REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Décisions à propos de sondages

Revue de statistique appliquée, tome 3, n° 4 (1955), p. 71-80

<http://www.numdam.org/item?id=RSA_1955__3_4_71_0>

DÉCISIONS A PROPOS DE SONDAGES

par

P. THIONET

Administrateur à l'Institut National de la Statistique et des Etudes Economiques
Agrégé de Mathématiques

M. Thionet, qui professe à l'Ecole d'Application de l'I.N.S.E.E. le cours de « Théorie et pratique des sondages », est un des rares spécialistes français de l'étude de ces questions, tant au point de vue des recherches théoriques que de celui de la réalisation; domaine dans lequel il faut concilier le souci de la précision à obtenir et la nécessité de tenir compte du prix de revient.

Ce sont là deux impératifs presque inévitablement contradictoires.

Dans l'exposé ci-après, présenté au séminaire de Recherche Opérationnelle (1). *M. Thionet a examiné la théorie des enquêtes par sondage, du point de vue des décisions à prendre par le statisticien placé face à cette double exigence.*

L'ensemble de ces décisions constitue le **"Plan de sondage"**. Les **décisions** sont de deux ordres (au moins) :

d'une part celles relatives à la **structure** de ce plan,
d'autre part celles concernant (pour une structure donnée) le **volume** à donner à l'échantillon.

Par exemple, le choix des strates, celui des unités de sondage relèvent des questions de **structure** Le choix des fractions de sondage dans telle ou telle strate relève des questions de **volume.**

Il n'y a plus de décision de structure dans les deux cas extrêmes, où il est décidé que le volume de l'échantillon sera nul (pas d'enquête du tout), ou bien qu'il sera maximum (l'enquête est, comme on dit, exhautive ou complète, ou encore est un recensement).

En fonction de quoi sont prises ces décisions ?

Nous n'ignorons pas qu'il existe des cas où l'on pourrait procéder à une enquête par sondage pour confirmer ou infirmer une hypothèse, une théorie susceptibles d'avoir d'importantes répercussions financières.

Par exemple, nous lisions récemment (2) que des études récentes semblaient montrer que les agriculteurs anglais avaient tendance à gaspiller les pommes de terre de semence, le prix de revient moyen du quintal de pommes de terre pourrait, semble-t-il, être réduit assez sensiblement si les semences étaient utilisées plus parcimonieusement; il en résulterait un moindre déficit de la balance des comptes et une moindre surcharge des transports intérieurs.

On voit bien, en pareil cas, ce que l'économie de la Grande-Bretagne a à gagner à une telle étude, en admettant que les entrepreneurs particuliers (quelque cent mille) qui produisent des pommes de terre soient informés convenablement de ces résultats et prennent correctement leurs décisions propres, en conséquen-

(1) Séminaire de Recherche Opérationnelle de l'Institut de Statistique de l'Université de Paris. Exposé présenté le 8 Juin 1955.

(2) Etude à nous communiquée par le Dr YATES, de Rothamsted.

ce - ce qui n'est pas du tout certain. Bref, une étude comme celle-ci est une recherche opérationnelle typique, mais à l'échelle d'une nation (et en économie non planifiée).

Ce n'est d'ailleurs pas un pur sondage; c'est une **étude**, qui comprend en particulier des sondages, mais aussi, des calculs de comptabilité agricole et des expérimentations. En tout cas on peut **mettre en balance le coût (assuré) de l'étude et le gain (aléatoire) à en attendre.**

Malheureusement nous n'avons jamais eu encore l'occasion d'organiser une enquête par sondage dont le but était aussi clairement exprimable en unités monétaires. Les enquêtes par sondage sont actuellement destinées essentiellement à fournir des statistiques, c'est-à-dire des estimations de certaines grandeurs, qu'on pourrait autrement connaître en effectuant des recensements ou encore des enquêtes exhaustives (autrement dit auprès de l'intégralité d'une certaine population que l'enquête concerne, population incluse comme on dit dans le champ de l'enquête). Ceci est d'ailleurs le cas, qu'il s'agisse d'enquêtes officielles ou d'enquêtes privées, destinées par l'exemple à connaître quel **"pourcentage"** du public exprime telle opinion favorable à l'égard (disons) d'une certaine marque de savon.

