JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Le problème théorique du plan d'échantillonnage

Journal de la société statistique de Paris, tome 89 (1948), p. 136-154

<http://www.numdam.org/item?id=JSFS_1948__89__136_0>

Le problème théorique du plan d'échantillonnage.

Chaque fois qu'il le peut, le statisticien substitue au recensement complet l'enquête statistique partielle et au dépouillement intégral d'une collection de documents celui d'un simple échantillon. Le choix de l'échantillon est souvent guidé par des nécessités impérieuses qui écartent *a priori* toute idée de *choix au hasard.* Par exemple, l'expérience a montré que, si l'on distribuait au public d'Australie des cahiers où dépenses et recettes devaient figurer, la proportion des répondants était de l'ordre de 5 %. Il serait alors vain de choisir au hasard un échantillon de familles afin d'y enquêter sur leurs budgets; et force est de se contenter d'un échantillon de famille qui n'a rien d'aléatoire et qu'aucun raisonnement logique n'autorise à considérer comme fournissant une image non déformée de l'ensemble des familles.

Mais le champ des enquêtes purement empiriques se réduit, à mesure que s'accroît le souci de rationalisme. Le professeur Bowley (B. I. I. S., 1926) semble avoir, l'un des premiers, présenté une justification théorique des sondages au hasard (telles ses enquêtes sociales à Londres) et de la méthode représentative en agriculture. Le professeur Neyman (J. S. R. S., 1934 et 1938) a apporté ensuite beaucoup de lumière sur ces questions. Puis l'intérêt du problème semble s'être transporté d'Angleterre aux États-Unis où, depuis 1940, travaux théoriques et réalisations pratiques se sont amoncelés (1). En France, on aborde les mêmes problèmes.

Une enquête statistique pose de nombreux problèmes et son aspect mathématique n'est sans doute pas le plus important. D'autre part, les mathématiciens français ont rarement l'occasion de quitter la statistique mathématique pure. Les recherches purement théoriques risquent parfois de se rapporter à des problèmes dont l'intérêt pratique est des plus restreints. En revanche, le fonctionnaire chargé de monter une enquête côtoie des obstacles qu'il aurait intérêt à franchir avec des moyens mathématiques appropriés; mais il n'en dispose guère. S'il connaît, en bon statisticien, les urnes de Bernouilli et de Poisson, il se rend compte du peu de ressemblance entre ces théories et la pratique et n'a guère l'occasion de s'en servir.

Le choix de l'échantillon est conditionné par les moyens d'enquête : nombre d'enquêteurs, moyens de transport, crédits disponibles, etc. Compte tenu de

(1) THIONET, *Méthodes statistiques modernes des administrations fédérales aux États-Unis.* Hermann, 1946.

ces moyens, on arrive à des résultats présentant une imprécision théoriquement calculable. Si cette marge d'imprécision est trop grande, il aurait mieux valu renoncer à l'enquête et économiser les crédits dépensés, car il est clair que les incertitudes tenant à la mauvaise qualité des réponses viennent encore s'ajouter à celles tenant au seul choix de l'échantillon.

Il en est comme d'un ingénieur chargé de construire un pont avec des moyens limités; il lui est possible de calculer, *grosso modo*, la charge maximum que peut supporter le pont: il doit pouvoir apprécier si des camions pourront passer sur le pont.

L'art de lancer des enquêtes, comme celui de lancer les ponts, relève donc de l'ingénieur, mais tout de même de l'ingénieur mathématicien. Combien de ponts ont dû s'effondrer avant que les Romains aient su utiliser les matériaux de l'époque (pierres ou pilotis) et combien peu de nos jours?

Aux travaux de Bernouilli, les travaux récents apportent peu qui intéresse spécialement le mathématicien pur. Pourtant il nous importe beaucoup de savoir raisonner avec précision sur des échantillons beaucoup moins *coûteux* à atteindre que celui sorti des urnes de Poisson et de pouvoir comparer soit les prix de revient de deux enquêtes d'égale précision dans leurs résultats, soit la précision de deux enquêtes de coûts identiques. On se rapproche ainsi dune manière très appréciable des échantillons qu'on peut naturellement atteindre; et on peut fixer des règles de choix au hasard de l'échantillon telles qu'on soit assuré d'améliorer la façon de faire par rapport à la routine.

Ces problèmes théoriques de technique sont loin d'être simples; par exemple, la méthode de choix de l'échantillon est liée au choix de la formule par laquelle on passera des observations sur l'échantillon aux résultats relatifs à l' « univers » dont l'échantillon est extrait. D'autre part, la France ne présente pas une structure particulièrement simple : les unités administratives (commune, département) ou économiques (exploitation agricole, par exemple) sont des êtres très disparates présentant des singularités innombrables et introduisant ainsi des *variances* considérables. Enfin, on possède encore actuellement très peu d'éléments chiffrés, par exemple pour évaluer les variances ou corrélations, ou encore pour tabler sur des prix de revient précis.

Quoi qu'il en soit, on se propose de faire connaître ici certaines difficultés théoriques rencontrées en 1947, à l'Institut national de la Statistique et des Études économiques, lors de la préparation du plan d'échantillonnage d'une enquête statistique que les services du ministère de l'Agriculture devaient effectuer sur la basse-cour (1). On ne disposait d'aucun recensement de la basse-cour qui aurait été fait antérieurement; seuls subsistaient les totaux par département de celui de 1929. On s'est proposé d'établir un plan d'enquête très simple valable pour l'agriculture en général et qui permette d'attendre le recensement agricole prévu pour 1949-1950, si d'autres enquêtes sur échantillon pris au hasard devaient avoir lieu d'ici-là.

On a accepté provisoirement les indications des services départementaux sur les régions agricoles (au nombre de près de 700) et sur les catégories d'exploi-

(1) Le ministère de l'Agriculture a renoncé en juillet 1947 à faire cette enquête pour laquelle il disposait d'abord de 3 millions de crédits destinés en quasi totalité à payer les enquêteurs et à les dédommager de leurs frais.

tations (très petites, petites, moyennes, grandes, très grandes) qu'il y aurait lieu d'y distinguer. Et on est parti d'un fichier d'exploitations obtenu par les déclarations individuelles de 1942 pour 86 départements et par des documents de 1945 ou 1946 pour la Corse, le Haut-Rhin, le Bas-Rhin, la Moselle et certaines fractions du territoire particulièrement remaniées depuis 1944 (certains cantons de la Meuse, des Ardennes, etc.). On remarquera en passant que le fichier de 1942 est tenu pour plus complet que toutes les collections postérieures de documents analogues, y compris les fiches d'exploitations du recensement général de 1946.

La fiche d'exploitation porte le département, la commune, les nom et prénoms de l'exploitant, la superficie. On a admis que l'enquêteur visiterait dans sa journée 5 exploitations, pourvu qu'elles soient dans la même commune; on a cependant fait une entorse au principe du choix au hasard; on tirera 10 exploitations par commune, dont 5 seulement sont à visiter (on tient ainsi compte des inexactitudes du fichier et des exploitants impossibles à joindre). Les crédits ne permettant d'enquêter que dans une exploitation sur 100, soit 25.000 sur 2 millions et demi, il faudra donc tirer au sort 5.000 des 38.000 communes françaises.

