JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

PIERRE THIONET

Quelques aspects de la théorie des sondages

Journal de la société statistique de Paris, tome 101 (1960), p. 99-112

<http://www.numdam.org/item?id=JSFS_1960__101__99_0>

VI

QUELQUES ASPECTS DE LA THÉORIE DES SONDAGES

Ire PARTIE

LE PROBLÈME DE GOODMAN ET KISH (et sa solution)

La technique statistique est un domaine aride et assez méprisé. En ce qui concerne les théories mathématiques autorisant tel ou tel procédé d'enquête par sondage, le statisticien mathématicien français lui-même en voit généralement mal l'intérêt; cependant que le statisticien tout court, l'économiste, l'administratif, le financier, ne s'intéressent d'abord qu'aux résultats de l'enquête, accessoirement parfois au travail des enquêteurs ou des « dépouilleurs » d'enquête; l'idée reçue est que le mathématicien interprète les données, on n'imagine pas qu'il ait à organiser l'obtention des données. A vrai dire le problème du meilleur des tirages au sort de 4 communes rurales par le directeur régional de l'I.N.S.E.E. de Lille ou de Strasbourg ne paraît pas susceptible de passionner qui que ce soit (et surtout pas un mathématicien). Telle est pourtant la question que nous allons reprendre (et mener à bien) après qu'elle ait intrigué pas mal de monde aux États-Unis, voire en Inde. Il est vrai que, dans bien d'autres domaines, tels la religion, les habitants de ces grands pays ont des centres d'intérêt fort éloignés de ceux de nos compatriotes.

L'intérêt que nous y trouvons personnellement n'est pas dans l'économie « de quatre sous » que nous ferions faire à nos directeurs régionaux (ce qu'un américain ferait sans doute ressortir) mais dans le pont que jette ce problème entre les théories mathématiques jusqu'ici profondément isolées les unes des autres.

Exposé d'un problème

Lors d'un exposé au Séminaire de Calcul des Probabilités du Professeur G. Darmois, le 20 janvier 1959 (1), nous signalions comment la théorie des Programmes linéaires s'était introduite en Inde dans la méthode des sondages à propos d'un problème de Lahiri. Nous disions comment Des Raj (2) avait vu qu'il s'agissait en fait d'un vieux problème (différemment posé par Keyfitz) et que le problème (resté sans solution théorique satisfaisante) de Goodman et Kish s'apparentait manifestement au même sujet. Nous ne comprenions d'ailleurs pas très bien pourquoi Des Raj abandonnait à d'autres le soin d'étudier le problème de Goodman et Kish. Il ne nous paraissait pas douteux que ledit problème serait bientôt réglé (par nous-même ou par d'autres). (Voir Note A en annexe.)

Il nous intéressait à divers titres. D'une part parce que nous avions eu jadis l'occasion d'essayer le procédé de Goodman et Kish dans le cadre des sondages français. Nos calculs sont publiés aux Bulletins intérieurs de l'I.N.S.E.E. (1951) et dans l'Étude Théorique nº 6

(1) Le Texte de ce Seminaire et publie au *Journal de la Société de Statistique de Paris* de 1959, nos 10-11-12. GOODMAN KISH, *J. Amer. Stat Association*, septembre 1950, *Deep stratification* (TEPPING, HURWITZ, DEMING), *J. Amer. Stat Assoc.*, March 1943

(2) DES RAJ, *On the method of overlapping maps in sample surveys* SANKHYA, 17, 1 (p 96, 99) (1956).

(p. 84-93) (1953). Nous n'avions jamais eu le loisir d'approfondir les problèmes qu'il pose, mais nous savions que, pour organiser rationnellement des sondages dans le cadre des 18 régions administratives de l'I.N.S.E.E., il aurait été utile d'employer la méthode a fond; nous nous y étions exercé surtout pour la région de Lille (Nord, Pas-de-Calais) (1).

D'autre part le procédé de Goodman et Kish dérive d'un essai américain bien antérieur appelé *deep stratification* (stratification en profondeur) (2), essai expérimental qui se solda par un échec, c'est-à-dire par des calculs montrant l'existence d'erreurs systématiques importantes dues à la méthode; ceci explique que le bureau du *Census* ait dès lors exclu de ses plans de sondage tout ce qui pouvait ressembler à ce procédé d'échantillonnage dont l'idée n'est au fond pas autre chose que celle d'un plan d'expérience classique.

Pour être plus concret reprenons notre exemple d'un sondage dans la région de Lille. Supposons les communes rurales de ses deux départements classés en 4 strates suivant la proportion de population vivant de l'agriculture. Supposons que les ressources en enquêteurs et les crédits disponibles limitent à 4 le nombre de communes rurales où se rendront les enquêteurs de Lille.

Si l'on tire au sort une commune de chaque strate il n'y a aucune raison que le hasard ne donne pas (disons) 4 communes du Pas-de-Calais; la loi des grands nombres ne joue pas à cette échelle. Il peut donc sembler préférable d'être assuré d'avoir dans l'échantillon 2 communes du Nord pour 2 du Pas-de-Calais, sans pourtant regrouper 2 par 2 les 4 strates de chaque département. *Par exemple on peut trouver* intéressant le schéma de sondage que voilà (assez équilibré) :

	Nord	Pas-de-Calais
Strate 1 . . .	1	0
Strate 2 . . .	0	1
Strate 3 . . .	0	1
Strate 4 . . .	1	0

Intéressant, mais ne cadrant pas avec la méthode orthodoxe des sondages; surtout si l'on pousse l'analogie avec l'expérimentation jusqu'à « interviewer » dans chacune des 4 communes *le même nombre de ménages* ou d'individus, auquel cas l'erreur systématique est grosse.