L'enquête par sondage apporte certaines **informations** et rien d'autre, y compris pour une **étude de marché** dont elle constituera peut-être la clé de voûte.

En l'absence de ces informations, les décisions sont prises tout de même, mais les risques d'erreur sont plus grands. Il faut mettre en parallèle :

Absence des informations	:	aucun coût d'enquête
Informations partielles	:	coût du sondage
Informations totales	:	coût d'une enquête complète (recensement)

Et ce n'est pas une même personne (ou un même organisme) qui prend les décisions et qui effectue l'enquête : les résultats du sondage sont (généralement) publiés, on ignore qui les utilisera et quel usage en sera fait.

Remarque : (juillet 1955)

Voici un autre exemple intéressant. On a choisi un échantillon de 5.000 Hollandaises, qui a été soumis à une série de mensurations. Ces données ont servi à la détermination de 14 tailles standard de vêtements de confection pour un grand magasin des Pays-Bas, qui économise de ce fait la plus grande partie des frais de "retouchage" tout en accroissant son chiffre de ventes. Il s'agit donc d'une recherche opérationnelle (sur la coupe des vêtements de confection), qui comprend une opération d'échantillonnage (laquelle ne semble d'ailleurs pas avoir été menée suivant des principes très orthodoxes). Il est question que la méthode soit étendue à l'ensemble des établissements de confection des Pays-Bas; il en résultera un bénéfice pour l'économie nationale, et sans doute aussi une crise grave pour la branche du vêtement sur mesure.

Nous avons l'impression qu'il serait un peu abusif de mettre ce bénéfice - et cette crise - à l'actif de la seule méthode de sondage employée.

Voici en fait le problème réel qui se pose le plus souvent au statisticien : On lui ouvre un crédit donné, par exemple de 5 millions de francs; moyennant quoi on lui demande de fournir les estimations les meilleures possibles sur un certain nombre de points (estimations faites sur un échantillon, bien entendu). (a)

Parfois le problème est posé en inversant les deux termes précédents : On demande de fournir des estimations avec une précision **donnée** (disons une marge d'erreur de 2 %) en procédant de façon à dépenser le moins d'argent possible (ce qui revient souvent à dire : en opérant avec le plus petit échantillon possible).

Si l'on parvient à définir alors deux fonctions, à savoir la précision f et le coût g, on sera donc ramené à résoudre un problème d'extremum lié:

(a) Une personne dans l'auditoire a confirmé que tel était bien le problème qu'on posait le plus fréquemment aux organismes d'étude de marché.

rendre f maximum pour g donné,
ou rendre g maximum pour f donné,
dans les deux cas écrire $df + \lambda\, dg = 0$.

C'est comme cela que le problème est traité, depuis des années, dans toute la littérature sur les sondages. Nous verrons plus loin ce qu'il faut en penser.

PREMIÈRE PARTIE

LA THÉORIE CLASSIQUE

Pour commencer nous suivons donc les divers manuels existant actuellement: de Yates, Deming, Hansen-Hurwitz-Madow, Sukhatme, Cochran; et les divers cours de sondage professés récemment. Voici comment on procède.

I. - LA PRÉCISION D'UN SONDAGE.

a) On définit la précision du sondage par **rapport à une variable x privilégiée.** Presque toujours l'enquête fournit simultanément des données concernant d'autres variables y. Par exemple les enquêteurs procèdent à des "interviews", au cours desquelles on remplit des **"questionnaires"** ; à chaque question du questionnaire (qui en comprend au moins dix, peut-être cent ou deux cents) correspond **grosso modo** une variable étudiée. On choisit parmi celles-ci la variable pratiquement la plus importante, celle avec laquelle on présume qu'un grand nombre d'autres seront en corrélation positive étroite.