Toutefois, on s'est aperçu qu'il ne fallait pas visiter uniformément 1 exploitation sur 100. En effet : 1° les très petites régions ont moins de 500 exploitations; 2° les régions de grande culture ne fourniraient à l'échantillon qu'une contribution insuffisante, le contraire se produisant pour les pays de petite culture; 3° la même observation est à faire pour les grandes et petites exploitations d'une même région; 4° la même observation s'applique encore pour les régions spécialisées dans l'élevage du poulet, du lapin ou de l'oie; 5° en outre, certains départements ont connu des scrupules exagérés et le moindre lopin de quelques ares (de vigne, par exemple) y est devenu une exploitation agricole.

Il importe de comprendre que l'effectif de basse-cour élevé dans une très petite ou petite exploitation est limité par la force des choses et comporte une *variance* assez faibles en soi, alors que le contraire se produit pour la grande ou très grande exploitation, qui, souvent, ne comporte aucune basse-cour, mais peut aussi en élever une très nombreuse.

On se propose, dans ce qui suit, de donner les règles de calcul nécessaires pour écrire une espérance mathématique ou une *variance;* puis de rappeler quelques résultats (déjà classiques chez les Anglo-Saxons), à titre d'exemple et d'application; enfin de passer au problème précis de l'enquête sur la basse-cour pour montrer les difficultés cachées qu'elle comportait, telle qu'on l'a entreprise. Pour terminer, on donnera des indications générales sur l'avenir des enquêtes agricoles par sondage en France.

1re Partie. — Règles de calcul (1).

I. — Soit un ensemble $\sum_k$ d'éléments $u_1\ u_2 \ldots u_k \ldots u_N$; soit $p\ (k)$ la probabilité de tirer u_k lorsqu'on tire au hasard l'*un* des u. On pose :

$$E\ [u_k] = \sum_k p\ (k)\ u_k; \quad V\ [u_k] = \sum_k p\ (k)\ [u_k - E\ u_k]^2$$

(1) Ce travail était déjà à l'impression quand M. Deming du *Bureau of the budget* a bien

et on suppose tiré un échantillon S de n éléments.

Formule 1. $\quad \mathrm{E}\left[\underset{k}{\mathrm{S}}\, u_k\right] = n \underset{k}{\Sigma}\, p\,(k)\, u_k = n\,\mathrm{E}\,[u_k].$

Cette formule est évidente si l'on remet dans l'urne l'élément tiré après chaque extraction. Lorsqu'il n'en est pas ainsi, on y aboutit de la façon suivante :

a) Si les u_k sont tous également probables, deux raisonnements sont possibles :

La probabilité de tirer u_k au premier tirage est $p\,(k) = \frac{1}{\mathrm{N}}$; si on ne l'a pas tiré $\left(\text{probabilité } \frac{\mathrm{N}-1}{\mathrm{N}}\right)$, celle de l'obtenir au deuxième tirage est $\frac{1}{\mathrm{N}-1}$, celle de ne pas l'obtenir $\frac{\mathrm{N}-2}{\mathrm{N}-1}$; alors la probabilité de l'obtenir au troisième tirage est $\frac{1}{\mathrm{N}-2}$; etc..., ce qui donne au total $\frac{\mathrm{N}}{n}$. D'où l'espérance mathématique (*formule 1'*).

$$\mathrm{E}\left[\underset{k}{\mathrm{S}}\, u_k\right] = \underset{k}{\Sigma}\, \frac{n}{\mathrm{N}}\, u_k = n \underset{k}{\Sigma}\, \frac{1}{\mathrm{N}}\, u_k \qquad \text{cqfd.}$$

On peut encore raisonner sur les combinaisons n à n des N éléments dont une, (S), est extraite :

$$\mathrm{E}\left[\mathrm{S}\, u_k\right] = \Sigma'\, \frac{1}{\mathrm{C}_{\mathrm{N}}^{n}}\, \underset{k}{\mathrm{S}}\, u_k,$$

Σ' étant la somme étendue à toutes ces combinaisons; celles qui renferment u_k sont au nombre de $\mathrm{C}_{\mathrm{N}-1}^{n-1}$, d'où la formule (1') :

$$\mathrm{E}\left[\underset{k}{\mathrm{S}}\, u_k\right] = \underset{k}{\Sigma}\, \frac{\mathrm{C}_{n-1}^{\mathrm{N}-1}}{\mathrm{C}_{n}^{\mathrm{N}}}\, u_k = \underset{k}{\Sigma}\, \frac{n}{\mathrm{N}}\, u_k \qquad \text{cqfd}$$

b) Si les u_k ne sont pas tous également probables, on imagine une loterie où l'élément u_k possède M_k billets ($\underset{k}{\Sigma}\, \mathrm{M}_k = \mathrm{M}$). On tire n billets (sans remettre après chaque coup le billet tiré dans l'urne). Il suffit d'introduire le Jeu de variables $u_{\mathrm{I}}\, u_{\mathrm{II}}, \ldots\, u_{\mathrm{K}} \ldots\, u_{\mathrm{M}}$ (où u_k est répété M_k fois) pour se ramener au cas précédent :

$$\mathrm{E}\left[\underset{k}{\mathrm{S}}\, u_k\right] = n \underset{\mathrm{K}}{\Sigma}\, \frac{1}{\mathrm{M}}\, u_{\mathrm{K}} \qquad (1')$$

$$= n \underset{k}{\Sigma}\, \frac{\mathrm{M}_k}{\mathrm{M}}\, u_k$$

$$= n \underset{k}{\Sigma}\, p\,(k)\, u_k \qquad (1) \text{ cqfd}$$

Il importe toutefois de remarquer que la formule n'est pas valable pour un choix où les gagnants sont exclus des coups suivants (ce qui peut se concevoir

---

voulu nous envoyer un exemplaire du livre édité par le *Bureau of the Census*, intitulé *a chapter in Population sampling*, dont la deuxième partie traite des mêmes questions de manière d'ailleurs beaucoup plus poussée.

du moment que l'on a $n \leqslant N$). La formule correcte serait alors (si $k_1\ k_2\ \ldots\ k_n$ sont les indices choisis, tous différents),

$$E\,[S_k\,u_k] = \Sigma'\, n!\, p\,(k_1)\,\frac{p\,(k_2)}{1-p\,(k_1)}\cdot\frac{p\,(k_3)}{1-p\,(k_1)-p\,(k_2)}\cdots$$

$$\frac{p\,(k_n)}{1-p\,(k_1)\ldots-p\,(k_2)-p\,(k_{n-1})}\,[u_{k_1}+u_{k_2}+\ldots+u_{k_n}]$$

expression qui redonne d'ailleurs $1/C_N^n$ lorsque $p\,(k) = 1/N$.

Dans la pratique de l'élaboration des plans d'échantillonnage, il arrive que l'on procède au tirage au sort d'éléments ayant une probabilité d'être tirés proportionnelle à leur taille. Il faudrait donc, en principe, procéder de la première façon (loterie) et accepter de faire figurer plusieurs fois dans l'échantillon le même élément, éventuellement, sans quoi les formules ne sont plus rigoureusement correctes. D'ailleurs la différence pratique est faible si les éléments sont nombreux et de tailles assez peu disparates.