Un premier compromis avec l'échantillonnage stratifié à 2 degrés classiques consiste à tirer dans chaque strate un *nombre de ménages proportionnel* à l'effectif de la strate. Mais il reste une erreur systématique qui peut faire perdre l'avantage évident de réduction de variance due à la méthode (la variance dans la strate i du Nord, ou dans la strate i du Pas-de-Calais est inférieure au plus égale à la variance dans la strate i de la Région de Lille).

Ceci conduit Goodman et Kish à leur procédé, lequel a le mérite d'éliminer le biais ou erreur systématique, mais comporte un calcul de variance, jugé par nous trop vite inextricable. (Voir note B en annexe.)

La lumière que nous apporte Des Raj, c'est que le problème posé ne relève plus guère du calcul des probabilités, mais de la théorie des programmes. A vrai dire les anciens traités de statistique mathématique (celui de Charles Jordan, 1925) comportait bien des chapitres (interpolation) d'où les probabilités étaient absentes; et parmi les traités modernes, celui de Cramér (H) contient des développements de calcul matriciel. Le chercheur

(1) Se rend-on bien compte de l'énergie et du mépris de l'opinion que cela nous demandait déjà? D'ailleurs les recherches furent commencées pour d'autres directions régionales.

(2) Voir P. THIONET, *Actualités Scientifiques et Indust.*, n° 1011, 1946, p. 50-51.

opérationnel ou l'économètre emprunte au probabiliste les processus stochastiques et lui prête en retour les programmes linéaires. Des professions sans doute trop spécialisées éclatent et s'interpénètrent. Telle est la rançon du progrès. Heureusement pour nous, G. Th. Guilbaud a su (en 1956) ouvrir nos yeux au monde du « programming ».

La Théorie mathématique nouvelle : Symboles matriciels figurant un Sondage.

a) Soit une matrice M à m lignes et m colonnes dont les cases ne renferment que 1 ou 0 dans chaque ligne une fois 1 et $(n - 1)$ fois 0 (i désigne une ligne et j une colonne).

$$\mathrm{M}_h = \left|\begin{array}{c|c} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{array}\right|$$

$$m = 4, \qquad n = 2$$

Soit $\mathrm{M}_1\ \mathrm{M}_2 \ldots \mathrm{M}_h \ldots$ la suite des n^m matrices M rangées dans un ordre fixe, d'ailleurs arbitraire.

On dira que les M_h constituent *l'énumération des sous-programmes possibles.* Chaque sous-programme est un certain *plan de sondage.*

Soit $p_{i1} + \ldots + p_{ij} + \ldots p_{in} = 1$ (pour $i = 1, 2, \ldots m$) une certaine distribution de probabilité *sur chaque ligne* de la matrice M*

$$\mathrm{M}^* = [p_{ij}]$$

b) On appellera *programme* Q tout vecteur $q\ (1)$, $q\ (2) \ldots q\ (h) \ldots q\ (n^m)$,

avec

$$\Sigma\, q\ (h) = 1, \qquad q\ (h) \geqslant 0,$$

(les $q(h)$ sont donc les probabilités respectives des sous-programmes M_h dans le programme Q) satisfaisant aux conditions linéaires symbolisées par l'équation

$$\mathrm{M}^* = \Sigma\, q\ (h)\ \mathrm{M}_h$$

Bien entendu, beaucoup de $q\ (h)$ peuvent être pris égaux à 0 puisque ces m grandeurs ne sont liées que par $[m\ (n - 1) + 1]$ relations. On dira alors qu'on a affaire à des *programmes frontières.*

On s'intéressera spécialement aux matrices M_h affectées d'une probabilité $q(h)$ *non nulle,* par ledit programme Q. On dira que Q a pour *composantes effectives* les sous-programmes M_h en question.

c) On définira comme suit un programme *optimum* vis-à-vis d'une variable X donnée (dont l'étude est supposée constituer le but essentiel du sondage).

On considère les *sommes* x_{ij} des valeurs que cette variable X prend sur tous les éléments x_{ijkl} de la case (ij), et la somme x pour l'ensemble du tableau.

On appelle *données* dans le programme Q_h, la matrice X_h se déduisant de $[x_{ij}]$ de la même façon que M se déduit de [1].

Exemple :

$$\mathrm{M}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad ; \quad [1] = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad ; \quad \mathrm{X}_1 = \begin{bmatrix} x_{11} & 0 \\ 0 & x_{22} \\ 0 & x_{32} \\ x_{41} & \end{bmatrix} \quad ; \quad [\mathrm{X}_{ij}] = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix}$$

On appelle *estimateur* dans le sous-programme Q_h l'expression y_h analogue à

$$y_1 = \frac{x_{11}}{p_{11}} + \frac{x_{22}}{p_{22}} + \frac{x_{32}}{p_{32}} + \frac{x_{41}}{p_{41}}$$

On vérifie facilement que l'estimateur y est *sans biais ;* c'est-à-dire que la relation

$$\Sigma q_h \, M_h = [p_{ij}]$$

entraîne bien

$$\Sigma q_h \, y_h = \Sigma\Sigma x_{ij} = x..$$

On appelle *écart* du sous-programme M_h la différence $(y_h - x\,..)$ et *variance* du programme Q l'expression

$$V\,(Q) = \Sigma q_h \, (y_h - x..)^2$$

On appelle enfin *programme optimum* Q^o celui qui correspond au plus petit des $V(h)$

$$min \; \Sigma \, (y_h - x..)^2 \, q_h = min \; \Sigma q_h \, y_h^2 - x^2..$$

d) Conclusion : Obtenir le programme optimum Q^o est un pur problème de programmation *linéaire* fort banale.