Par exemple, pour un questionnaire concernant une entreprise industrielle ou commerciale, le **chiffre d'affaires** joue un rôle essentiel; on prendra x_i égal au montant du chiffre d'affaires de l'entreprise (i).

b) On définit assez arbitrairement une **mesure** de la précision : On considère un certain estimateur correct X du chiffre d'affaires total $\sum x_i$, à partir des données échantillon; par exemple si l'échantillon a été tiré au sort à la manière des boules d'une urne, on pourra prendre :

$$X = \frac{1}{f} S\, x_i$$

: S = symbolise une sommation étendue à l'échantillon
: f = fraction sondée

On adopte comme mesure de la précision du sondage l'inverse de la **variance** de cet estimateur soit :

$$1/V(X)$$

Et ceci n'a de sens, bien entendu, que si X est une variable **aléatoire** connue, de façon qu'on puisse expliciter l'expression de V (X), connaissant la façon dont l'échantillon est tiré au sort et la forme même de l'estimateur X.

De même, tant qu'on ne connait pas la manière dont l'échantillon est **désigné,** la notion d'estimateur correct n'a pas de sens, car "correct" signifie que l'espérance mathématique de l'estimateur X n'est autre que la quantité à estimer.

II. - LE COÛT D'UN SONDAGE.

On trouve dans les manuels des formules relativement simplistes pour représenter le coût; c'est ce qu'on appelle les "fonctions de coût".

Par exemple on peut schématiser à l'extrême :

Si l'on tire $n_1\ n_2\ \ldots\ n_h$ unités de sondage des strates N° 1, 2, ..., h et si le coût moyen par unité soumise à l'enquête est $c_1\ c_2\ \ldots\ c_h$, la fonction de coût est

$$C = c_1\, n_1 + c_2\, n_2 + \ldots + c_h\, n_h$$

A l'opposé on peut multiplier les conventions sur les composantes du coût.

Si l'on effectue un sondage à 2 degrés, le coût comprend :

- une partie représentant les frais pour transporter les enquêteurs d'une unité primaire à l'autre (frais de transport + heures perdues),

- une partie représentant les frais d'enquête à l'extérieur de l'unité primaire (frais de transport + heures perdues)

Les heures sont elles-mêmes décomposées en :

temps d'interview
temps morts et temps de déplacement à l'intérieur de l'unité primaire

etc .....

Chaque composante du coût sera alors évaluée empiriquement (d'après les données concernant des enquêtes antérieures, par exemple).

III. - LA NOTION DE PLAN DE SONDAGE OPTIMUM.

On dira que le plan de sondage est optimum :

- lorsque, pour un coût donné, il conduit à l'estimation la plus précise;
- ou bien lorsque, pour une précision donnée de l'estimation, il minimise le coût.

Autrement dit : ou bien $C = k$ & V minimum
ou bien C minimum & $V = k$

Dans le cas où C et V dépendent de paramètres $\lambda\ \mu\ \nu$, on est conduit à un problème classique d'extremum lié, et à un système d'équations

$$\frac{\frac{\partial C}{\partial \lambda}}{\frac{\partial V}{\partial \lambda}} = \frac{\frac{\partial C}{\partial \mu}}{\frac{\partial V}{\partial \mu}} = \frac{\frac{\partial C}{\partial \nu}}{\frac{\partial V}{\partial \nu}} \qquad (1)$$

Ouvrons une parenthèse sur la nature des paramètres qui interviennent ici :

1) λ , μ , ν sont des paramètres de "**volume**" qui achèvent de **définir** le plan de sondage (par exemple fractions de sondage). Le système d'équations (1) ne permet donc de définir un optimum qu'à **l'intérieur d'un plan de sondage donné.**

Par exemple on se donne le découpage de la population en unités de sondage, classées en diverses strates. On se pose seulement la question : combien faut-il prélever d'unités échantillon de chaque strate ?

2) La variance V (X) renferme des paramètres liés à la structure de la population sondée; tels que :

- les **écarts-types** , à l'intérieur de chaque strate, des variables considérées,

- les **coefficients de corrélation** entre la variable x et une variable auxiliaire y (sur laquelle on aurait des informations supplémentaires).

