobre-décembre 1959, pp. 279-295.

[4] Thionet (P.). — Quelques aspects de la théorie des sondages. *J. S. S. P.*, avril-juin 1960, pp. 99-111.

[5] Dalenius (Tore). — Recent advances in sample survey theory and methods. *A. M. S.*, 33-2, June 1962, pp. 325-349.

[6] Murthy (M. N.). — Some recent advances in sampling theory. *J. A. S. A.*, 58, september 1963, pp. 737-755.

[7] Godambe (V. P.). — A review of the contributions toward a unified theory of sampling from finite population. *R. I. I. S.*, 33/2, 1965, pp. 242-258.

[8] THIONET (P.). — Une façon d'exploiter un mauvais échantillon. *B. I. I. S.*, Belgrade, 1965.

[9] SUDMAN (Seymour). — Probability sampling with quotas. *J. A. S. A.*, 61, septembre 1966, pp. 749-771.

[10] CHAHINE (Jacques). — Une généralisation de la loi binomiale négative. *R. S. A.*, 13-4, 1965, pp. 33-43.

[11] THIONET (P.). — Sur l'estimation de variance dans le cas d'échantillons systématiques. *R. S. A.*, 13-4, 1965, pp. 51-60.

[12] THIONET (P.). — Le sondage systématique au sens large (à paraître dans la *R. S. A.*).

[13] TORNQVIST (L.). — The theory of replicated systematic cluster sampling with random start. *R. I. I. S.*, 31-1, 1963, pp. 11-23.

[14] ZINGER (A.). — Estimations de variances avec échantillonnage systématique. *R. S. A.*, 11-2, 1963, pp. 89-97.

[15] THIONET (P.). — Quelques points se rapportant aux calculs de variance des sondages. *R. S. A.*, 13-4, 1965, pp. 45-50.

[16] W. EDWARDS DEMING. — Sample design in business research. 1960, 630 pages (J. Wiley).

[17] TATE (R. F.), KLETT (G. W.). — Optimal confidence intervals for the variance of a normal distribution. *J. A. S. A.*, September 1959, pp. 674-682.

[18] THIONET (P.). — La théorie de l'estimation et les sondages. *Estadistica*, 65, décembre 1959, pp. 702-715.

[19] DES RAJ. — Some remarks on a simple procedure of sampling without replacement. *J. A. S. A.*, June 1966, pp. 391-396.

[20] HARTLEY (H. O.), RAO (J. N. K.). — Sampling with unequal probabilities and without replacement. *A. M. S.*, 33, June 1962, pp. 350-374.

RAO (J. N. K.), HARTLEY (H. O.), COCHRAN (W. G.). — On a simple procedure of unequal probability sampling without replacement. *J. R. S. S.*, B 24-2, 1962, pp. 482-491.

[21] RAO (J. N. K.). — On the procedures of unequal probability sampling without replacement. *J. A. S. A.*, 58, March 1963, pp. 202-215.

HARTLEY (H. O.). — Systematic sampling with unequal probability and without replacement. *J. A. S. A.*, September 1966, pp. 739-748.

[22] HORVITZ (D. G.), THOMPSON (D. J.). — A generalization of sampling without replacement from a finite universe. *J. A. S. A.*, 47, December 1952, pp. 663-685.

[23] GOODMAN (Leo A.). — On the exact variance of products. *J. A. S. A.*, 55, December 1960, pp. 708-713.

[24] ROY (J.), CHAKRAVARTI (I. M.). — Estimating the mean of a finite population. *A. M. S.*, 31, June 1960, pp. 392-398.

MURTHY (M. N.), etc. — Ordered and unordered estimators in sampling without remplacement. *Sankhya*, 18-3/4, September 1957, pp. 379-390.

[25] GODAMBE (V. P.), JOSHI (V. M.). — Admissibility and Bayes estimation in sampling finite populations I, II, III. *A. M. S.*, 36, December 1965, pp. 1707-1742.

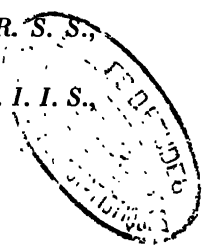
AGGARWAL (O. M. P.). — Bayes and minimax procedures for estimating the arithmetic mean of a population with two stage sampling. *A. M. S.*, 37, October 1966, pp. 1186-1195.

AGGARWAL (O. M. P.). — Bayes and minimax procedures in sampling from finite and infinite populations. *A. M. S.*, 30, March 1959, pp. 206-218.

[26] PATHAK (P. K.). — Sufficiency in sampling theory. *A. M. S.*, 35, June 1964, pp. 795-808.

GODAMBE (V. P.). — A new approach to sampling from finite population I-II. *J. R. S. S.*, B 28-2, 1966, pp. 310-328.

[27] LAHIRI (D. B.). — A method of sample selection providing unbiased ratio estimates. *B. I. I. S.*, 33/2, 1951, pp. 133-140.



[28] Durbin (J.). — Estimation of parameters in time-series regression mode. *J. R. S. S.*, B 22-1, 1960, pp. 139-153.

