REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Un problème de Tore Dalenius sur les sondages

Revue de statistique appliquée, tome 14, n° 4 (1966), p. 45-67

<http://www.numdam.org/item?id=RSA_1966__14_4_45_0>

UN PROBLÈME DE TORE DALENIUS SUR LES SONDAGES

P. THIONET
Faculté des Sciences de Poitiers

INTRODUCTION

1 - LE PROBLEME TEL QU'IL EST POSE

Nous nous référons d'un document de DALENIUS intitulé : Potential research objects in sample survey theory and methods. (Sujets de recherches en puissance, concernant la théorie et les méthodes de sondage), communication à un colloque tenu à Budapest 18-22 juin 1963 (Colloquium on Applications of Mathematics in economics).

Le document donne une liste de sujets, faisant (en principe) l'objet des recherches de DALENIUS et de ses élèves au Centre de la Recherche Scientifique de Suède (Swedish Social Science Research Council).

Nous allons en étudier le point 7.1. Estimating restricted parameters que nous citons d'abord.

"Dans de nombreuses applications, on est pratiquement certain que le paramètre satisfait par exemple à une inégalité $P_b \leqslant P \leqslant P_h$. Pour rester concret, envisageons l'exemple suivant. Un sondage doit être effectué en Suède pour estimer la proportion P de chômeurs dans la population active. Quiconque connaît la situation présente du marché du travail en Suède est absolument certain que P ne dépasse pas 10 % ou même 5 % ; c'est-a-dire qu'il est pratiquement certain que P satisfait à l'inégalité $0 \leqslant P \leqslant 0,10$.

Il semble parfaitement raisonnable de supposer qu'il "doit" y avoir moyen d'utiliser ce fait pour établir un plan de sondage plus efficace que celui qu'on établirait autrement. Pourtant la littérature statistique ne renferme sur ce sujet qu'extrêmement peu de choses"(1).

Nous avons obtenu de l'auteur confirmation verbale qu'il se refusait à employer une estimation Bayesienne du paramètre P, à partir d'une hypothèse qui ferait de P une variable aléatoire dont la loi de probabilité est donnée a priori, le sondage survenant seulement pour remplacer la loi a priori par une loi de probabilité a posteriori de P. Pour DALENIUS, P est un paramètre, inconnu mais non aléatoire ; c'est

---

(1) "Bien entendu les partisans de la définition subjective des probabilités n'auront pas le même point de vue".

dit-il, le point de vue de NEYMAN. Il ne semble pas croire que la méthode de Bayes Laplace apporte une solution valable au problème qu'il se posait ; en tous cas ce n'est pas cette solution là (dépendant d'une loi a priori arbitraire) qui l'intéresse.

Nous voici donc parfaitement informé sur le sujet de recherches 7.1, sinon sur les résultats auxquels ces recherches conduiront M. DALENIUS. Nous allons ici indiquer quelles réflexions ce sujet nous a suggérées.

2 - ETAT PRESENT DES METHODES

Nous faisons abstraction des techniques destinées à éviter les erreurs de définition du chômeur, erreurs susceptibles de fausser de 100 pour cent et plus toute évaluation). D'ores et déjà tout organisateur de sondage sur l'emploi (en Suède ou ailleurs) s'il sait que le chômage est rare, n'essaiera pas de le déterminer par sondage. Il est à conseiller que le chômage soit cerné par enquête exhaustive. Si ces enquêtes sont (par mesure d'économie) trop espacées (par exemple si les chômeurs ne sont recensés qu'au Recensement Général de population), il semble que le sondage doive plutôt servir à évaluer dans quelle proportion le chômage a augmenté ou diminué depuis le dernier recensement (lequel constitue l'information supplémentaire indispensable pour valoriser le sondage).