3) Le coût C renferme des paramètres divers, tels que :

coût moyen d'un sondage par questionnaire, dans une strate,
prix du km de chemin de fer;
prix moyen de la nuit d'hôtel;
nombre moyen de km à parcourir pour visiter une commune échantillon.

Les paramètres de types 2 et 3 figurent implicitement dans le système d'équations (1), donc dans les solutions trouvées pour $\lambda\ \mu\ \nu$; on ne pourra **expliciter** $\lambda\ \mu\ \nu$ que si l'on possède déjà des **évaluation des paramètres** de **types 2 et 3,** c'est-à-dire certains coûts unitaires, certains écarts types, etc...

Choix entre plusieurs structures de plans de sondage.

Lorsqu'on a à choisir, **à coût égal** , entre divers plans de sondage, il est bien clair que l'on retient celui qui donne la **plus petite variance.**

Par exemple, une certaine stratification des unités de sondage permettra-t-elle de réduire la variance de 10 % ? On aura intérêt à tirer l'échantillon à l'intérieur des strates en question; mais il est bien rare qu'une stratification, un perfectionnement ne supposent pas un accroissement du **coût de préparation** ou de **dépouillement**, dont il faut aussi tenir compte.

Comme on étudie toujours plusieurs variables x simultanément, telle stratification réduira V (X_1), sera insensible sur V (X_2), accroîtra V (X_3), etc... et il faudra par exemple sacrifier la variable x_3 à x_2 jugée plus importante.

Enfin, on n'a aucun moyen de trouver ainsi le plan de sondage le **plus précis** parmi tous ceux (de structures différentes) qui demeurent possibles à coût constant. Si l'on tient encore compte du fait que, lorsque l'enquête est terminée, il est souvent possible d'en tirer des estimations de variance plus ou moins faible (grâce à des "informations supplémentaires") on voit que le problème n'est pas simple, encore qu'il s'énonce tout simplement : obtenir le **maximum d'informations** (*) pour un coût donné.

*

Nous allons, dans une deuxième partie, indiquer quelques aménagements qu'il convient (à notre avis) d'apporter actuellement à la théorie classique.

DEUXIÈME PARTIE

AMÉNAGEMENTS DIVERS QU'IL CONVIENT D'APPORTER A LA THÉORIE CLASSIQUE

I. - COÛT MOYEN OU COÛT MARGINAL ?

La théorie classique présente un aspect assez primaire - elle n'est manifestement pas due à un économiste. Le statisticien qui prend sa décision en fonction de son seul coût total (ou coût moyen) semble ignorer le coût marginal.

Historiquement, le premier problème "classique" doit avoir été posé par Neyman (Journ. Royal Stat. Soc. 1934) **de façon encore plus simple,** - en ne distinguant pas un coût moyen différent pour chaque strate. - On voulait simplement **répartir au mieux** (problème de l' "optimum allocation") un échantillon d' effectif total donné ($n_1 + n_2 + \ldots + n_h$ = constante) entre des strates (où les unités étaient tirées au sort à la façon des boules de l'urne de Bernoulli).

En tous cas on suppose toujours que chaque problème de sondage est indépendant des autres; on n'imagine jamais qu'on confie l'enquête à un **organisme** spécialisé, ayant des frais généraux qu'il peut avoir intérêt à alléger en travaillant parfois "à perte"; c'est-à-dire qu'à la fonction de coût des frais directs il convient d'ajouter une fraction des **frais généraux** - mais une fraction élastique plutôt qu'une quote-part. Que dire d'ailleurs du cas (qui n'est pas tellement rare) de 2 enquêtes **jumelées,** les frais de déplacement par exemple servant à la fois aux 2 enquêtes ? Que devient alors la fonction de coût ?

Si l'on passe du cas de l'organisme spécialisé à celui du service de statistique effectuant **entre autres travaux** des enquêtes par sondage (à titre accessoire), il devient encore plus difficile de parler de **"fonction de coût"** (c'est ce que nous reverrons à propos de l'INSEE). En tous les cas il est clair que c'est le **coût marginal** (autrement dit les **frais ajoutés** correspondant à un sondage de plus, tâche minime à côté de la masse des autres travaux) qui devrait intervenir dans la théorie et non pas le **coût total.**

II. - CARACTÈRE ALÉATOIRE DU COÛT.

a - La théorie développée dans les manuels ne tient pas compte du fait que, dans les problèmes pratiques, le coût est un élément aléatoire.