e) Remarque : En pratique Q^o restera optimum pour tout un domaine de variation pour les y_h^2, ce qui autorise à penser que le caractère d'optimum n'est pas limité à la seule variable X — ce qui est intéressant à savoir en pratique.

f) On sait que Q est défini dans un domaine polyédral régulier convexe et qu'il prend sa position Q^o en un certain *sommet* de ce domaine. Nos recherches seront donc limitées en pratique aux sommets, c'est-à-dire aux *programmes frontières* (et en fait à certains d'entre eux). Il n'est pas exclu qu'on puisse trouver l'optimum sans méthode classique trop compliquée (simplexe, potentiel logarithmique, multiplexe).

g) Nous avons décrit ailleurs (Étude théorique n° 6, etc) comment s'y prendre pour obtenir des programmes frontières. Nous ne savions pas reconnaître le meilleur. Il suffira de comparer les $\Sigma y_h^2 = W\,(Q)$; et le meilleur de deux programmes Q_1 Q_2 sera celui dont le W (Q) sera le plus petit ; on peut donc ne pas ambitionner un optimum absolu et se contenter d'un optimum relatif.

h) A notre successeur à l'I.N.S.E.E. de procéder à quelque application numérique de notre méthode ; espérons qu'elle n'exigera pas un calculateur électronique comme c'est le cas pour beaucoup de programmes linéaires.

Raccord avec le sondage stratifié à deux degrés.

Il reste à raccorder la théorie précédente à celle des sondages. Supposons que l'on tire au 1er degré des *communes* (avec probabilités proportionnelles à leur taille) et au 2e degré des *ménages*. Tirons pour la strate *i* une commune du *Nord* (disons) avec les sous-programmes Q_1, Q_3 ... Q_{15} et une commune du *Pas-de-Calais* avec les sous-programmes Q_2, Q_4 ... Q_{16} ; dans cette commune tirons N_k ménages au sort pour les interviewer. Nous voilà revenus à l'horizon familier du statisticien. De toute façon on sait écrire :

l'estimateur $S_i \, x_{ijkl}/N_k$ de la moyenne par ménage, strate (i).
Comme le nombre de ménages de (ij) est connu, nous en déduisons :
l'estimateur x'_{ij} de x_{ij} comme dans la théorie banale et classique.

Si l'on s'en était tenu là, la commune tirée au sort appartiendrait à la colonne

$$j = 1 \text{ avec probabilité } p_{i1}$$
$$j = 2 \text{ avec probabilité } p_{i2} \quad \text{etc...}$$

et notre x'_{ij} serait le x'_i classique, dont la variance est connue (V_i). Avec le procédé de Goodman et Kish (au contraire), le 1er degré de sondage normal fait place en fait à un *degré avant premier*, c'est-à-dire le tirage de Q_h avec probabilité $q(h)$, que suit un 1er degré banal mais dans le cadre d'un seul département.

Donc il convient de distinguer :

1° *Variance liée* par l'avant-premier tirage Q_h (dans les m sous-strates retenues, les *tirages sont indépendants*) $V(Y \mid Q_h)$;

2° *Variance additionnelle*, due à *l'avant-premier tirage :* V (Q).

1. La première est de la forme

$$\mathcal{V}(Y \mid Q_h) = \Sigma\, \lambda_i\, \Sigma\, p_{ij}\, \sigma_{ij}^2 \qquad \text{au lieu de } \Sigma\, \lambda_i\, \sigma_i^2,$$

avec une relation de la forme :

$$\sigma_i^2 = \Sigma_j\, p_{ij}\, \sigma_{ij}^2 + \Sigma_j\, p_{ij}\, (\overline{x_{ij}} - \overline{x}_i)^2$$

Autrement dit : on gagne de la variance, proportionnellement aux écarts entre les moyennes x_{ij} et la moyenne $\overline{x}_i$ (moyennes *par ménage*).

(C'est seulement si la moyenne du Pas-de-Calais diffère beaucoup de celle du Nord qu'il est indiqué de *penser* au procédé de Goodman et Kish.)

2. La seconde de ces variances n'est pas autre chose que V (Q) défini plus haut. Toute la justification de notre méthode réside donc dans le fait que la *variance dite liée par* Q_h ne dépend pas du tout de notre choix en faveur de tel ou tel programme Q.

L'effet de notre choix de Q se concentre sur V (Q) = variance additionnelle due à l'avant-premier tirage, — composante qu'il convenait donc bien de rendre minimum.

Une fois Q choisi, tout se passe comme si on tirait la commune échantillon de (i) dans la colonne (j) avec la probabilité p_{ij} *indépendamment* de tout ce qui se passe dans les autres strates $i' \neq i$ Nous appréhendions une ruée des covariances entre les strates; et elles n'existaient pas en fait.

NOTES ANNEXES

Note A. Des Raj à vrai dire a étudié dans Sankaya (1956) un tout autre problème : il s'est imposé de rendre minimum le trajet des enquêteurs visitant les communes. Il reconnaît d'ailleurs que rendre minimum la variance à coût constant serait plus intéressant.

En France, le coût d'enquête dans quatre communes rurales d'une même direction régionale est en gros formé d'un terme proportionnel au nombre de questionnaires, plus des frais de transports ne dépendant guère de l'implantation des communes. L'enquête a lieu en outre dans de nombreuses communes urbaines, et le même enquêteur visite à la fois des communes urbaines et des rurales regroupées de façon à réduire les frais. L'optimum pour un nombre donné de communes équivaut en France à l'optimum pour un coût donné de sondage. Le problème qu'avait traité Des Raj n'est pas sans intérêt mais n'a jamais constitué à nos yeux le problème de Goodman et Kish.