Comme le chômage paraît lié à l'implantation géographique, on enquêtera dans un échantillon de communes. Il est clair qu'une sous-stratification de la population suivant l'âge et la profession serait efficace. Mais il est vraisemblable qu'on y renoncera, car on tirera un échantillon de logements et non pas un échantillon d'individus - l'existence d'une base de sondage adéquate étant la justification de ce choix.

En résumé le seul fait de savoir qu'on doit faire une enquête pour rechercher des cas rares conditionne les plans de sondage d'une façon classique qui, semble-t-il, ne laisse guère de place à de nouvelles recherches.

3 - NATURE TROP PARTICULIERE DU RENSEIGNEMENT $0 \leqslant P \leqslant 0,10$

L'exemple donné par DALENIUS paraît d'ailleurs mal choisi pour une autre raison. Alors que jusqu'ici on n'a pas trouvé de technique propre pour utiliser pleinement une information telle que $0 \leqslant \overline{X} \leqslant A$, $\overline{X}$ désignant la moyenne (à estimer) des valeurs x_i prises sur les unités de sondage i par la variable x étudiée, on sait très bien en revanche utiliser une information telle que :

$$0 < \sigma < B$$

σ désignant l'écart-type des mêmes x_i : On en déduit en effet que, pour un échantillon de n unités tirées à la manière des boules d'une urne, l'erreur $|\overline{x} - \overline{X}|$ d'échantillonnage dépassera moins d'une fois sur 20 le nombre $2B/\sqrt{n}$.

Or si l'on considère une variable x, égale à 0 ou 1, dans les proportions P et (1 - P), sa moyenne est P et son écart-type $\sqrt{P(1-P)} = \sigma$.

La fonction $\sigma(P)$ est croissante de 0 à 0,5 et décroissante de 0,5 à 1. Ainsi le renseignement relatif à $\overline{X} = P$

$$0 < P < 0.10$$

se traduit par :

$$0 < P(I - P) < 0,09$$

ou

$$0 < \sigma < 0.3$$

Ainsi l'erreur d'échantillonnage dépassera moins d'une fois sur 20 le nombre :

$$0,6/\sqrt{n}$$

Il va de soi que DALENIUS n'avait pas en vue un résultat aussi connu et banal quand il écrivait son point 7.1. Il songeait naturellement à une égalité telle que $0 < \overline{\overline{X}} < A$. C'est du moins l'inégalité dont nous nous occuperons dans ce qui suit, et (on va le voir) elle laisse encore place à diverses interprétations.

4 - LES DIVERSES SIGNIFICATIONS DE L'INEGALITE $\boxed{0 < \overline{X} < A}$

a) La première confusion qui risque de se produire au sujet de cette inégalité est qu'au lieu d'y voir une information supplémentaire destinée à améliorer le sondage, on peut l'interpréter comme une condition supplémentaire imposée à l'estimation $\hat{x}$ de $\overline{X}$ déduite du sondage.

Car si nous tirons au sort, par exemple un échantillon de n unités x_i, il n'y a aucune raison pour que leur moyenne $\bar{x}$ soit inférieure à A, même lorsqu'on sait par ailleurs la moyenne générale $\overline{X}$ inférieure à A. Du moment qu'il existe suffisamment d'unités x_i supérieures à A, la probabilité d'obtenir $\bar{x} > A$ n'est pas nulle.

C'est alors un bouleversement des méthodes usuelles d'estimation qui s'impose, dans l'optique : condition supplémentaire. Nous ne nous en occuperons pas ici et ce n'est pas d'ailleurs un problème neuf[1].

b) Si nous avons obtenu ($\bar{x} > A$), l'utilisation de l'information ($X < A$) dépendra du but recherché.

- Ou bien nous voulons essayer d'abord de savoir pourquoi on a obtenu

$$\bar{x} > A$$

et employer cette information pour juger la qualité de l'échantillon.

---

(1) Voir par exemple :

- BRUNK (H.D.) On the estimation of parameters restricted by inequalities, Ann. Math. Stat. 29-2 june 1958 p. 437-454.