(*) Ce qui, pour moi, est synonyme de : réduire le plus possible la variance de l'estimateur X.

Signalons que Birnbaum et Sirken (≠), dans un article de 1950, ont utilisé déjà un coût aléatoire C, avec, par exemple :

$$E(C) = 359 \text{ dollars}$$
$$\sigma(C) = 8{,}0 \text{ dollars}.$$

Il est certain que, si C est aléatoire, il convient de **réviser** la théorie. Par exemple, au lieu de C = k on imposera

$$E(C) = k \text{ , pour } V(X) \text{ min;}$$

et c'est bien le plus simple; mais il y aura des cas où il sera plus indiqué d'imposer, par exemple : $C \leqslant k$ (avec une quasi certitude).

Autrement dit : $E(C) + 3\,\sigma(C) = k$

b - Nous avons ainsi le moyen de traiter des **problèmes nouveaux** que la théorie classique laissait de côté. **En voici un exemple très simple :**

Soit 2 strates comprenant m et m' unités, dont on tire n et n' respectivement à la façon de Bernoulli. La théorie classique donne

$V(X) = m^2 \frac{\sigma^2}{n} + m'^2 \frac{\sigma'^2}{n'}$: avec l'estimateur $X = m\,\overline{X} + m'\,\overline{X}'$

$C = c\,n + c'\,n' = c_o$: ($\overline{X}$ et $\overline{X}'$ étant les moyennes échantillons dans chaque strate).

et l'optimum $$\frac{n}{\frac{m\sigma}{\sqrt{c}}} = \frac{n'}{\frac{m'\sigma'}{\sqrt{c'}}} = \frac{c_o}{m\sigma\sqrt{c} + m'\sigma'\sqrt{c'}}$$

Or ce calcul ne tient pas compte du fait que (tirage de Bernoulli) certaines unités de sondage sont probablement tirées plusieurs fois, tout en n'étant visitées qu'une seule fois par l'enquêteur. Ainsi

$$c\,n + c'\,n' = C_o$$

est le **plafond** du coût, non le coût lui-même (probablement assez inférieur à c).

Il est facile de voir que l'aléatoire C a pour espérance mathématique :

$$E(C) = cm\left[1 - \left(1 - \frac{1}{m}\right)^n\right] + c'm'\left[1 - \left(1 - \frac{1}{m'}\right)^n\right]$$

On peut donc changer de problème et lier V (n,n') par une condition du type

$$E(C) = \text{constante}$$

qui conduit à un "optimum" différent du précédent.

III. - INDICATIONS SUR UNE EXPÉRIENCE DE RECHERCHE PRATIQUE DE FONCTIONS DE COÛT.

a - En France, pour l'INSEE, ces dernières années, nous avons essayé de poser logiquement notre problème du choix du plan des enquêtes par sondage. Pour cela il a été procédé non seulement à des calculs de variance, mais aussi à des essais de détermination de nos fonctions de coût. La nature aléatoire du coût nous a paru évidente.

A vrai dire, il faut distinguer (pour un office de statistique donné) les enquêtes de type courant, qui se ressemblent beaucoup et pour lesquelles on peut faire facilement une prévision assez précise du coût d'exécution par cet office, et des enquêtes de type occasionnel, pour lesquelles on ne peut guère faire des pronostics sérieux.

(≠) Voir J. Am. Stat. Assoc. March 1950, p. 98-110, ou : Etude théorique n° 6 de l'INSEE, p. 137.

Nous avions fait une série d'enquêtes assez comparables entre elles en 1950 et 1951; et nous en avons tiré des prévisions de coût pour les sondages des années 1952, 1953 et début 1954, sondages comparables aux précédents. La question est étudiée en détail dans un article qui va paraître incessamment dans la Revue de l'Institut International de Statistique.