Note B. Goodman et Kish ont le grand mérite d'avoir vu qu'en tirant au sort parmi un ensemble convenable Q de Matrices du type M_h (schématisant un plan de sondage) affectées de probabilités convenables $q(h)$ on pouvait s'arranger pour :

— avoir un estimateur sans biais;
— avoir une forte probabilité de tirer un plan de sondage intéressant;
— avoir une faible probabilité de tirer un plan de sondage néfaste.

Nous avions constaté qu'on pouvait construire de tels ensembles Q de bien des façons.

Mais le calcul de variance et le choix du plan Q optimum restaient à élucider, ce qui est fait ici.

2e PARTIE

POUR UNE MODIFICATION DE LA NOTION CLASSIQUE DE PLAN DE SONDAGE OPTIMAL

Nous allons développer à présent une conception que nous croyons nouvelle du plan de sondage optimum ou optimal. Il s'agit au fond d'une transposition au domaine des enquêtes par sondage, d'une théorie d'économétrie bien connue : celle de la répartition optimum des ressources dans un organisme de production.

Voici ce que nous écrivons sur cette question dans un article à paraître dans *Estadistica :*

« Neyman avait imposé à son estimateur (d'une population stratifiée) une liaison très simple : il supposait que *l'effectif* total de l'échantillon était donné. C'est Yates (1935) qui introduisit l'hypothèse d'un *coût* moyen différent d'une strate à l'autre et lia l'estimateur par la condition du coût total constant. Neyman fit sienne cette conception (1938).

« Pour les sondages à plusieurs degrés, les divers auteurs (par exemple Thionet, *J.S.S.P.*, 1948) ont admis, pour simplifier, qu'on ne maintenait plus constant ni l'effectif de l'échantillon, ni le coût du sondage, mais bien l'*espérance mathématique* de cet effectif ou de ce coût devenu *aléatoire.*

« Finalement il s'agit toujours de rendre minimum une certaine fonction $V(p_i)$ de paramètres p_i liés par une condition $C(p_i) = 0$, problème classique d'*extrêmum lié* qu'on résout en introduisant une variable supplémentaire L appelée multiplicateur de Lagrange. On ne semble guère avoir eu jusqu'ici l'occasion d'imposer aux p_i simultanément plusieurs conditions $C_j = 0$, et par conséquent de faire intervenir plusieurs multiplicateurs de Lagrange L_j; cependant nous avons signalé dans la *Revue de l'Institut International de Statistique* (1954) qu'en France (et vraisemblablement ailleurs) la liaison à la Neyman-Yates n'était pas tres réaliste. Une enquête par sondage publique nécessite la mise en œuvre de moyens divers en personnel, matériel, crédits, qui ne sont pas du tout interchangeables; même si c'est finalement la puissance publique qui finance en totalité l'enquête (ce qui n'est pas forcément le cas, puisque bien des sondages sont d'intérêt mixte, à la fois public et privé), il n'est pas du tout indifférent de savoir que les enquêteurs sont pour partie des employés à temps complet des offices locaux de statistique, et pour partie du personnel occasionnel payé en fonction du nombre de questionnaires remplis. Et même lorsque les frais de voyage des enquêteurs du 1er type et le salaire des enquêteurs du 2e type sont payés par la même Caisse publique, il s'agit de deux opérations comptables distinctes, pour lesquelles un budget sain prévoit des imputations sur des crédits distincts.

« C'est pourquoi nous pensons que la théorie des sondages doit prévoir la maximisation d'une expression de la forme :

$$\mathcal{V}(p_i) + L_1 \cdot \mathcal{C}_1 \cdot (p_i) + L_2 \cdot \mathcal{C}_2 (p_i) + L_3 \cdot \mathcal{C}_3 (p_i) = F (p_i, L_j)$$

« Il faut d'ailleurs ajouter, pour être complet, que, si les *ressources* dont dispose le statisticien pour son enquête sont toutes limitées, il n'est pas prouvé qu'on doive *épuiser* la totalité de ces ressources pour atteindre le minimum de $\mathcal{V}(p_i)$. Les conditions réellement imposées sont de la forme $\mathcal{C}_j (p_i) \leqslant 0$; le signe « inférieur à » est remplacé par le signe « égal à » quand le type j de ressources est épuisé. »

Donnons ici quelques détails de plus que dans *Estadistica.*

Pour simplifier, limitons-nous aux 3 types de ressources suivantes constituant un modèle assez réaliste.

$j = 1$: crédits pour frais de mission et de déplacements du personnel permanent.

$j = 2$: ressources en personnel permanent (exprimées en journées de travail).

$j = 3$: crédits pour payer des enquêteurs supplémentaires, les charges sociales et les remboursements de frais y afférents.

On impute en fait sur les ressources de type $j = 1$ et 2 les instructeurs et inspecteurs qui s'occupent des enquêteurs supplémentaires ($j = 3$). Mais on peut admettre pour un calcul théorique que les ressources c_1 et c_2 disponibles sont comptées après déduction des servitudes accessoires.

La particularité de ce modèle est que les frais de type 1 et 2 sont *liés et complémentaires;* sauf dans la zone où résident les enquêteurs permanents, dans laquelle il n'y a pas de frais de type 1. D'autre part les frais de type 3 sont *substituables* à ceux de type 1 + 2 ou de type 2 seul.

En effet les enquêteurs supplémentaires se justifient :

— soit dans de très grandes villes (telles Paris) où il y a insuffisance d'enquêteurs permanents;

— soit au contraire dans des localités excentriques, éloignées des résidences des enquêteurs permanents (c'est-à-dire des offices de statistique).