Formule 2.

$$\begin{cases} V\left[\underset{k}{S}\,u_k\right] = n\,\sigma^2 = n\,V\,[u_k] \text{ (éléments réunis dans l'urne).} \\ V\left[\underset{k}{S}\,u_k\right] = n\,\sigma^2\,\dfrac{N-n}{N-1} \text{ (éléments non remis dans l'urne).} \end{cases}$$

où σ^2 est la *variance* des u. En particulier, $V = \sigma^2$ si $n = 1$.

a) Ces formules sont classiques et supposent essentiellement que les u_k sont tous également probables;

b) Lorsqu'il n'en est pas ainsi, il faut raisonner comme pour la formule (1), ce qui conduit à :

$$V\,[S_k\,u_k] = n\,\sigma^2\,\frac{M-n}{M-1}$$

en posant

$$M\,\sigma^2 = \sum_k [u_k - E\,u_k]^2$$

$$= \sum_k M_k\,[u_k - E\,u_k]^2$$

donc

$$\sigma^2 = V\,[u_k].$$

II. — Soit un ensemble $\Sigma\Sigma$ formé d'éléments u_{jk}; soit $p\,(jk) = p\,(j)\,p_j\,(k)$ la probabilité d'en tirer au hasard l'élément u_{jk};

$$E\,[u_{ik}] = \sum_j \sum_k p\,(jk)\,u_{jk};\quad V\,[u_{jk}] = \sum_j \sum_k p\,(jk)\,[u_{jk} - E\,u_{ik}]^2$$

et on suppose tiré un échantillon (à deux degrés) S S. On se propose de calculer :

$$E\left[\underset{j}{S}\,\underset{k}{S}\,u_{jk}\right];\qquad V\left[\underset{j}{S}\,\underset{k}{S}\,u_{jk}\right]$$

a) Lorsque les n éléments composant l'échantillon sont extraits l'un après l'autre, et chacun remis dans l'urne immédiatement, il est évident que l'on a :

$$E\,(S_j\,S_k\,u_{jk}] = n\,\Sigma\Sigma\,p\,(jk)\,u_{jk} = n\,E\,[u_{ik}] \qquad (3)$$

$$V\,[S_j\,S_k\,u_{jk}] = n\,V\,[u_{jk}] \qquad (4)$$

Si l'on introduit les probabilités liées $p_j(k)$ et les espérances mathématiques liées qui en résultent, il vient, d'une part :

$$\mathrm{E}\,[\mathrm{S}_j\,\mathrm{S}_k\,u_{jk}] = n\,u \quad (3\ bis) \quad \text{en posant} \begin{cases} u_j = \mathrm{E}_j\,u_{jk} \\ u = \mathrm{E}\,u_j \end{cases}$$

d'autre part :

$$\begin{aligned} \mathrm{V}\,[u_{jk}] &= \mathrm{E}\,[u_{jk} - u_j + u_j - u]^2 \\ &= \mathrm{E}\,[u_{jk} - u_j]^2 + \mathrm{E}\,[u_j - u]^2. \end{aligned}$$

les doubles produits disparaissant (calcul classique); d'où :

$$\begin{aligned} \mathrm{V}\,[u_{jk}] &= \mathrm{V}\,[u_j] + \mathrm{E}\,[\mathrm{V}_j\,u_{jk}[\\ \mathrm{V}\,[u_{jk}] &= \mathrm{V}\,\mathrm{E}_j\,[u_{jk}] + \mathrm{E}\,\mathrm{V}_j\,[u_{jk}]. \end{aligned}$$

ce qui permet de retenir au passage l'expression symbolique intéressante :

$$\boxed{\mathrm{V} = \mathrm{V}\,\mathrm{E}_j + \mathrm{E}\,\mathrm{V}_j} \qquad \textit{Formule 5.}$$

On passe de là à :

$$\mathrm{V}\,[\mathrm{S}_j\,\mathrm{S}_k\,u_{jk}] = n\,\mathrm{V}\,\mathrm{E}_j\,[u_{jk}] + n\,\mathrm{E}\,\mathrm{V}_j\,[u_{jk}] \qquad (4\ bis).$$

Ces formules ne sont plus correctes, dès que l'on ne remet pas les éléments dans l'urne.

b) Calcul de la valeur moyenne.

Supposons les u_{jk} disposés en échiquier (j) étant un numéro de ligne, (k) un numéro de colonne.

Si la ligne (j) comprend N_j éléments également probables, dont on extrait n_j, on a :

$$p_j\,(k) = \frac{1}{\mathrm{N}_j},$$

L'espérance mathématique liée au choix de j est, d'après (1') :

$$\begin{aligned} \mathrm{E}_j\,\mathrm{S}_k\,u_{jk} &= \frac{n_j}{\mathrm{V}_j}\sum_k u_{jk} \\ &= \varphi_j \end{aligned}$$

La probabilité de tirer φj est $p\,(j)$. Si l'ensemble comprend L lignes également probables, dont l sont choisies, on a, d'après (1') :

$$w = \mathrm{E}\,\mathrm{S}_j\,\varphi_j = \frac{l}{\mathrm{L}}\sum_j \varphi_j$$

et au total :

$$w = \mathrm{E}\,\mathrm{S}_j\,\mathrm{S}_k\,u_{jk} = \frac{l}{\mathrm{L}}\sum_j \frac{n_j}{\mathrm{N}_j}\sum_k n_{jk} \qquad (\textit{Formule 3'})$$

où apparaît bien :

$$p\,(j\,k) = \frac{1}{\mathrm{L}}\cdot\frac{1}{\mathrm{N}_j}$$

et qui se ramène à (3) lorsque tous les n_j sont égaux $\left(\text{à } \frac{n}{l}\right)$.

Si les lignes sont choisies avec des probabilités proportionnelles au nombre N_j de leurs éléments, on a, d'après (1′) :

$$w = E\,S_j \varphi_j = l \sum_j \frac{N_j}{N} \varphi_j \qquad (N = \sum_j N).$$

$$= l \sum \frac{n_j}{N} \sum_k u_{jk} \qquad (\textit{Formule } 3'')$$

où apparaît bien :

$$p\,(j\,k) = \frac{1}{N}$$

et qui se ramène encore à (3) lorsque tous les n_j sont égaux $\left(\text{à } \frac{n}{l}\right)$. La formule (3″) sous-entend d'ailleurs que le choix des lignes est fait de telle sorte que (1′) soit valable.

Plus généralement, on a :

$$E\left[\sum_j \sum_k u_{jk}\right] = l \sum_j p_n(j)\, n_j \sum_k p_j\,(k)\, u_{jk} \qquad (3''')$$

à condition que le choix des indices (j), puis des indices (k) soit fait conformément à la méthode de loterie décrite en (I 1 *b*) ci-dessus.

Quant à la formule valable dans le cas général, c'est évidemment :

$$E\left[\sum_j \sum_k u_{jk}\right] = \Sigma' P\,(j_1\, j_2 \ldots j_l)\, P\,(k_1\, k_2 \ldots k_n) \sum_j \sum_k u_{jk} \qquad (3'''')$$

la sommation Σ' étant étendue à tous les couples de combinaisons $(j_1 \ldots j_l)$ $(k_1 \ldots k_n)$ possibles.

Calcul de la variance.

Raisonnons sur un exemple.

Soit 3 variables également possibles, u, φ, w, centrées respectivement autour de a, b, c (avec, pour simplifier, $a + b + c = 0$) et dont on tire 2 au hasard. On a :

$$E\,[S\,u) = \frac{a+b}{3} + \frac{b+c}{3} + \frac{c+a}{3} = \frac{2}{3}(a+b+c) = 0$$

$$V\,[S\,u] = \frac{1}{3} E\,(\varphi + w + b + c)^2 + \frac{1}{3} E\,(w + u + c + a)^2 + \frac{1}{3} E\,(u + \varphi + a + b)^2$$

$$= \frac{1}{3} E\,[(\varphi + w)^2 + (w + u^2) + u + \varphi)^2] + \frac{1}{3}[(b + c)^2 + (c + a)^2 + a + b)^2]$$

$$+ \left[\frac{2}{3}(b + c)\, E\,(\varphi + w) + \frac{2}{3}(c + a)\, E\,(w + u) + \frac{2}{3}(a + b)\, E\,(u + \varphi)\right].$$

La dernière partie de l'expression disparaît, car $Eu = E\varphi = Ew = 0$.