- SKIBINSKY (M.) & COTE (L.) On the inadmissibility of some standard estimate in the presence of prior information. Ann. Math. Stat. 34-2 June 1963 p. 537-548.

- Ou bien nous désirons avant tout des règles de calcul susceptibles de nous fournir des estimations utiles dans la majorité des cas. Mais alors les moyens mathématiques à employer ne sont pas indépendants des causes que nous attribuons au résultat $x > A$. Et les règles de calcul dépendent des buts que nous poursuivons.

1e PARTIE. JUGEMENT SUR LA QUALITE DE L'ECHANTILLON

1 - INTRODUCTION

L'information $\overline{X} < A$, sachant $A < B = \overline{x}$ conduit à se poser la question : l'échantillon est-il bon ?

Supposons l'échantillon tiré au sort à la manière des boules d'une urne (avec ou sans remise) ; soit n la taille de cet échantillon. L'échantillon nous fournit d'autres renseignements que la valeur de $\overline{x}$: L'histogramme des x_i donne une idée de la forme de la distribution des $\{x\}$, et d'autant mieux que la taille n de l'échantillon est grande.

Notamment on sait estimer au moyen des x_i l'écart type σ de la distribution. Soit s^2 la variance des x_i échantillons, on sait que :

$$E\left(\frac{ns^2}{n-1}\right) = \frac{N\sigma^2}{N-1}$$

ns^2 serait un $\boxed{\sigma^2\chi^2}$ à $(n-1)$ degrés de liberté si la distribution des $\{x\}$ était gaussienne.

2 - ETUDE D'UN CAS EXTREME

Commençons par nous placer dans un cas, non réaliste mais qui présente l'avantage d'être très simple et de donner une idée de ce qui se passe en fait : supposons connue la variance σ^2 de la distribution des x_i ; et supposons en outre n très grand.

On sait qu'alors la variable $(\overline{x} - \overline{X})\sqrt{n}/\sigma = t$ est à la limite une variable de Laplace-Gauss réduite. Posons :

$\delta = A - X$: cette marge (entre X et sa borne) est inconnue. L'échantillon connu nous donne seulement $B(= \overline{x})$. Donc $(B - A)\sqrt{n}/\sigma$ est connu :

$$B - \overline{X} = B - A + A - \overline{X} = B - A + \delta$$

La Table suivante nous montre comment varie $(B - A)\sqrt{n}/\sigma$ en fonction de $\delta\sqrt{n}/\sigma$ et de P_o, avec :

$$P_o = \text{Probabilité } (\overline{x} > B)$$

Valeur de $\frac{(B - A)\sqrt{n}}{\sigma}$ en fonction de $\frac{\delta\sqrt{n}}{\sigma}$ et du niveau de probabilité P_o

Probabilité \ $\frac{\delta\sqrt{n}}{\sigma}$	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0	2,1	2,2	2,3	2,4
0,0968	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0											
0,0808	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0										
0,0668	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0									
0,0548	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0								
0,0446	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0							
0,0359	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0						
0,0287	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0					
0,0228	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0				
0,0179	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0			
0,0139	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0		
0,0107	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0	
0,0092	2,4	2,3	2,2	2,1	2,0	1,9	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,0

On peut lire cette même table en diagonale : elle indique alors comment pour une valeur donnée de $(B - A)\sqrt{n}/\sigma$ la Probabilité P_o varie en fonction de $\delta\sqrt{n}/\sigma$

Exemple : $n = 2.500 \Longrightarrow \sqrt{n} = 50$;

$A = 750$ Francs ; $B = 760$ Francs

$\sigma = 400$ Francs

$$\frac{(B - A)\sqrt{n}}{\sigma} = \frac{10 \times 50}{400} = 1,25$$

a) Nous consultons les diagonales 1,2 et 1,3 de la Table.