[29] Thionet (P.). — Sur une extension de l'estimation sans biais (à paraître dans la *R. S. A.*).

[30] Robson (D. S.), Vithayasai (C.). — Unbiased componentwise ratio estimations. *J. A. S. A.*, June 1961, pp. 350-358.

[31] Fisher (R. A.). — Moments and product moments of sampling distributions. *Proc. London Math. Soc.* 30, 1928, pp. 199-238.

[32] Tukey (J. W.). — Some sampling simplified. *J. A. S. A.*, 45, 1950, pp. 501-519.

[33] Wishart (J.). — Moment coefficients of the k-statistics in sample from a finite population. *Biometrika*, 39, 1952, pp. 1-13.

[34] Dayhoff (E.). — Generalized polykays, an extension of simple polykays and bipolykays. *A. M. S.*, 37, February 1966, pp. 226-241.

[35] Schaeffer (Esther), Dwyer (Paul S.). — Computation with multiple k-statistics. *J. A. S. A.*, 58, March 1963, pp. 120-151.

[36] Tukey (J. W.). — Keeping moment-like sampling computations simple. *A. M. S.*, 27, March 1956, pp. 37-54.

Tukey (J. W.). — Variances of variance components I. *A. M. S.*, 1956, pp. 722-736 (II, III, etc.).

[37] Dalenius (Tore). — The mode, a neglected statistical parameter. *J. R. S. S.*, A, 128-1, 1965, pp. 110-117.

[38] Mc Carthy (P. J.). — Stratified sampling and Distribution-free confidence intervals for a median. *J. A. S. A.*, 60, September 1965, pp. 772-783.

[39] Cox (D. R.). — Some application of exponential ordered scores. *J. R. S. S.*, B 26-I, 1964, pp. 103-110.

[40] Ali (M. M.), Chang (L. K.). — Some bounds for expected values of ordered statistics, *A. M. S.*, 36, June 1965, pp. 1055-1057.

[41] Thionet (P.). — Comment choisir un échantillon dans une population où les sujets sont d'importances très différentes, pp. 29-37.

Desabie (J.). — Sur un problème d'échantillon optimum pp. 37-43.

Rééditions dans Revue française de Marketing, 3e trimestre 1965.

Voir aussi : *J. S. S. P.*, juillet-septembre 1955, pp. 192-206, une version plus complète.

[42] Ekman (G.). — An approximation useful in univariate stratification. *A. M. S.*, 30, March 1959, pp. 219-229.

Ekman (G.). — Approximate expressions for the conditional mean and variance over small intervals of a continuous distribution. *A. M. S.*, 30, December 1959, pp. 1131-1134.

Ekman (G.). — On the sum $\Sigma p_h^i \sigma_h^i$. *R. I. I. S.*, 31-1, 1963, pp. 67-80.

[43] Ghosh (S. P.). — Optimum stratification with two characters. *A. M. S.*, 34-3, 1963, pp. 866-872.

[44] Bagai (O. P.). — The distribution of the generalized variance. *A. M. S.*, 36, February 1965, pp. 120-130.

Wilks (S. S.). — Certain generalizations in the analysis of variance. *Biometrika*, 26, 1932, pp. 471-94.

[45] Kokan (A. R.). — Optimum allocation in multivariate surveys. *J. R. S. S.*, A 126/4, 1963, pp. 557-565.

[46] Jagannathan (R.). — A method for solving a nonlinear programming problem in sample surveys. *Econometrica*, 33, October 1965, pp. 841-846.

[47] Folks (J. L.), Antle (C. E.). — Optimum allocation of sampling units to strata, when there are R responses of interest. *J. A. S. A.*, 60, March 1965, pp. 225-233.

[48] Yates (F.). — Sampling methods for censuses and surveys. 3e édition, Londres, 1960.

[49] SEDRANSK (J.). — A double-sampling scheme for analytical surveys. *J. A. S. A.*, 60, December 1965, pp. 985-1004.

[50] GHOSH (S. P.). — Post cluster sampling. *A. M. S.*, 34-2, June 1963, pp. 587-597.

[51] FISHER (W. D.). — On a pooling problem from the statistical decision view point. *Econometrica*, 21, 1953, pp. 567-585.

FISHER (W. D.). — On grouping for maximum homogeneity. *J. A. S. A.*, 53, December 1958, pp. 789-798.

COX (D. R.). — Note on grouping. *J. A. S. A.*, 52, December 1957, pp. 453-547.

WARD (J. H.). — Hierarchical grouping to optimize an objective function. *J. A. S. A.*, 58, March 1963, pp. 236-244.

LAZAR (Philippe). — Partition d'un groupe hétérogène en sous groupes homogènes. *R. A. S.* 14-1, 1966, pp. 39-43.

[52] GOLDMAN (A. S.), ZEIGLER (R. K.). — Comparisons of some two-stage sampling methods. *A. M. S.*, 37, August 1966, pp. 891-897.

[53] HOEFFDING (W.). — On the distribution of the number of successes in independent trials. *A. M. S.*, 27, 1956, pp. 713-721.