Il fallait d'abord tenir compte des variations incessantes des tarifs de chemin de fer et des taux de remboursement des frais de mission.

En même temps nous devions procéder à des aménagements de nos plans d'enquête, pour les ajuster aux crédits disponibles pour chaque enquête et aussi de façon à nous rapprocher de l'emploi optimum des crédits, dont nous avions constaté être assez éloignés en 1950-51.

3.b - Il est arrivé tout d'abord ceci, que la théorie classique ne s'appliquait pas du tout. Le coût du sondage se composait d'éléments absolument distincts, tels que :

- frais de mission de déplacement,
- frais de timbres poste,
- traitements et salaires des enquêteurs fonctionnaires,
- vacations des enquêteurs extérieurs.

Ces frais sont imputés sur des crédits **distincts;** il est impossible pratiquement de virer des crédits d'un poste à l'autre du Budget, en vertu des règles comptables de l'administration française, règles qui (vous le savez) sont particulièrement odieuses.

Au lieu d'avoir une condition unique liant V ($\lambda \mu \upsilon$), à savoir

$$c(\lambda \mu \upsilon) = k$$

nous avions donc plusieurs conditions que nous appellerons les **goulots d'étranglement**

$C_1(\lambda \mu \upsilon) = k_1$ (frais de mission et déplacement)

$C_2(\lambda \mu \upsilon) = k_2$ (enquêteurs extérieurs)

Et nous avions, en outre, des conditions **élastiques,** du type :

$C_3(\lambda \mu \upsilon) < k_3$ (heures de fonctionnaires)

qui sont à vrai dire bien commodes, puisqu'elles laissent une certaine **souplesse** au dispositif.

Voici à quoi correspondait par exemple une telle inégalité. Nous pouvions accroître la précision de l'échantillonnage, réduire les pertes de temps des enquêteurs, etc... par certains "travaux en salle" avant enquête, ou encore améliorer les résultats par certaines complications des travaux mécanographiques de dépouillement. Nous pouvions encore augmenter la fraction de sondage dans certaines agglomérations, sièges des directions régionales et par conséquent réduire la part des erreurs d'échantillonnage correspondant à certaines grandes villes, **sans dépenser pour autant** ni frais de mission ou déplacement ni crédits pour enquêteurs extérieurs.

Néanmoins, nous ne pouvions pas aller très loin dans cette voie, car les sondages seraient vite entrés en concurrence avec les autres travaux en salle de l'INSEE et on nous aurait imposé une limite k_3 à ne pas dépasser.

3.c - Il est arrivé en outre que les types de fonction de coût, en usage à l'étranger et indiqués dans les manuels, ne s'appliquèrent pas du tout en France.

Considérons par exemple les frais de déplacement, exprimés en kilomètres SNCF de 3° classe (pour éliminer les hausses de tarif) et assimilables par conséquent à la somme des trajets parcourus par les enquêteurs. D'après une formule "classique", cette distance serait proportionnelle à **la racine carrée du nombre** de points d'enquête. Quant à nous, il n'en était rien. Après avoir éliminé les sièges des Directions Régionales et leur banlieue, nous avons constaté une certaine proportionnalité avec le **nombre de points d'enquête** lui-même. Ce qui tendrait à

faire penser que, si l'on augmente les points d'enquête, l'Administration met en mouvement un nombre croissant d'enquêteurs. En revanche on a constaté, d'une enquête à l'autre, un abaissement lent mais très régulier du taux de proportionnalité (toutes choses égales d'ailleurs), comme si l'Administration faisait des progrès **dans l'art d'organiser des circuits** et de faire visiter aux enquêteurs plus de communes-échantillon pour le même nombre de kilomètres parcourus.

Pour les frais de mission, nous avons eu également des surprises (et cette fois une tendance durable à leur accroissement). Mais nous n'insisterons pas davantage.

En résumé, il n'est pas question de se placer exactement à l'optimum (en admettant qu'on sache le définir avec rigueur), mais seulement de s'en approcher, en espérant qu'on ne gagnerait plus guère de précision à aller au-delà.