On sait d'autre part (compte tenu de $a + b + c = 0$) que :

$$(b + c)^2 + (c + a)^2 + (a + b)^2 = a^2 + b^2 + c^2 = 3\,\sigma^2.$$

Enfin, on pose :

$$E\,[u\,\varphi] = \rho_{uv} \cdot \sigma_u \cdot \sigma_v; \quad E\,[\varphi\, w] = \rho_{vw} \cdot \sigma_v \cdot \sigma_w; \quad E\,[w\,u] = \rho_{wu} \cdot \sigma_w \cdot \sigma_u$$

$$E\,[u]^2 = \sigma_u^2 \quad ; \quad E\,[\varphi^2] = \sigma_v^2 \quad ; \quad R\,[w^2] = \sigma_w^2.$$

On peut d'ailleurs admettre en général que les ρ sont nuls (indépendance des variables deux à deux). Il vient alors :

$$V\,[u] = \sigma_u^2,$$
$$V\,[S\,u] = \frac{2}{3}(\sigma_u^2 + \sigma_v^2 + \sigma_w^2) + \sigma^2$$

σ^2 désignant la variance des espérances mathématiques des variables.

Passons au cas des éléments u_{jk} disposés en échiquier et où les éléments de la ligne (j) ont des chances égales d'être choisis si la ligne j est choisie. On a déjà posé :

$$w = E\,S_j\,\varphi_j$$
$$\varphi_j = E_j \sum_k u_{jk}$$
$$= \frac{n_j}{N_j} \sum_k u_{jk} = n_j\,u_j$$

On se propose d'évaluer :

$$V\left[\sum_j \sum_k u_{jk}\right] = E\left[\sum_j \sum_k u_{jk} - w\right]^2$$
$$= E\left[\sum_j \sum_k u_{jk} - \sum_j \varphi_j + \sum_j \varphi_j - w\right]^2$$
$$= E\left[S_j\left(\sum_k u_{jk} - \varphi_j\right)\right]^2 + E\left[\sum_j \varphi_j - w\right]^2$$

les doubles produits disparaissant; on a évidemment :

$$E\left[\sum_j \varphi_j - w\right]^2 = V\,[S_j\,\varphi_j],$$

que l'on sait expliciter en fonction de $p\,(j)$.

D'autre part, on sait écrire :

$$E_j\,[S_k\,u_{jk} - \varphi_j]^2 = V_j\,[S_k\,u_{jk}] = n_j\,\sigma_j^2\,\frac{N_j - u_j}{N_j - 1}$$

et on a le développement :

$$[S_j\,(S_k\,u_{jk} - \varphi_j)]^2 \equiv S_j\,[S_j\,u_{jk} - \varphi_j]^2 + 2\,S'_{hh'}\,[S_k\,u_{jk} - \varphi_h]\,[S_k\,u_{jh'} - \varphi_{h'}].$$

Enfin, on sait expliciter en fonction de $p\,(j)$:

$$E\,S_j\,[S_k\,u_{jk} - \varphi_j]^2 = E\,S_j\,V_j\,[S_k\,u_{jk}]$$

tandis que la quantité :

$$2\,E\,S'_{h\,h'}\,[S_k\,u_{hk} - \varphi_h]\,[S_k\,u_{h'k} - \varphi_{h'}]$$

peut s'exprimer à l'aide des :

$$E_{h\,h'}\,(u_{hk} - u_h)\,(u_{h'k} - u_k) = \rho_{h\,h'}\,\sigma_h\,.\,\sigma_{h'}$$

et est négligeable dans la même mesure que les coefficients de corrélation
Il reste au total :

$$V\,[S_j\,S_k\,u_{jk}] = V\,[S_j\,E_j\,S_k\,u_{jk}] + E\,[S_j\,V_j\,S_k\,u_{jk}] \qquad (\textit{Formule 6}).$$

à rapprocher de :

$$V\,[u_{jk}] = V\,[E_j\,u_{jk}] + E\,[V_j\,u_{jk}] \qquad (\textit{Formule 5}).$$

Rappelons enfin que les formules bien connues :

$$\mathrm{E}\,[a\ u] = a\ \mathrm{E}\,[u]$$
$$\mathrm{V}\,[a\ u] = a^2\,\mathrm{V}\,[u]$$

permettent de calculer à présent des expressions telles que :

$$\mathrm{E}\,[a\ \mathrm{S}_j\ b_j\ \mathrm{S}_k\ u_{jk}] = a\ \mathrm{E}\,[\mathrm{S}_j\ \mathrm{S}_k\ (b_j\ u_{jk})]$$
$$\mathrm{V}\,[a\ \mathrm{S}_j\ b_j\ \mathrm{S}_k\ u_{jk}] = a^2\ \mathrm{V}\,(\mathrm{S}_j\ \mathrm{S}_k\ (b_j\ u_{jk})].$$

2e Partie. — Application à quelques problèmes d'échantillonnage.

I. — *Un problème classique.*

Une population est supposée divisée en « strates » (par exemple groupes professionnels) désignées par l'indice :

$$1\ 2\ \ldots\ i\ \ldots\ \mathrm{L}$$

Dans chaque strate, on procède au tirage au hasard de n_i des N_i individus la composant. On note, par exemple, le salaire x_{ik} de chacun de ces individus. Une estimation bien naturelle du salaire moyen de la population est :

$$\bar{x} = \frac{1}{\mathrm{N}} \sum_i \frac{\mathrm{N}_i}{n_i} \mathrm{S}_k\ x_{ik}$$

avec

$$\mathrm{N} = \sum_i \mathrm{N}_i$$

On peut établir que cette estimation est *la meilleure*, en ce sens que c'est la seule qui :

soit une fonction linéaire des x_{ik} de l'échantillon;

ait pour espérance mathématique E (x_{ih}). Les Anglo-Saxons qualifient une telle estimation de *unbiased* (sans biais, c'est-à-dire sans erreur systématique). Les Français, tenant compte en outre de ce que son écart-type est fini et de ce qu'elle converge en probabilité vers son espérance mathématique, donc vers la valeur correcte, la nomment souvent estimation *absolument correcte.*

On se propose de choisir les paramètres nj de manière que la *variance* de l'estimation soit minimum. Celle-ci est (aux $\rho_{i\,i'}$ près) :

$$\mathrm{V}\,(x) = \frac{1}{\mathrm{N}^2} \sum_i \left(\frac{\mathrm{N}_i}{n_i}\right)^2 n_i \frac{\mathrm{N}_i - n_i}{\mathrm{N}_i - 1} \sigma_i^2$$

en posant :

$$\mathrm{V}_i\,[x_{ik}] = \sigma_i^2$$

Supposons fixé le nombre total $\sum_i n_i = n$ d'éléments sur lesquels portera l'enquête (c'est-à-dire en gros les frais d'enquête). Rendre V $(\bar{x})$ minimum, compte tenu de la condition imposée aux ni, c'est (conditions du premier ordre) un problème d'extremum lié qu'on résout au moyen d'un multiplicateur de Lagrange. On écrit l'identité en dn_i :

$$d\ \mathrm{V}\,(\bar{x}) + \lambda\ d\left(\sum_i n_i - n\right) \equiv 0$$

et comme :

$$\frac{1}{n_i^2}\, n_i\, \frac{N_i - n_i}{N_i - 1} \equiv \left(\frac{N_i}{n_i} - 1\right) \frac{1}{N_i - 1}$$

il vient :

$$-\frac{1}{n_i^2}\,\frac{N_i^2}{N^2}\,\frac{N_i}{N_i - 1}\,\sigma_i^2 + \lambda = 0 \qquad (i = 1, 2 \ldots L)$$

ou :

$$n_i = \sqrt{\frac{\lambda\, N_i}{N^2 (N_i - 1)}}\, N_i\, \sigma_i \doteq \frac{\sqrt{\lambda}}{N}\, N_i\, \sigma_i$$