Admettons qu'on puisse distinguer 3 grandes strates dans l'univers sondé :

Grande Strate 1 : localités où les frais de type 1 et 2 *vont de pair;*

Grande Strate 2 : localités où les frais de type 2 sont seuls;

Grande Strate 3 : localités où les frais de type 3 sont seuls;

et admettons qu'il s'agit bien de *strates* d'échantillonnage, de sorte que la variance d'échantillonnage est de la forme

$$\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 + \mathcal{V}_3$$

C'est un cas un peu fictif, car on ne décide pas *a priori* que telle strate est réservée aux enquêteurs supplémentaires. Mais c'est assez commode pour raisonner.

Que devient alors la théorie de Neyman? Il vient :

$$d\,(\mathcal{V} + \Sigma\, L\, \mathcal{C}) = d\,(\mathcal{V}_1 + \mathcal{V}_2 + L_1\, \mathcal{C}_1 + L_2\, \mathcal{C}_2) + d\,(\mathcal{V}_3 + L_3\, \mathcal{C}_3)$$

On réalise séparément :

$$d\,(\mathcal{V}_3 + L_3\, \mathcal{C}_3) = 0$$
$$d\,(\mathcal{V}_1 + \mathcal{V}_2 + L_1\, \mathcal{C}_1 + L_2\, \mathcal{C}_2) = 0$$

La première de ces relations justifie notre affirmation (dans *Estadistica*), au sujet des enquêteurs extérieurs (1); mais à *l'intérieur de la grande strate 3* il y a lieu de respecter la répartition optimale banale; si l'enquête à Paris coûte en moyenne par questionnaire 4 fois plus cher que l'enquête dans une petite localité, il faudra prendre proportionnellement deux fois d'enquêtés à Paris que dans la petite localité.

(1) « Il est facile de comprendre qu'on réduit la variance $\mathcal{V}$ en utilisant jusqu'à épuisement les crédits c_3. »

La seconde de ces relations va nous créer au contraire quelques difficultés :

a) Si les ressources en frais de déplacement (type 1) n'existaient pas; on aurait une répartition optimum de Neyman.

$$d\,(\mathcal{V}_1 + \mathcal{V}_2) + L_2\, d\,\mathcal{C}_2 = 0$$

avec

$$d\,\mathcal{C}_2 = d\,\mathcal{C}_{21} + d\,\mathcal{C}_{22}$$

Les frais $\mathcal{C}_2$ étant ventilés entre les strates 1 et 2, ceci donne une ventilation des ressources c_2

$$c_2 = \mathcal{C}_{21} + \mathcal{C}_{22}$$

Répartition optimum $= c_{21} + c_{22}$.

Les frais de déplacement $\mathcal{C}_1$ sont en fait étroitement liés aux frais d'enquête $\mathcal{C}_{21}$; posons

$$\mathcal{C}_1 = f\,(\mathcal{C}_{21})$$

d'où

$$\overline{c_1} = f\,(c_{21})$$

Par rapport à l'optimum de Neyman $\overline{c_1}$ de frais de déplacement, les ressources réelles c_1 sont en dessous, au-dessus ou à niveau.

Si $c_1 = \overline{c_1}$, l'optimum de Neyman convient au problème posé.

Si $c_1 > \overline{c_1}$ l'optimum de Neyman n'utilise pas toutes les ressources.

Si $c_1 < \overline{c_1}$ l'optimum de Neyman ne peut être atteint, les ressources de type 1 constituant un goulot d'étranglement.

b) Si les resources de type 1 entrent *seules en jeu*, et si on veut les épuiser, on aura :

$$d\,\mathcal{V}_1 + L_1\, d\,\mathcal{C}_1 = 0$$

ce qui conduit à utiliser une certaine quantité de ressources c_2, définie par

$$c_1 = f\,(\overline{c_2})$$

Si $\overline{c_2} = c_2$, il n'y a plus moyen d'enquêter dans les strates 2.

Si $\overline{c_2} > c_2$, les ressources de type 2 constituent le goulot d'étranglement et empêchent d'épuiser celles de type 1.

Si $c_{21} < c_2 < c_2$, l'échantillon des strates 2 est plus petit que l'optimum de Neyman.

Si $c_{21} = \overline{c_2}$, donc $c_1 = f\,(c_{21})$, l'échantillon des strates 2 est égal à l'Optimum de Neyman.

Si $c_{21} > c_2$, l'échantillon des strates 2 est plus grand que l'Optimum de Neyman.

c) L'optimum consistera à abaisser $(\mathcal{V}_1 + \mathcal{V}_2)$ le plus possible, avec

$$f\,(\mathcal{C}_{21}) = \mathcal{C}_1 \leqslant c_1, \qquad \mathcal{C}_{21} + \mathcal{C}_{22} \leqslant c_2$$

sans épuiser nécessairement les ressources. Pour simplifier, traitons seulement le cas suivant (que Hajek qualifierait de canonique).

$$\mathcal{C}_{21} = \sum_i a_i\, p_i; \qquad \mathcal{C}_{22} = \sum_j a_j\, p_j; \qquad \mathcal{C}_1 = K\,\mathcal{C}_{21};$$

$$\mathcal{V}_1 = \sum_i \frac{b_i}{p_i}; \quad \mathcal{V}_2 = \sum_j \frac{b_j}{p_j}.$$

Il vient :

$$\sum_{i}\left(\frac{-b_i}{p_i^2}+L'_1 a_i\right)dp_i+\sum_{j}\left(\frac{-b_j}{p_j^2}+L_2 a_j\right)d p_j=0$$

avec

$$L'_1=K L_1+L_2$$

d'où (1)

$$L'_1=\frac{b_i}{a_i p_i^2}=\frac{V_1}{c_{21}}=\frac{K V_1}{c_1};\ L_2=\frac{b_j}{a_j p_j^2}=\frac{V_2}{c_{22}}.$$

Donc : la répartition optimale des ressources à l'*intérieur* des grandes strates 1 et 2 (*séparément*) est celle de Neyman.