IV. - LE PROBLÈME DE LA STRUCTURE OPTIMUM.

Nous venons de nous occuper surtout du terme C. Il y aurait aussi beaucoup à dire sur le terme V. La théorie classique ne permet pas, nous l'avons vu, de rechercher le plan de sondage présentant la **structure** optimum.

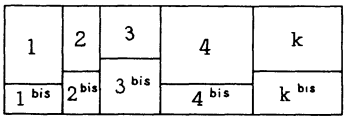
A vrai dire nous sommes rarement maître d'agir dans un tel sens. Mais il est des cas où l'action est possible et il est essentiel d'en prendre conscience.

Reprenons par exemple le problème de la "répartition optimum" (de l'échantillon ou des crédits) entre deux strates. Ce problème suppose les 2 strates données. Dans bien des cas, on reste parfaitement libre de les choisir, autrement dit de se donner la frontière qui les séparera. Pour simplifier on peut supposer que les éléments de la population sont disposés sur un axe (ce qui suppose déjà un choix) : on reste encore libre de **choisir le point limite** entre les 2 strates.

On peut donc chercher à se donner les strates conduisant à la plus haute précision, ou au plus faible coût; et ces problèmes sont du domaine de la recherche la plus actuelle.

De 1950 à 1952 **Tore Dalenius** leur a consacré 3 mémoires dans le **Journal des Actuaires scandinaves** (1); et il est bien loin d'avoir épuisé la question.

M. DESABIE a étendu (en 1954) le dernier résultat de DALENIUS, relatif à 2 strates seulement, à une population de 2 k strates



n° 1 et 1 bis
2 et 2 bis
---
k et k bis

Les frontières entre les k couples de strates sont fixes, mais les frontières entre les 2 strates de chaque couple sont mobiles; on les détermine de façon à réaliser l'optimum, c'est-à-dire la plus faible variance, les strates bis étant "enquêtées à 100 %".

Pratiquement les unités de sondage sont des établissements industriels et commerciaux, chaque couple de strates est un **groupe d'activité** (Travail des métaux, bâtiment, construction électrique, etc...), les strates bis sont formées des "gros" établissements.

Nous-mêmes avons eu à nous occuper (1954, Direction générale des Impôts) du problème symétrique du précédent : obtenir, pour **chaque groupe** d'activité, des résultats ayant une précision déterminée, moyennant un échantillon de volume convenable. Comment choisir la frontière entre petits et grands établissements pour réduire le plus possible cet échantillon ? On avait besoin de résultats concrets pour plusieurs centaines de cas; ce qui nous a conduit à une méthode très approchée, mais facile à appliquer.

Ce problème nous a conduits d'ailleurs à quelques réflexions sur lesquelles nous allons terminer.

(1) Tore Dalenius : The problem of optimum Stratification - Skandinavisk Aktuarietidskrift 1950, p. 203-213; 1951, p. 133-148 (avec Margaret Gurney), p. 61-70, 1952 -

Nous avions trouvé finalement que l'échantillon à retenir, pour obtenir, avec coefficient de variation de 1 % l'estimation du chiffre d'affaires total de **chaque** groupe d'activité (y compris les strates bis retenues en totalité) représentait 25% de la population au total.

En revanche en acceptant une précision moindre, à savoir un coefficient de variation de 2 %, on trouvait au total un échantillon représentant 17 % de la population.

Alors nous retrouvions un problème de décision à prendre, du type de ceux que nous avions voulu éviter au début du présent exposé. Valait-il mieux retenir 25 %, ou descendre jusqu'à 17 % ? Notre sentiment personnel était que "le jeu n'en valait pas la chandelle". Il faut dire que nous avions alors le point de vue du technicien. Un sondage est difficile à **organiser** matériellement et à faire exécuter correctement par les agents de l'administration; il ne faut pas croire que l'échantillon de 25 % se substituant à une enquête à 100 % va réduire des 3/4 notre charge de travail; seuls quelques ateliers mécanographiques (chiffrement, perforation) auront leur tâche réduite des 3/4; mais par exemple on va compliquer beaucoup la tâche des agents chargés de la collecte des renseignements dans des documents fiscaux, tâche qui se limitait jusqu'alors à la transmission au centre mécanographique des liasses de déclarations des redevables des T.C.A.