1,2 : Quand $\frac{\delta\sqrt{n}}{\sigma}$ varie de 0 à 1,2, P_o varie de 0,1151 à 0,0092

1,3 : " " " 0 à 1,1, P_o varie de 0,0968 à 0,0092

b) Traduisons le résultat ; la probabilité que $\bar{x}$ dépasse 760 Francs est égale à 10 % si $\delta = 0$, c'est-à-dire si $\bar{X} = 750$; cette probabilité n'est plus que de 1 %,

si $$\delta = 1,15\ \sigma/\sqrt{n},$$

c'est-à-dire : $$\delta = 9 \ ; \ \bar{X} = 741.$$

Si nous commençons à tenir pour "significatif" un écart correspondant à $P_o < 5$ %, il lui correspond (interpolation) :

$$\delta \geq 0,45\ \sigma/\sqrt{n} = 3,6$$

Remarque : Lorsqu'on n'aura aucune opinion sur δ (sinon que $\delta \geqslant 0$) on devra raisonner dans le cas le plus défavorable, c'est-à-dire supposer $\delta \geqslant 0$. Les formules suivantes offrent sensiblement le même intérêt que le tableau :

$$\text{Prob}\left(\delta + B - A \geqslant 1,64 \frac{\sigma}{\sqrt{n}}\right) = 5\ \%$$

$$\text{Prob}\left(\delta + B - A \geqslant 2,05 \frac{\sigma}{\sqrt{n}}\right) = 2\ \%$$

$$\text{Prob}\left(\delta + B - A \geqslant 2,33 \frac{\sigma}{\sqrt{n}}\right) = 1\ \%$$

3 - CAS OU n EST GRAND ET OU σ N'EST PAS CONNU (TIRAGE DANS UNE POPULATION FINIE)

On suppose que la distribution des x_i n'est pas Gaussienne.

Dans la pratique des sondages on conserve n grand. Il s'en suit que :

- la moyenne $\bar{x}$ suit toujours une loi de Laplace-Gauss (à la limite) ;

- la moyenne $\frac{2ns^2}{n-1}$ des $(x_i - x_j)^2$ échantillons suit aussi une loi de Laplace-Gauss-limite, quelle que soit la distribution des $(x_i - x_j)^2$ de la population.

- ces 2 lois-limite sont indépendantes,

- les variables s et $(\bar{x} - \bar{X})\sqrt{n}$ suivent des lois de Gauss limite; leur rapport R en suit une également, et on peut voir qu'il s'agit d'une loi normée.

$$R = \frac{(\bar{x} - \bar{X})\sqrt{n}}{s} : \quad \mathcal{E}R \longrightarrow 0 \quad \mathcal{V}R \longrightarrow 1$$

Ainsi : la distribution de probabilité limite (n très grand) de R coïncide avec celle donnée au n° 2 (loi de Laplace Gauss normée).

Remarque 1 - Si l'échantillon provient d'une loi théorique, il est possible qu'on puisse estimer $\bar{x}$ et σ par des estimateurs meilleurs que $\bar{x}$ et s.

Remarque 2 - Si l'échantillon provient d'une loi de distribution théorique, il n'y a aucune raison pour qu'on obtienne toujours le résultat ci-dessus . Le calcul complet fait intervenir le moment d'ordre 4 de x ; on aura et donc des difficultés si le moment d'ordre 4 est infini (a fortiori si c'est le moment d'ordre 2, comme avec une loi de Pareto, laquelle n'a pourtant rien d'exceptionnel pour un statisticien). Bien entendu les échantillons d'un sondage proviennent toujours d'une population finie et aucun de leurs moments ne peut donc être infini en soi ; mais si cette population a l'allure d'un superéchantillon d'une loi de PARETO, on peut penser que R restera très éloigné de sa loi-limite.

4 - CAS OU n N'EST PAS TRES GRAND MAIS OU LA DISTRIBUTION DES x_i EST DE LAPLACE GAUSS

Lorsque les x_i suivent une loi de Gauss, la loi suivie par leur moyenne x est une loi de Gauss quelque soit n (et non plus pour n très grand).