Autrement dit : l'échantillon fournissant la *variance* la plus petite s'obtient en adoptant des n_i proportionnels aux $N_i\, \sigma_i$. En particulier, ce n'est pas l'*échantillon représentatif*, où les n_j sont proportionnels aux N_i. Sans doute, ce résultat perd-il beaucoup de sa valeur si l'on n'a pas, avant l'enquête, calculé déjà les σ_i (au moyen d'un rencensement antérieur), ce qui est un cercle vicieux. On peut en faire état cependant quand les éléments d'une certaine strate sont de natures très diverses et que l'on doit raisonnablement s'attendre à une *variance* beaucoup plus grande dans cette strate que dans les autres. C'est qu'il suffit, en effet, d'avoir une idée des ordres de grandeur respectifs des *variances* pour utiliser le résultat; autour de son minimum, la *variance* générale varie d'ailleurs lentement et on n'a d'intérêt réel à ne pas prendre un échantillon représentatif que lorsque les σ_i sont nettement différents les uns des autres.

II. — *Le problème classique du choix des communes.*

Une région est supposée divisée en communes (1 2 .. j ... L) dont les populations respectives sont $N_1\ N_2 \ldots N_j \ldots N_L$. On tire au sort ($l$) de ces communes et, à l'intérieur de celles-ci, des individus en nombre $n_1\ n_2 \ldots n_j \ldots n_L$. On recueille, par exemple, le salaire xj_k de ces individus.

a) Une estimation très naturelle du salaire moyen de la population est :

$$\bar{x} = \frac{1}{N} \cdot \frac{L}{l} \sum_j \frac{N_j}{n_j} \sum_k x_{jk}$$

avec :

$$N = \sum_j N_j$$

Supposons les communes tirées avec d'égales probabilités. On a :

$$\begin{aligned} E[\bar{x}] &= \frac{1}{N}\,\frac{L}{l}\, l \sum_j p\,(j)\,\frac{N_j}{n_j}\,\frac{n_j}{N_j} \sum_k x_{jk} \\ &= \frac{1}{N} \sum_j \sum_k x_{jk} \qquad\qquad \text{si } p\,(j) = \frac{1}{L} \\ &= E\,[x_{jk}]. \end{aligned}$$

De même :

$$V[\bar{x}] = \frac{L^2}{N^2 l^2} \left\{ l\, V\left[\sum_k x_{jk}\right] \frac{L - l}{L - 1} + \frac{l}{L} \sum_j \left(\frac{N_j}{n_j}\right)^2 n_j\, V_j[x_{jk}]\, \frac{N_j - n_j}{N_j - 1} \right\}$$

Telle serait donc la « meilleure » estimation (au sens précédent) si l'on

rend it minimum la *variance* V $[\bar{x}]$, lorsque les n_j varient, en choisissant les n_j optima, à savoir proportionnels à $N_j \sqrt{V_i [x_{jk}]}$.

b) Cependant, le bon sens fait sentir que pareille estimation ne peut être très bonne. On s'aperçoit alors que la première partie de la *variance* (où les n_j ne figurent pas) est très importante, à cause de V $[\sum_k x_{jk}]$; les variations que sont susceptibles de subir les $\sum_k x_{jk}$ sont considérables, du seul fait que le nombre d'habitants par commune est très variable dans une même région. On n'est certainement pas en possession de la meilleure formule d'estimation au sens propre.

On a avantage à choisir les communes qui contribueront à l'échantillon au moyen d'une loterie où chaque commune a N_j billets et où il y a l billets gagnants, et à adopter la formule :

$$\bar{x} = \frac{1}{l} \sum_j \frac{1}{n_j} \sum_k x_{jk}$$

avec :

$$\begin{aligned} E[\bar{x}] &= \frac{1}{l} \sum_j l \frac{N_j}{N} \left\{ \frac{1}{n_j} \cdot \frac{x_j}{N_j} \sum_k x_{jk} \right\} \\ &= E\,[x_{jk}] \end{aligned}$$

et avec

$$V[\bar{x}] = \frac{1}{l^2} \left\{ l\, V \left[\frac{1}{N_j} \sum_k x_{jk} \right] \frac{N - l}{N - 1} + l \sum_j \frac{N_j}{N} \cdot \frac{1}{n_j^2} n_j V_j [x_{jk}] \frac{N_j - n_j}{N_j - 1} \right\}$$

La détermination des *nj* optima peut s'effectuer comme précédemment, la condition d'*extremum* conduisant cette fois à des valeurs proportionnelles à $\sqrt{N_j V_j [x_{jk}]}$.

Mais on constate que la première partie de la *variance* renferme V $[E_j\, x_{jk}]$, *variance* d'une quantité qui n'est pas susceptible de grandes variations d'une commune à l'autre et devrait même rester très stable si la région est assez homogène. Au total, la *variance* sera beaucoup plus faible que dans le premier cas.

On a donc un gros intérêt à choisir les communes avec des probabilités proportionnelles à leur population et à adopter la meilleure formule d'estimation correspondante.

3e Partie. — L'enquête sur la basse-cour qui fut projetée en France en 1947.

1. Dans une région agricole donnée, on désignera :

— une exploitation agricole par les trois indices : (*i*) catégorie (très petite, petite, ... très grande); (*j*) commune; (*k*) exploitation à l'intérieur de la catégorie et de la commune;

— les nombres d'exploitations par N dans la région et *n* dans l'échantillon;

— le nombre de communes par L dans la région et dans l'échantillon;

— l'indice de sommation par E dans la région et S dans l'échantillon;

— la caractéristique étudiée (par exemple nombre de poules) par X ou x.

Dans la région, le nombre de poules, par exemple, s'écrira :

$$X = \Sigma \Sigma \Sigma x_{ijk}$$

Le problème théorique qui se pose est :

— de choisir pour X une formule d'estimation présentant certaines qualités;

— de préciser la manière de choisir les communes de l'échantillon;

— de fixer les nombres n d'exploitations à choisir.

Quant à la probabilité de choisir une entreprise de catégorie (i) dans la commune (j) si la commune (j) est choisie, elle est forcément la même pour toutes les exploitations (ij) quel que soit k.