DARROCH (J. N.). — On the distribution of the number of successes in independent trials. *A. M. S.*, 35, September 1964, pp. 1317-1321.

SAMUELS (S. M.). — On the number of successes in independent trials. *A. M. S.*, 36, August 1965, pp. 1272-1278.

[54] HAJEK (J.). — Asymptotic theory of rejective sampling with varying probabilities from a finite population. *A. M. S.*, 35, December 1964, pp. 1491-1523.

[55] THIONET (P.). — Sur le moment d'ordre (-1) de la distribution binomiale tronquée. *P. I. S. U. P.* 12-3, 1963, pp. 93-102.

ABRÉVIATIONS BIBLIOGRAPHIQUES EMPLOYÉES

A. M. S. : Annals of Mathematical Statistics.
J. A. S. A. : Journal of the American Statistical Association.
J. R. S. S. : Journal of the Royal Statistical Society (Londres).
B. I. I. S. : Bulletin de l'Institut international de Statistique.
R. I. I. S. : Revue de l'Institut international de Statistique (La Haye).
P. I. S. U. P. : Publications de l'Institut de Statistique de l'Université de Paris.
R. S. A. : Revue de Statistique appliquée (Paris).
J. S. S. P. : Journal de la Société de Statistique de Paris.

DISCUSSION

M. le président GIBRAT remercie M. Thionet de sa communication qu'il juge difficile; puis il donne la parole à l'auditoire.

M. PRÉVOT Jean, chef du bureau central de Statistique industrielle, serait désireux de s'enquérir sur les sondages par quotas faits par M. Chahine auprès des industriels, pour des études de marché.

M. THIONET répond que M. Chahine est sans doute (comme saint Paul) de ces gens qui sentent deux hommes en eux. D'une part son occupation professionnelle lui donne

l'occasion de se poser des problèmes mathématiques et d'en résoudre certains, fort proprement. D'autre part sa profession de statisticien demande de lui une activité au fond très éloignée des mathématiques; d'après certains travaux dont nous avons eu connaissance, ce jeune libanais semble s'y être parfaitement adapté. Nous n'avons pas dit que les sondages par quotas étaient ou n'étaient pas mauvais, nous avons dit que tout calcul relevant des Probabilités était d'une entreprise délicate en ce qui les concerne : puisque ces calculs envisagent ce qui se produirait si l'échantillon avait une distribution de probabilité, alors qu'il n'en a en fait aucun. On nous a fait faire jadis, à l'I. N. S. E. E., des raisonnements du même type : par exemple si les points de vente de la viande où les prix sont observés étaient tirés au sort, un calcul conduit (pour les fluctuations de l'indice des prix) à certaines conclusions. Or nous savons très bien que ces points de vente ne sont pas et ne peuvent être tirés au sort. D'où il suit que les résultats du calcul dont nous parlons doivent être utilisés avec prudence, sans aller pourtant jusqu'à dire qu'ils seraient fallacieux (voir le *Bulletin de Statistique*, janvier-mars 1953, p. 81).

M. le président GIBRAT dit qu'il aurait souhaité avoir des précisions sur les *polykays*. M. Thionet (tout en lui remettant un document à ce sujet) lui répond qu'il s'agit d'une théorie d'algèbre (les semi-invariants algébriques) à laquelle on a trouvé des applications en statistique, mais non d'une théorie statistique proprement dite. Soit à estimer sur échantillon les moments d'une population de N éléments. Le passage des moments d'échantillon aux moments de population, en substituant brutalement les seconds aux premiers, serait tout à fait incorrect : de tels estimateurs ne valent rien. On imagine alors l'existence d'une distribution de probabilité, dont la population serait un grand échantillon. Cette distribution possède des cumulants κ, c'est-à-dire que sa fonction caractéristique a pour logarithme une fonction que nous supposons développable en série entière. Soit κ_1, $\kappa_2/2$, $\kappa_3/6$, ... les coefficients de cette série. κ est la lettre grecque kappa minuscule.

On sait estimer correctement chaque κ par K (lettre latine majuscule) avec les données du grand échantillon et par k avec celles du petit échantillon.

Il se trouve qu'alors les k sont les estimateurs corrects des K quand on ne connaît que le (petit) échantillon.

La lettre K se dit *Kay* en anglais. On peut donc dire que les « monokays » (où K a un seul indice) sont les *cumulants* de la population et k ceux de l'échantillon.

Il existe des formules de passage entre moments et cumulants, suivant le schéma :

$$\begin{array}{ccc} m & \xrightarrow{\text{formule}} & k \\ & & \downarrow \text{ estimation} \\ \mathrm{M} & \xleftarrow{\text{formule}} & \mathrm{K} \end{array}$$

m : moments d'échantillon, M : moments de population.

Ainsi finirait-on pour trouver l'estimation correcte des M en fonction des *m*. Mais Tukey a introduit des k et K à plusieurs indices (alors que les cumulants d'une distribution d'une seule variable n'ont qu'un indice). Il s'agit d'un artifice de calcul pour faciliter l'obtention des formules utiles; et *tout un calcul symbolique* se greffe là-dessus. Mais ceci nous entraînerait trop loin.