La loi suivie par $(\bar{x} - \bar{X})\sqrt{n}/s = t$ est connue sous le nom de loi de Student Fisher.

Exemple	$P_o \leqslant 5\ \%$	$P_o \leqslant 1\ \%$
	$n = 20,\ t \geqslant 1.73$	$t \geqslant 2.54$
	$n = 30,\ t > 1.70$	$t \geqslant 2.46$
	$n = 60,\ t \geqslant 1.67$	$t \geqslant 2.39$
au lieu de	$t \geqslant 1.64$	$t \geqslant 2.33$

Remarque

1/ Lorsque les x_i suivent une distribution très différente de la loi de Gauss, - par exemple la loi de Pareto. - la loi de $(\bar{x} - \bar{X})\sqrt{n}$ n'a plus aucune ressemblance avec celle qui précède.

5 - CAS D'UN PLAN DE SONDAGE USUEL

Les calculs précédents supposent un échantillon sans strate et à un seul degré. En pratique on ne procède jamais ainsi ; et il y a lieu de substituer à s^2, dans la formule du n° 2, la variance d'échantillonnage habituelle (dans la mesure où on sait l'évaluer valablement).

6 - PROBLEME : Que fera-t-on si l'écart $(\bar{x} - A)$ paraît "significatif" ?

Si $(\bar{x} - A)$ paraît "significatif", c'est que :

- ou bien l'erreur d'échantillonnage a une taille exceptionnelle, comme cela peut arriver naturellement 1 fois sur 100, 1 fois sur 1000, etc ;

- ou bien la variance d'échantillonnage estimée est très différente de la variance théorique,

- ou bien cette variance ne correspond pas au plan d'échantillonnage réel mais à un plan théorique dont les exécutants se sont éloignés, volontairement ou non : le plan théorique pouvait par exemple être bon en théorie mais inexécutable, c'est-à-dire donner disons 50 % et plus de non-réponses pour certaines unités de sondage ; l'échantillon réel est alors fortement déformé par la répartition inégale des non-réponses.

- il peut encore se faire que l'écart excessif $(\bar{x} - A)$ soit du davantage aux erreurs de mesure et d'observation qu'aux erreurs d'échantillonnage,

- enfin plusieurs causes peuvent agir simultanément, soit dans le même sens d'où $\bar{x} - A > 0$ soit en sens contraire (d'où $\bar{x} - A < 0$ malgré une forte variance).

Il peut être puéril de vouloir "disculper" telle cause d'erreurs aux dépens de telle autre - alors qu'on manque d'informations exogènes pour y voir clair.

Un point essentiel est de savoir si, quand on écrit les 2 relations :

$$\bar{X} < A, \quad A < \bar{x}$$

$\bar{x}$ est bien comparable à $\bar{X}$; c'est-à-dire si les données échantillons x utilisées pour calculer $\bar{x}$ figureraient aussi, telles-quelles, dans $\bar{X}$ au cas où l'on pourrait calculer $\bar{X}$.

Par exemple lors d'un sondage sur les conditions de vie des personnes âgées (1948) nous avions trouvé que les nombres de personnes âgées, percevant l'allocation aux économiquement faibles ou la retraite des vieux travailleurs, étaient sensiblement concordants, qu'on l'ait estimé sur échantillon, ou qu'il soit fourni par le budget de l'Etat. Mais si l'on considérait séparément les 2 postes : Vieux travailleurs - Economiquement faibles, le désaccord entre le sondage et le budget était flagrant. On pense que beaucoup d'enquêtés avaient confondu de bonne foi les deux expressions (surtout les gens les plus âgés) ; et il aurait fallu pouvoir leur demander le montant du dernier mandat pour vérifier l'exactitude de leur déclaration. (On rappelle qu'il s'agit de deux prestations dont le cumul par une même personne est interdit).

Parmi toutes les causes d'erreur, on a surtout tendance à penser à l'abondance des refus ou non-réponses (dans certaines classes sociales, etc) apportant à l'échantillon des déformations qu'on cherche à redresser.