2. Lorsqu'on suppose les communes de la région toutes extraites avec d'égales probabilités (1/L) la formule suivante d'estimation présente de grandes qualités :

$$x = \sum_i \frac{L}{l} S_j \frac{N_{ij}}{n_{ij}} S_k x_{ijk}.$$

a) Si l'on y remplace S par Σ, l par L, n par N, on obtient X; autrement dit x devient X si l'échantillon vient à englober toutes les exploitations de toutes les communes (les Anglo-Saxons disent que cette estimation est *consistante*);

b) On a E $[x]$ = X; autrement dit, l'estimation est sans *biais ;*

c) La *variance* de x se déduit de l'expression calculée au *II a*). ci-dessus par la relation :

$$x = \sum_i N_i \bar{x}_i$$

donc (aux $\rho_{ii'}$ près) :

$$V(x) = \Sigma N_i^2 V(\bar{x}_i)$$

ou

$$V(x) = \left\{\frac{L^2}{l}\frac{L-l}{L-1}\sum_i \sigma_i^2\right\} + \left\{\frac{L^2}{l}\sum_i\sum_j \frac{N_{ij}-n_{ij}}{N_{ij}-1}\frac{(N_{ij})^2}{n_{ij}}\sigma_{ij}^2\right\}$$

avec

$$\sigma_i^2 = V_j\left[\sum_k x_{ijk}\right]$$

$$\sigma_{ij}^2 = V_{ij}\left[x_{ijk}\right]$$

On peut rendre minimum la seconde partie de cette *variance* pour un jeu convenable de n_{ij}, avec les conditions :

$$\sum_i n_{ij} = 5 \qquad \text{(pour toutes les valeurs de } j\text{)}.$$

A cet effet, on écrit l'identité en n_{ij} :

$$d\,V(x) + \sum_i\sum_j \lambda_j\, d\, n_{ij} = 0$$

qui conduit à :

$$\frac{L}{l}\frac{N_{ij}}{N_{ij}-1}\cdot\left(\frac{N_{ij}\,\sigma_{ij}}{n_{ij}}\right)^2 = \lambda_j.$$

En confondant N_{ij} et $N_{ij} - 1$ (ce qui n'est pas d'ailleurs absolument correct pour les petites communes), on trouve donc que les n_{ij} doivent être proportionnels à $N_{ij}\,\sigma_{ij}$.

Mais, ainsi qu'on l'a déjà remarqué au *II a*) ci-dessus, la première partie de la *variance* est beaucoup plus importante que la seconde, en raison surtout des

grandes différences entre les nombres d'exploitations composant chaque commune. On a donc dû renoncer à adopter cette formule d'estimation.

3. On a vu (au *II b*) qu'il était tout indiqué de choisir les communes avec des probabilités inégales. Posons :

$$\sum_j N_{ij} = N_i \; ; \quad \sum_i N_{ij} = N_{oj}.$$

On peut facilement tirer au sort l communes, à une loterie où chacune a N_{oj} billets; mais on s'aperçoit que la « meilleure » formule d'estimation correspondante ne convient pas bien; ce serait :

$$x = \mathrm{S}_j \frac{1}{l} \frac{N}{N_{oj}} \sum_r \frac{N_{ij}}{n_{ij}} \mathrm{S}_k x_{ijk} = \frac{1}{l} \mathrm{S}_j \mathrm{S}_k \left[\frac{N}{N_{oj}} \sum_i \frac{N_{ij}}{n_{ij}} x_{ijk} \right]$$

avec

$$\mathrm{E}(x) = \sum_j \frac{N}{N_{oj}} \cdot \frac{N_{oj}}{N} \sum_i \sum_k x_{ijk} = \mathrm{X}.$$

Cette formule d'estimation tient sans doute compte des nombres variables d'exploitations par communes, mais ne tient pas compte des extrêmes différences de structure existant d'une commune à l'autre, à l'intérieur de la même région. Le dépouillement du fichier des exploitations montre, en effet, que telle commune n'a aucune très grande ni grande exploitation, alors que telle autre, qui est sa voisine, n'en aura aucune petite ou très petite. On comprend donc que la *variance* de x soit grande. Effectivement, on a :

$$\mathrm{E}_j \mathrm{S}_k \left[\frac{N}{N_{oj}} \sum_i \frac{N_{ij}}{n_{ij}} x_{ijk} \right] = \frac{N}{N_{oj}} \sum_k \sum_i x_{ijk}$$

$$\mathrm{V}_j \mathrm{S}_k \left[\frac{N}{N_{oj}} \sum_i \frac{N_{ij}}{n_{ij}} x_{ijk} \right] = \left(\frac{N}{N_{oj}} \right)^2 \sum_i \left(\frac{N_{ij}}{n_{ij}} \right)^2 n_{ij} \mathrm{V}_{ij}(x_{ijk}) \frac{N_{ij} - n_{ij}}{N_{ij} - 1}$$

donc

$$\mathrm{V}(x) \equiv \mathrm{V}[\mathrm{S}_j \mathrm{E}_j \mathrm{S}_k] + \mathrm{E}[\mathrm{S}_j \mathrm{V}_j \mathrm{S}_k]$$

$$= \mathrm{N}^2 l \frac{N - l}{N - 1} \mathrm{V} \left[\frac{\sum_i \sum_k x_{ijk}}{N_{oj}} \right] + \sum_j l \left(\frac{N}{N_{oj}} \right) \sum_i \frac{(N_{ij})^2}{n_{ij}} \mathrm{V}_{ij}(x_{ijk}) \frac{N_{ij} - n_{ij}}{N_{ij} - 1}.$$

On peut, comme précédemment, rendre minimum la deuxième partie de la *variance*, mais la première partie (qui ne renferme pas les n_{ij}) ne serait peu importante que si l'effectif moyen par exploitation $[\sum_i \sum_k x_{ijk} / \sum_i N_{ij}]$ variait peu d'une commune à l'autre. On n'est pas, en général, placé dans ce cas, en raison des variations de structure entre les communes d'une même région. On a donc une formule qui, tout en étant meilleure que la première, ne donne pas entièrement satisfaction.

4. Il est cependant possible *dans certaines régions agricoles*, d'avoir une bonne estimation en appliquant *séparément*, à chaque catégorie d'exploitations, les formules du paragraphe *II b*), à savoir ici :

$$\begin{cases} x_i = \dfrac{N_i}{l} \mathrm{S}_j \dfrac{\mathrm{I}}{n_{ij}} \mathrm{S}_k x_{ijk} \\ \mathrm{V}(x_i) = \dfrac{N_i^2}{l^2} \left\{ l \, \mathrm{V}_i \left[\dfrac{\sum_k x_{ijk}}{N_{ij}} \right] \dfrac{N_i - l}{N_i - 1} + l_j \sum \dfrac{N_{ij}}{N_i} \dfrac{1}{n_{ij}} \mathrm{V}_{ij}[x_{ijk}] \dfrac{N_{ij} - n_{ij}}{N - 1} \right\} \end{cases}$$

$$\begin{cases} x = \sum_i x_i \\ \mathrm{V}(x) = \sum \mathrm{V}(x_i) \end{cases} \qquad \text{(aux } \rho \textit{ it' près)}$$

Autrement dit, on tirera au sort (à la loterie, chaque commune ayant N*ij* billets) *l* communes qui contribueront *à l'échantillon de la catégorie i.* On recommencera une loterie différente pour chaque catégorie, sur les mêmes communes. Les communes tirées seront donc en général différentes suivant la catégorie.

Le premier inconvénient du procédé est sa complication; mais il ne faut surtout pas oublier que, aux nos 2 et 3, il faut déjà choisir 10 exploitations-échantillons par commune (5 + 5 supplémentaires pour pallier les exploitations disparues, inaccessibles ou les exploitants absents ou hostiles); un inconvénient des plus sérieux dans certaines régions est le trop grand nombre de communes de moins de 10 exploitations. Or, dans le cas présent, il est extrêmement rare qu'une région ne comprenne que des grandes communes ayant au moins 10 grandes, au moins 10 moyennes et au moins 10 petites exploitations (sans parler des très grandes et des très petites que, par commodité, on peut absorber dans les catégories voisines). Ainsi, les singularités se multiplieraient, il faudrait convenir pour les communes trop petites (les plus nombreuses) d'empiéter sur les communes voisines, ce qui est certainement plus facile à faire sur le terrain pour l'enquêteur que sur le fichier pour l'employé qui tirera les adresses. En résumé, le procédé ne vaut que pour certaines régions très rares où chaque commune a au moins 100 exploitations.