Répartition optimale des ressources entre strates 1 et strates 2.

On va d'abord poser

$$c_2-c_1/k>0$$

c'est-à-dire $c_2>\bar{c}_2$, sans quoi il est impossible d'épuiser les ressources. On a donc, en supposant les ressources intégralement utilisées

$$\frac{V_1}{L'_1}=\frac{c_1}{K};\ \frac{V_2}{L_2}=c_2-c_1/k$$

d'où

$$\frac{L'_1}{K}=\frac{V_1}{c_1};\ \frac{L_2}{K}=\frac{V_2}{K c_2-c_1}$$

d'où

$$L_1=\frac{L'_1-L_2}{K}=\frac{V_1}{c_1}-\frac{V_2}{K c_2-c_1}$$

d'où

$$L_1 c_1+L_2 c_2=c_1\frac{V_1}{c_1}-\frac{c_1 V_2}{K c_2-c_1}+\frac{K c_2 V_2}{K c_2-c_1}$$

$$=V_1+V_2$$

Autrement dit :

$$\text{Min}\,(\mathcal{V}_1+\mathcal{V}_2+L_1\mathcal{C}_1+L_2\mathcal{C}_2)=2\,(V_1+V_2)$$

Notons que c_1/K est justement la quantité $\bar{c}_2$ définie en (*b*) ci-dessus.

Il n'est pas du tout forcé que c_1/K et c_2-c_1/K soient égaux à c_{21} et c_{22}, valeurs optimales de $\mathcal{C}_{21}$ et $\mathcal{C}_{22}$ définies d'abord en (*a*) en répartissant les enquêteurs c_2 sans se soucier des frais de déplacement c_1.

On peut très bien avoir

$$c_{21}>\bar{c}_2$$

c'est-à-dire un *optimum* qui diffère de celui de Neyman par une densité d'enquêtes plus grande dans les villes où résident les enquêteurs. On peut aussi avoir l'inverse : $c_{21}<\bar{c}_2$.

(1) V désigne le minimum lié de $\mathcal{V}$.

Remarque : On passe du schéma précédent à un schéma plus réaliste en précisant :

— que la strate 2 est constituée en fait d'un grand nombre de strates locales ayant chacune des ressources locales c_{21}, c_{22}, c_{23} ... qui ne sont que grossièrement liées aux besoins optimaux.

— que la strate 3 comprend également des ressources (faibles) en enquêteurs permanents (type 2) ressources qu'on commence par épuiser avant de recourir aux ressources de type 3; c'est ce qui a lieu pour les enquêtes à Paris.

Ces derniers points ne changent rien aux résultats énoncés. En revanche, si l'on examinait le cas de communes rurales qu'on pourrait atteindre :

— soit par des ressources de types 1 + 2;

— soit en implantant des enquêteurs à temps partiel (type 3);

on verrait que le choix optimum varie suivant les cas (le choix réel varie d'ailleurs suivant les régions).

Pour ce qui est des frais de type 1 et 3 (en espèces) : parmi tous les *optima* qu'on vient d'évoquer, il en est qui sont *meilleurs* que d'autres : si les ressources de type 1 et 3 peuvent être bloquées *en un fonds commun*, on réduit le risque de voir le type 1 constituer un « goulot d'étranglement ». Mais pour les frais de type 2 (définis en journées de travail) leur introduction dans le fonds commun ne serait qu'une fiction comptable sans signification profonde.

3e PARTIE

SUR LES ESTIMATEURS SANS BIAIS DU TYPE RATIO

Depuis que nous avons remis au Journal de la S.S.P. le texte (déjà mis à jour) du séminaire 1959 sur les sondages, un nouveau travail sur les estimateurs sans biais du type ratio nous est parvenu. A vrai dire, une partie de cet article de Mickey (1) avait déjà été lue à un congrès en 1954 et figurait avec la mention *unpublished* dans les bibliographies américaines.

Le principe de ces estimateurs est le même que chez Hartley et Goodman avec une beaucoup plus grande variété dans les applications. Si par exemple on a tiré un échantillon de $n = 4$ unités de sondage, sur N (sans remise), on s'arrange pour avoir un estimateur sans biais de

$$\sum_N y_i / \sum_N x_i$$

en combinant linéairement

$$y_1 + y_2 + y_3 + y_4/x_1 + x_2 + x_3 + x_4$$

avec

$$y_1/x_1;\ y_2/x_2;\ y_3/x_3;\ y_4/x_4 \text{ (comme Hartley)}$$

et aussi

$$y_1 + y_2/x_1 + x_2;\ y_1 + y_3/x_1 + x_3;\ \text{etc...}$$

et encore

$$y_1 + y_2 + y_3/x_1 + x_2 + x_3,\ \text{etc...}$$

Précisons que le biais est éliminé *en toute rigueur.*

(1) M. R. Mickey, *Some finite population unbiased ratio and regression estimators.* Journal Americ. Stat. Assoc., septembre 1959, p. 594.

On va montrer ici plus simplement comment on peut éliminer dans le même esprit *la partie principale du biais;* et nous nous en tiendrons aux tirages *avec remise* (ou, ce qui revient au même, au cas d'une urne renfermant un très grand nombre de boules).