La pratique conduit alors à découper a posteriori la population en classes, et à faire comme si l'échantillon était bien tiré au sort avec d'égales probabilités à l'intérieur de chacune de ces classes. Cette hypothèse est bien entendu commode mais gratuite.

J'ajoute, par expérience, que lorsqu'un échantillon est très mauvais, le redressement d'échantillon est totalement inefficace.

Le même redressement appliqué à un échantillon peu déformé, est à peu près inoffensif et d'un effet nul. Restent les cas intermédiaires, qui d'ailleurs sont les plus fréquents.

2e PARTIE : INFORMATION SUPPLEMENTAIRE EN VUE DE L'ESTIMATION

POSITION DU PROBLEME

Par hypothèse on a :

$$\overline{Y} \leqslant A, \quad \overline{Z} \leqslant B \qquad (I)$$

où A, B sont connus (pour certaines variables de contrôle).

Un sondage conduit à des estimations sur échantillon :

$$\overline{x} = \text{est. } \overline{X}, \quad \overline{y} = \text{est. } \overline{y}, \quad \overline{z} = \text{est. } Z$$

Comment utiliser (I) pour améliorer ces résultats ? c'est-à-dire : pour remplacer $\overline{x}$ par x^* qui soit une meilleure estimation de $\overline{X}$?

La théorie classique des sondages apprend à utiliser des informations supplémentaire à l'un des trois stades que voici(1) :

A) soit lorsque le sondage est terminé, les informations s'introduisant dans l'estimateur ou la formule d'estimation $x^*(x_1\ x_2 \ldots x_n)$ substituée à $\overline{x}(x\ x_2 \ldots x_n)$;

B) soit au cours du sondage, les informations contribuant à la désignation d'un échantillon meilleur que celui qu'on obtiendrait si on en était privé ;

C) soit avant le sondage, les informations étant incorporées dans les calculs du plan de sondage.

On peut de même se proposer de chercher à utiliser des informations du type nouveau :

$$\overline{Y} \leqslant A, \quad \overline{Z} \leqslant B$$

---

(1) Toute l'information disponible peut être utilisée à un seul des stades, de sorte qu'on ne gagne rien à intervenir simultanément à plusieurs stades.

à l'un des stades (A) (B) (C). On va se limiter d'ailleurs au cas d'une seule inégalité.

Il va de soi qu'une information se présentant sous la forme d'inégalités ne peut être aussi riche, aussi "informative" que s'il s'agissait d'égalités.

A - UTILISATION A POSTERIORI DE L'INFORMATION. ESTIMATEUR PAR RATIO OU REGRESSION

1/ Procédons par analogie. Si l'on sait la condition suivante vérifiée :

$$\overline{Y} = A$$

on peut substituer à l'estimation $\bar{x}$ de $\overline{X}$, l'estimation par ratio

$$x^* = \overline{Y} \frac{\bar{x}}{\bar{y}} = \frac{\overline{Y}}{\bar{y}} \bar{x}$$

sachant que pour l'échantillon la moyenne $\bar{y}$ est en général différente de A.

Un autre procédé classique (moins employé parce que plus compliqué) consiste à substituer A à $\overline{Y}$ dans une formule de régression.

$$\bar{x} - \overline{X} = \beta (\bar{y} - \overline{Y})$$

Si le coefficient β de régression est estimé sur l'échantillon qui fournit aussi $\bar{x}$ et $\bar{y}$, il vient :

$$x^* = \text{est. } \overline{X} = \bar{x} + (\text{est. } \beta)\ (A - \bar{y})$$

Ces calculs sont intéressants si la variable à l'étude x est fortement corrélée avec la variable connue (ou variable de contrôle) y.