On en arrive ainsi, dans le cas général, à renoncer à une formule « sans *biais* » et à lui préférer une expression dont la *variance* soit petite et le « *biais* » encore plus petit.

Malheureusement, on en est réduit à conjecturer que tel sera bien le cas, attendu que les éléments positifs manquent pour calculer dès à présent écart-type et espérance mathématique de certaines distributions de la volaille dans une région (d'ailleurs il est peu vraisemblable que *chaque espèce* de volaille admette précisément la même meilleure formule).

Considérons d'abord la formule d'estimation :

$$x = \sum_i \frac{N_i}{n_i} \sum_j \sum_k x_{ijk}$$

Elle est manifestement consistante, mais son espérance mathématique n'est pas identique à X. D'ailleurs cette formule n'est pas d'un type étudié plus haut.

On a bien :

$$\begin{aligned} \mathrm{E} \sum_j \sum_k x_{ijk} &= \frac{n_{ij}}{N_{ij}} \sum_k x_{ijk} \\ &= n'_{ij} \overline{x_{ij}} \end{aligned}$$

Mais

$$\begin{aligned} \mathrm{E}\,[x] &= \sum N_i \, \mathrm{E} \left[\frac{1}{n_i} \sum_j n_{ij} \overline{x_{ij}} \right] \\ &= \sum_i N_i \, \mathrm{E} \left[\frac{\sum_j n_{ij} \overline{x_{ij}}}{\sum_j n_{ij}} \right] \end{aligned}$$

renferme une expression du type

$$\mathrm{E} \left[\frac{S\, u}{S\, v} \right]$$

ce qui nous écarterait des méthodes de calcul élémentaires décrites.

Posons $\sum_j n_{ij} = \nu_i$ (qui diffère de $S_j n_{ij} = n_i$). Une formule d'estimation exempte du défaut précédent est :

$$x = \sum_i \frac{L}{l} \frac{N_i}{\nu_i} S_j S_k x_{ijk}$$

avec

$$E(x) = \sum_i \frac{N_i}{\nu_i} \sum_j n_{ij} \overline{x_{ij}}$$

si les communes ont des chances égales d'être choisies.

On a, d'autre part,

$$x = \sum_i \sum_j N_{ij} \overline{x_{ij}}$$

de sorte que les deux expressions coïncideraient si coïncidaient les deux moyennes pondérées suivantes des $\overline{x_{ij}}$:

$$\frac{\sum_j N_{ij} \overline{x_{ij}}}{\sum N_{ij}} \qquad \text{et} \qquad \frac{\sum_j n_{ij} \overline{x_{ij}}}{\sum_i n_{ij}}.$$

Or, on sait que les x_{ij}, pour une même valeur de i (catégorie i), diffèrent peu dans une région agricole donnée, qui est sensée être très homogène. Ainsi le *biais* E (x) — X n'est pas très conséquent. Calculons la *variance* (aux $\rho_{i\,i'}$ près) :

$$V[x] = \sum_i \frac{L^2 N^2}{l^2 \nu_i^2} \left\{ l V_i \left[n_{ij} \overline{x_{ij}} \right] \frac{L - l}{L - 1} + \sum_j \left(\frac{N_{ij}}{n_{ij}} \right)^2 n_{ij} V_{ij} [x_{ijk}] \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right\}$$

On constate que la première partie de la variance est notablement plus petite que si $V\left[\sum_k x_{ijk}\right] = V\left[N_{ij} \overline{x_{ij}}\right]$ y figurait, comme c'était le cas au n° 2, car n_ij est en pratique 0, 1 ou 2, alors que Nij est susceptible de variations bien plus considérables.

Toutefois, la présence de ν_i et de $V[n_{ij} \overline{x_{ij}}]$ empêche une détermination simple des n_{ij} *optima*, rendant minimum la variance.

6. Aussi considérons une nouvelle formule d'estimation :

$$x = \sum_i \frac{N_i}{l} S_j \frac{1}{n_{ij}} S_k x_{ijk}$$

toujours dans le cas où les communes ont des probabilités égales d'être choisies. On a :

$$E(x) = \sum_i \frac{N_i}{L} \sum_j \frac{1}{N_{ij}} \sum_k x_{ijk}$$

$$= \sum_i \frac{\sum_j N_{ij}}{L} \sum_i \overline{x_{ij}}.$$

On aurait donc :

$$E(x) = X$$

si coïncidaient les moyennes simple et pondérée des x_{ij} :

$$\frac{\sum_j N_{ij} \overline{x_{ij}}}{\sum N_{ij}} \qquad \text{et} \qquad \frac{\sum_j \overline{x_{ij}}}{L}$$

Cette coïncidence est d'autant plus parfaite que les x_{ij}, pour i fixe, j variable, sont moins dispersés, c'est-à-dire que les catégories d'exploitations sont plus homogènes. Calculons la variance :

$$V(x) = \sum_i \frac{(N_i)^2}{l^2} \left\{ l\, V_i[\overline{x_{ij}}] \frac{L-l}{L-1} + \sum_j \frac{l}{L} n_{ij} V_{ij} \left[\frac{x_{ijk}}{n_{ij}}\right] \frac{N_{ij}-n_{ij}}{N_{ij}-1} \right\}$$

On voit qu'elle comprend encore deux termes :

— le premier, indépendant des nij, renferme $V_i[\overline{x_{ij}}]$ expression d'autant plus petite que les catégories d'exploitations sont plus *homogènes*.

— le second est minimum, compte tenu des conditions :

$$\sum_i n_{ij} = 5$$

lorsque l'on a :

$$d\,V(x) + \sum_j \lambda_i \sum_j d\cdot n_{ij} = 0$$

c'est-à-dire que n_{ij} est proportionnel à $\sqrt{V_{ij}[x_{ijk}]}$.

7. Dans le cas où les communes ont d'égales probabilités d'être choisies, on vient d'examiner trois formules d'estimation; la partie de la variance indépendante des n_{ij} représente les écarts suivants :

(2) $$\frac{L}{\sqrt{l}} \sqrt{V[\overline{x_{ij}}\, N_{ij}]}$$

(5) $$\frac{1}{\sqrt{l}} \sqrt{V[\overline{x_{ik}}\, n_{ij}]} \cdot \frac{L\, N_i}{\nu_i}$$

(6) $$\frac{N_i}{\sqrt{l}} \sqrt{V[\overline{x_{ij}}]}$$

Naturellement, compte tenu des relations $E(x) = X$, à savoir :

(5) $$\frac{N_i}{\nu_i} n_{ij} = N_{ij}$$

(6) $$\frac{N_i}{L\, N_{ij}} = 1$$

ces écarts sont tous trois équivalents. Mais ceux qui varient le moins sont les derniers et ceci sans aucune hésitation possible.

En conséquence, on a décidé d'adopter pour formule d'estimation la dernière formule :

$$x = \sum_i \frac{N_i}{l} S_j \frac{1}{n_{ij}} S_k x_{ijk}$$

qui sera d'autant meilleure que les *régions* et les *catégories* seront plus *homogènes*, c'est-à-dire que les exploitations de même catégorie dans la même région seront plus interchangeables quelle que soit leur commune.