Les coefficients de variation des x_l et des y_l et leur coefficient de corrélation étant désignés respectivement par :

$$\frac{\sigma(x)}{\overline{X}} = \gamma, \quad \frac{\sigma(y)}{\overline{Y}} = \gamma', \quad \frac{\mathrm{Cov}(x, y)}{\overline{X}\,\overline{Y}} = \rho$$

Le ratio y_l/x_l a pour biais :

$$b = \mathcal{E}\left(\frac{y_l}{x_l}\right) - \frac{\Sigma_N y_l}{\Sigma_N x_l}$$

$$= \frac{\overline{Y}}{\overline{X}}(\gamma^2 - \rho\,\gamma\,\gamma' + \ldots)$$

La partie principale de ce biais est, en valeur relative

$$B = \gamma^2 - \rho\,\gamma\,\gamma'$$

Celle de l'estimateur par ratio de type classique $\sum_n y_l / \sum_n x_l = \bar{y}/\bar{x}$ est $\frac{B}{n}$ tandis que celle du ratio moyen r

$$\bar{r} = \frac{1}{n}\sum_n \frac{y_l}{x_l}$$

est

$$\frac{1}{n}(n\,B) = B$$

Par suite, si l'on combine linéairement les 2 estimateurs $\bar{y}/\bar{x}$ et $\bar{r}$, soit

$$r' = K\frac{\bar{y}}{\bar{x}} + K'\,\bar{r}$$

le biais de l'estimateur r' sera du second ordre en $(1/n)$ si :

$$K\frac{B}{n} + K'\,B = 0$$

$$K + K' = 1$$

d'où :

$$K = n/n - 1, \quad K' = -1/n - 1$$

$$\boxed{r' = \frac{n}{n-1}\left(\frac{\bar{y}}{\bar{x}} - \frac{\bar{r}}{n}\right)}$$

l'estimateur de Hartley et C°, totalement dépourvu de biais. est (à titre de comparaison) :

$$\bar{r} = \frac{(N-1)}{N}\frac{n}{(n-1)}\left\{\frac{\bar{y}}{\bar{x}} - \bar{r}\left[1 - \frac{(n-1)}{n}\frac{N}{(N-1)}\frac{\overline{X}}{x}\right]\right\}$$

Calculons à présent la variance de r'. On aurait :

$$\mathcal{V}(r') = \left(\frac{n}{n-1}\right)^2\left[\mathcal{V}\left(\frac{\bar{y}}{\bar{x}}\right) + \frac{1}{n}\mathcal{V}(\bar{r}) - \frac{2}{n}\mathrm{Cov}\left(\frac{\bar{y}}{\bar{x}}\,\bar{r}\right)\right]$$

Il est bien connu que les premiers termes du développement de $C\mathcal{V}^2(y/\bar{x})$ sont

$$C\,\mathcal{V}^2(\bar{y}/\bar{x}) \sim \frac{A}{n} = \frac{1}{n}(\gamma^2 - 2\rho\gamma\gamma' + \gamma'^2)$$

Il est clair qu'on a :

$$\mathcal{V}(y_i/x_i) = A\,(\bar{Y}/X)^2$$
$$\mathcal{V}(\bar{r}) = \mathcal{V}\left(\frac{1}{n}\sum_n \frac{y_i}{x_i}\right) = \frac{1}{n^2}\,n\,\mathcal{V}\left(\frac{y_i}{x_i}\right)$$

La corrélation R entre $\bar{y}/\bar{x}$ et $\bar{r}$ est certainement proche de $+1$ en même temps que la corrélation entre x_i et y_i; on va poser $R = 1 - e$.

$$\mathrm{Cov}\left(\frac{\bar{y}}{x}, \bar{r}\right) = \sigma\left(\frac{\bar{y}}{x}\right)\sigma(\bar{r}) \cdot (1 - e)$$

avec

$$o \leqslant e \leqslant 2$$

D'où la première approximation, suivante du coefficient de variation de r'

$$C.V^2 \cdot r' \sim C.V^2\left(\frac{\bar{y}}{x}\right) + \frac{1}{n^2}\,C.V^2(\bar{r}) - \frac{2}{n}(1 - e)\,C.V \cdot \left(\frac{\bar{y}}{x}\right) C.\,V(\bar{r})$$
$$\sim \left(\frac{n}{n-1}\right)^2\left[\frac{A}{n} + \frac{A}{n^3} - 2\frac{A}{n^2}(1 - e)\right]$$
$$= \frac{A}{n} + 2\frac{A\,e}{(n-1)^2} = \frac{A}{n}\left[1 + 2\frac{n\,e}{(n-1)^2}\right]$$

Comme e est positif, on a donc (à ce degré de précision) :

$$CV^2(r') > CV^2(\bar{x}/\bar{y})$$

Si l'on convient de mesurer la précision d'un estimateur biaisé par la somme de la variance et du carré du biais (comme le font généralement les auteurs et bien que ce soit, à notre avis, une convention contestable) il faudra comparer à $CV^2(r')$ l'expression $CV^2\left(\frac{\bar{y}}{x}\right) + b^2$; ou à ce degré de précision :
comparer

$$\frac{A}{n} + \frac{2\,A\,e}{(n-1)^2} \quad \text{à} \quad \frac{A}{n} + \frac{B^2}{n^2}$$

autrement dit

$$\frac{2\,A\,e}{(n-1)^2} \quad \text{à} \quad \frac{B^2}{n^2}\,; \quad \text{ou} \quad (e) \quad \text{à} \quad \boxed{B^2/2\,A}$$

avec

$$A = \gamma^2 - 2\rho\gamma\gamma' + \gamma'^2$$
$$B = \gamma^2 - \rho\gamma\gamma'$$

Supposons alors que γ et γ' soient du même ordre de grandeur (convention souvent faite, parfois à la légère) :

$$A = 2\gamma^2(1 - \rho); \qquad B = \gamma^2(1 - \rho) = A/2$$

Pour avoir $2\,A\,e < B^2$, il convient que e soit inférieur à $A/8$:

$$R = 1 - e > 1 - \frac{A}{8} = 1 - \frac{\gamma^2}{4}(1 - \rho)$$

Il s'agit de la corrélation R entre $\bar{y}/\bar{x}$ et $\bar{r}$; il n'est pas absurde de s'intéresser surtout aux distributions des x_i et y_i dont le coefficient de variation est de l'ordre de 1 ou 2; autrement dit $\gamma^2/4$ peut être compris entre 1/10 et 1.