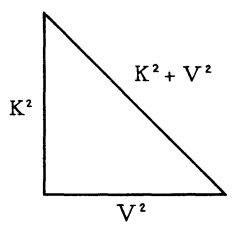
Pratiquement les échantillons de 17 % ou 25 % représenteront pour l'administration des difficultés presque égales. Autant faire un effort et obtenir un coefficient de variation de 1 %.

Ma réaction ne fut pas celle de notre Inspecteur des Finances, lequel avait manifestement raison en nous demandant de distinguer entre :

- les groupes d'activité de l' **Industrie,** pour lesquels il est intéressant de chercher une précision élevée parce que les données (chiffres d'affaires fiscaux) sont relativement bonnes;

- les groupes d'activité du **commerce** pour lesquels la précision était inutile, parce que les déclarations fiscales étaient notoirement inexactes.

Nous souhaitons qu'il existe un jour prochain une théorie mathématique qui puisse nous guider mieux dans le choix entre divers plans d'enquête, donnant des résultats de plus en plus précis pour un coût de plus en plus élevé. Notre règle empirique actuelle, c'est d'essayer d'avoir un coefficient de variation du même ordre de grandeur que la précision des données.



Si l'on admet que x est connu à K % près **sans erreur systématique**, les carrés du coefficient de variation V et de K % s'ajoutent [1] lorsqu'on parle de la précision avec laquelle est connue l'estimation X. En somme on a intérêt à ne pas gaspiller ses moyens à réduire V^2, tant qu'il n'est pas possible de réduire K^2 (en améliorant les données de base).

Dans le cas présent, il s'agissait au contraire d'erreurs systématiques, mais dont l'importance est très mal connue et très variable d'un redevable à l'autre. Si l'on corrige les déclarations relatives aux Commerces alimentaires par exemple, en les multipliant toutes par 2 par la pensée, on peut parler à nouveau d'erreurs accidentelles (et considérables).

REMARQUES.

I - La précision des données joue même un rôle non négligeable dans la comparaison entre enquête par sondage et enquête exhaustive (recensement).

Il n'y a pas de comparaison possible entre la qualité des renseignements qu'on recueille par sondage avec des enquêteurs professionnels bien instruits,

(1) Ne vaudrait-il pas mieux comparer K et 2V, donc considérer $K^2 + 4V^2$?, voire $K^2 + 9V^2$?

bien contrôlés, et celle des renseignements collectés par une armée hétérogène d'agents recenseurs occasionnels.

II - L'exposé qui précède est loin d'avoir fait le tour de tous les problèmes de décisions concernant les sondages. Par exemple un problème important (que Dalenius a affronté récemment en Suède) est de **décider** s'il vaut mieux organiser un service d'enquête par sondage sur la base d'enquêteurs peu nombreux, employés à temps complet et très mobiles ou au contraire sur la base de correspondants travaillant à temps partiel au voisinage de leur domicile. Bien entendu tout organisme de sondage a eu un jour à prendre cette décision, mais l'a fait,guidé par des considérations administratives ou humaines; il n'est pas sans intérêt d'essayer de poser rationnellement le problème, en fonction du coût et de la précision des résultats des enquêtes.

BIBLIOGRAPHIE GÉNÉRALE

THIONET P. - La théorie des sondages - Institut National de la Statistique - Paris 1953.

THIONET P. - Application des sondages aux enquêtes statistiques - Institut National de la Statistique - Paris 1953.

YATES. - Méthodes de sondages pour recensements et enquêtes - (Traduction française) - Dunod, Paris 1951.

DEMING. - Some theory of sampling - Wiley, New-York, 1950.

HANSEN, HURWITZ and MADOW. - Sample Survey methods and theory - (2 vol.) Wiley, New-York, 1953.

COCHRAN. - Sampling techniques - Wiley, New-York, 1953.