Quant aux *variances* elles-mêmes, il est clair qu'on ne les connaît pas, et on peut seulement tabler sur le fait qu'elles grandissent quand on passe des petites aux moyennes et des moyennes aux grandes exploitations. On peut penser que les écarts-types sont entre eux à peu près dans le même rapport que les superficies moyennes de chaque catégorie d'exploitations; mais ce

n'est là qu'un ordre de grandeur très approximatif ; on s'est contenté d'accroître la proportion des grandes et très grandes exploitations dans l'échantillon, aux dépens des petites et très petites, sans l'exagération qui aurait consisté à faire à peu près disparaître ces dernières de l'échantillon.

8. Considérons à présent un ensemble de régions d'indice (h) et cherchons à répartir au mieux entre ces régions un nombre fixe a (5.000) de communes. On placera un indice h devant les indices ($i\ j\ k$) de toutes les expressions précédentes. La caractéristique de l'ensemble est :

$$x = \sum_h x_h$$

avec, par exemple :

$$x_h = \sum_i \frac{N_{hi}}{l_h} S_j \frac{1}{n_{hij}} S_k x_{hijk}$$

On a :

$$V[x] = \sum_h V[x_h]$$

$$= \sum_h \frac{1}{l_h} \frac{L_h - l_h}{L_h - 1} \sum_i N_{hi}^2 V_{hi}[x_{hij}]$$

$$+ \sum_h \frac{1}{l_h L_h} \sum_i \sum_j \frac{N^2_{hi}}{n_{hij}} \frac{N_{hij} - n_{hij}}{N_{hi} - 1} V_{hij}[x_{hijk}]$$

et

$$\sum_h l_h = a$$

Il s'agit de déterminer les valeurs optima des l_h

Les conditions d'*extremum* lié s'écrivent :

$$\frac{\partial}{\partial l_h} V[x] + b \frac{\partial}{\partial l_h} \left[\sum_h l_h - a\right] = 0$$

de la forme (en confondant L_h et $L_h - 1$) :

$$-\frac{1}{l_h^2}\left(A_h + \frac{B_h}{L_h}\right) + b = 0$$

ou, en négligeant le deuxième terme de la parenthèse :

$$l_h = \sqrt{\frac{A_h}{b}}$$

avec

$$A_h = \Sigma_i N_{hi}^2 V_{hi}[x_{hij}]$$

Il est permis, en première approximation, d'admettre que l'écart-type $\sqrt{V_{hi}[x_{hij}]}$ varie proportionnellement à la moyenne x_{hi} de sorte que $N_{hi}\sqrt{V_{hi}[x_{hij}]}$ varie, en très gros, proportionnellement à $\sum_j \sum_k x_{hijk}$. Or $\sqrt{A_h}$ est la moyenne (quadratique) des expressions précédentes (au facteur 1/5 près, s'il y a 5 catégories 1, 2, .. i .. 5). Au total, les l_h ont un système de valeurs optima, assez grossièrement proportionnelles à $\sum_i \sum_j \sum_k x_{ijk}$.

Comme une seule enquête devait servir à déterminer les effectifs de lapins et de chaque espèce de volaille, les nombres d'œufs pondus et d'animaux sacrifiés, on s'était contenté de rendre, par département, le nombre de

communes où aurait lieu l'enquête à peu près proportionnel au poids total des animaux de basse-cour en 1929. C'est ainsi que les résultats de cette vieille enquête devaient finalement servir; on avait pourtant rejeté le chiffre relatif aux Bouches-du-Rhône qui a paru exagéré. On n'avait pas jugé nécessaire non plus de serrer de trop près ces chiffres de 1929 tant du fait de leur incertitude que de leur éloignement; on s'était contenté d'adopter un quotient d'échantillonnage *simple* (100, 200, 300, 150, 50) pour passer du nombre total d'exploitations du département à celui de l'échantillon (l'Ain comportait deux quotients distincts, celui de la Bresse étant moindre que l'autre). Le plus grand nombre des quotients était égal à 100 qui était d'ailleurs le quotient moyen de la France entière (1).

Indiquons, pour terminer qu'on n'escomptait pas obtenir une estimation détaillée par département, mais des effectifs globaux pour la France entière; cependant, on devait rechercher, le moment venu, si certains résultats, par grandes régions, avaient assez de précision pour être publiés.

4e Partie. — Indications générales sur l'avenir des enquêtes agricoles par sondage en France.

Ce premier essai en vue d'appliquer à un cas complexe des résultats théoriques dont chacun est simple, permet de dégager quelques enseignements pratiques. La France doit effectuer un recensement agricole en 1950, à la suite duquel un nouveau fichier d'exploitations sera constitué, qu'on substituera à celui de 1942. Ce fichier doit être prévu dès à présent, constitué de manière à se prêter facilement aux futures enquêtes sur échantillon pris au hasard, enquêtes agricoles ou enquêtes d'économie rurale, voire enquêtes démographiques.

Il y aura lieu pour les statisticiens de veiller à ce que les questionnaires soient remplis avec autant de soin que ceux antérieurs à 1943 et qu'aucune exploitation importante ne soit oubliée. Et il faudra également se pencher sur le problème très délicat des limites inférieures à fixer, en dessous desquelles il n'y a plus lieu defaire une fiche d'exploitation agricole.

Toutefois, les problèmes qui attirent plus spécialement l'attention du « sondeur » sont les suivants :

D'une part, la section de commune à laquelle appartiennent les bâtiments principaux de l'exploitation, le hameau, le lieudit où ils se trouvent devraient figurer sur les documents de base, ceux-ci pourraient ainsi être regroupés à l'intérieur des communes (sauf pour les plus petites communes); un découpage du territoire en *unités* assez homogènes deviendrait possible, les communes comprenant un nombre entier d'*unités*; aux États-Unis, de telles *unités* contiennent au maximum 10 logements, dont 4 ou 5 exploitations agricoles; en France, il semble que le hameau offrirait, dans bien des régions, les mêmes avantages, à savoir possibilité d'accroître l'échantillon à peu de frais en inter-

(1) Ajoutons qu'en juillet 1947, on avait préparé en outre une enquête sur les salaires agricoles qu'une partie des mêmes enquêteurs auraient effectuée simultanément sur un échantillon plus réduit, le nombre de communes où elle aurait eu lieu étant à peu près proportionnel aux effectifs de main-d'œuvre agricole dans chaque département.

rogeant à la fois tous les gens de l'*unité ;* la théorie de l'échantillon en grappe, dont il n'a pas été question ici, montre l'intérêt substantiel que l'on a souvent à procéder ainsi.

Ainsi, l'unité d'échantillonnage ne serait plus l'exploitation agricole, qui, d'ailleurs, ne peut plus jouer ce rôle dès qu'on quitte le domaine de l'enquête purement agricole, ce serait le hameau.

D'autre part, le rapprochement (au moyen des résultats du recensement) des *unités* similaires permettrait peut-être de constituer des « *strates* » sensiblement plus homogènes que ne le sont les régions agricoles actuelles, désignées assez empiriquement. Pour la publication des résultats d'une enquête par sondage, il paraît, en outre, nécessaire de pouvoir partager le territoire en un petit nombre de grandes régions ou plutôt de grandes zones (avec des solutions de continuité et des enclaves) vers le choix rationnel desquelles le recensement pourra nous guider. Enfin, la nécessité de distinguer des unités secondaires, composées d'unités primaires groupées en sous-strates, s'imposera sans doute ; et on pense pouvoir choisir ces unités rationnellement au lieu des unités empiriques et artificielles que constituent les communes.

P. Thionet.