Si la corrélation ρ entre les x_i et y_i est positive et très forte ($1 - \rho$ très petit), il est raisonnable de penser que celle, R, entre $\bar{y}/\bar{x}$ et $\bar{r}$ très proche de l'unité; et si $(1 - \rho)$ est au contraire assez grand, la condition imposée à R n'est pas bien restrictive. Mais il est difficile d'être plus affirmatif (tant qu'on n'explicite pas R).

La structure de la distribution (x, y) est loin d'être définie par la corrélation ρ. Le nuage des points (x_i, y) peut être rectiligne ($\rho = 1$) mais ne pas être vertical et ne pas s'aligner avec l'origine; alors $\bar{y}/\bar{x}$ et $\bar{r}$ peuvent différer notablement et R sera sûrement trop petit : r' sera moins précis que $\bar{y}/\bar{x}$.

Notre analyse confirme d'ailleurs l'exemple numérique donné dans l'article de Goodman et Hartley (J.A.S.A., June 1958, fig. p. 501) où l'estimateur sans biais est favorable parce qu'il existe une corrélation ρ *négative* entre x et y. Mais on sait bien qu'en pareil cas on n'aurait pas employé l'estimateur par ratio $\bar{y}/\bar{x}$.

Ainsi on se demande encore dans quel cas les nouveaux estimateurs de Hartley-Mickey trouveront leur champ d'application naturel (le cas d'un nuage de points effilé et quasi *vertical*, où $\bar{r}$ coïncide avec $\bar{y}/\bar{x}$ ne semble pas avoir d'intérêt).

Personnellement ces estimateurs nous intéressent surtout du point de vue théorique; car dans nos propres recherches sur la perte d'information par sondage (1), les estimateurs biaisés posent une énigme. Nous avons étendu notre théorie aux estimateurs par ratio classique $(\bar{y}/\bar{x})$, qui font partie de ce que nous appelons les estimateurs *isomorphes*. Mais nous ne considérons pas $\bar{r}$ comme un vrai estimateur; si n tend vers N, l'échantillon coïncide avec l'univers sondé mais $\bar{r}$ diffère foncièrement de $\bar{Y}/\bar{X}$; ce n'est pas un estimateur *cohérent*. Lorsque N est infiniment grand, $\bar{r}$ ne converge pas en probabilité vers $\bar{Y}/\bar{X}$ si n croit indéfiniment.

La variance de l'estimateur r' a (*asymptotiquement*) les caractères d'une perte d'information; celle de l'estimateur sans biais $\bar{r}$ est une vraie perte d'information. La connaissance de renseignements sur la variable auxiliaire x apporte un certain gain d'information : Lequel?

Il est nécessaire de savoir quel est celui des estimateurs de Mickey qui possède *la plus petite variance ;* d'après un théorème très général déjà évoqué ailleurs il ne doit pas dépendre de *l'ordre* dans lequel sont rangés les indices des unités de sondage échantillon; il s'en suit que c'est justement $\bar{r}$, l'estimateur de Hartley (cf. Mickey, p. 598). La variance de $\bar{r}$ est connue (Goodman et Hartley, *J. A. S. A.*, June 1958) mais fâcheusement compliquée.

On peut espérer que ces questions se simplifieront d'elles-mêmes quand on leur aura fait faire des progrès substantiels.

Le calcul des variances de ces nouveaux estimateurs est facilité par une méthode

(1) Voir Étude théorique de l'INSEE, nº 7 (1958) et Bulletin de l'Inst. International de Statistique. Congrès de Stockholm, 1957.

générale très puissante (et très lourde), celle des *poly-Kays* que nous connaissons encore très mal et dont M. Fonsagrive pourrait vous parler mieux que nous-même.

***

Terminons ici ce petit exposé. Nous avons cherché à montrer que la théorie des sondages est en train de s'étendre grâce à des apports extérieurs. De plus en plus une culture mathématique élargie est nécessaire pour traiter des problèmes qui semblaient réservés aux spécialistes. Nous n'avons pas encore eu besoin des Corps de Galois, dont on sait qu'ils jouent un rôle essentiel dans la théorie moderne des carrés latins et plans d'expérimentation. Mais personne ne peut savoir ce que l'avenir nous réserve.

P. THIONET.

DISCUSSION

M. DUGUÉ : Fait observer que, dans la mesure où les plans de sondages deviennent des cas spéciaux des plans d'expérimentation, la théorie d'Algèbre moderne des Corps de Galois devrait effectivement être appelée à y jouer un grand rôle, comme dans toute la théorie des Plans d'expérimentation.

M. THIONET : Est bien d'accord avec M. Dugué que, si on ne les a pas encore rencontrés en sondage, il n'est pas prouvé que les Corps de Galois seront toujours hors de question. Le vrai problème est que les mathématiciens qui s'intéressent aux sondages acquièrent bien (le moment venu) les connaissances d'Algèbre moderne ou autres nécessaires à leurs recherches.