Faisons l'hypothèse de structure (dite de régression linéaire forte) qu'il existe des constantes $\alpha\ \beta$ telles que :

$$x = \alpha + \beta\ y + \varepsilon, \quad E\varepsilon = 0$$

Sans être entièrement nécessaire, cette hypothèse a le mérite de la clarté. Elle implique :

$$\overline{X} = \alpha + \beta\ \overline{Y}$$

$$\bar{x} = \alpha + \beta\ \bar{y} + \bar{\varepsilon}$$

d'où $$\overline{X} = \bar{x} + \beta(\overline{Y} - \bar{y}) - \bar{\varepsilon}$$

Posons $$\text{est } \overline{X} = \bar{x} + (\text{est } \beta)\ (\overline{Y} - \bar{y})$$

$$\text{est } \overline{X} - \overline{X} = (\text{est}\,\beta - \beta)\ (\overline{Y} - \bar{y}) + \bar{\varepsilon}$$

L'estimation par ratio s'en déduit comme cas particulier si l'on a : $\alpha = 0$; $\beta = \overline{X}/\overline{Y}$; est $\beta = \bar{x}/\bar{y}$

2/ Supposons à présent que $\overline{Y}$ ne soit plus exactement connue, mais qu'on sache $\overline{Y} \leqslant A$ (où A est positif)

- Si l'échantillon fournit $\bar{y} \leqslant A$, on en restera là et on aura :

$$\text{est } \bar{X} = \bar{x}$$

- Si au contraire on a $A < \bar{y}$, on doit penser que $\bar{x}$ est, lui aussi, trop grand (à supposer que β soit positif).

On aura intérêt à corriger $\bar{x}$ en écrivant

non plus $\quad \text{est. } \bar{X} = \bar{x} - (\text{est. } \beta).\ (\bar{y} - \bar{Y})$

mais $\quad \text{est. } \bar{X} < \bar{x} - (\text{est. } \beta).\ (\bar{y} - A)$

La pente β de la droite de régression sera évaluée sur l'échantillon des $(x_i\ y_i)$.

Cette formule de régression peut être remplacée par la formule de ratio suivante :

$$\text{est } \bar{X} < A\frac{\bar{x}}{\bar{y}}$$

Si ce mode d'estimation par une inéquation au lieu d'une équation paraît choquante, il faudra introduire une variable supplémentaire δ (inconnue)

$$\bar{Y} = A - \delta, \quad \boxed{\delta \geqslant 0}$$

$$\Longrightarrow \begin{cases} \text{est } \bar{X} = \bar{x} - (\text{est. } \beta)\ (\bar{y} - A - \delta) & \text{estimation par régression} \\ \text{est } \bar{X} = (A - \delta)\ \bar{x}/\bar{y} & \text{estimation par ratio} \end{cases}$$

Remarques

1/ Les cas où β est négatif - ou encore A négatif - sont analogues.

2/ Comme on dispose des couples échantillons $(x_i\ y_i)$, on devra d'abord s'en servir pour apprécier le bien-fondé de l'hypothèse de structure. Il sera éventuellement possible de substituer à y une fonction $\varphi(y)$ choisie de façon à améliorer la colinéarité avec x.

3/ On pourrait passer de là au cas où l'information supplémentaire concerne deux variables (ou davantage) y, z. On pourrait faire des estimations par ratio (cf. la communication de Jaroslav HAJEK, au Congrès de l'I.I.S. de Bruxelles 1958). Il sera plus commode de postuler l'existence d'une structure ad hoc.

$$x = \alpha + \beta\ y + \gamma\ z + \varepsilon$$

d'où $\quad \text{est. } \bar{X} = \bar{x} - (\text{est. } \beta)\ (\bar{y} - X) - (\text{est. } \gamma)\ (\bar{z} - \bar{Z})$

Posons $\quad \bar{Y} = A - \delta, \ \bar{Z} = B - \delta', \ \boxed{\delta\ ,\ \delta' \geqslant 0}$

$$\Longrightarrow \text{est. } \bar{X} = \bar{x} - (\text{est. } \beta)\ (\bar{y} - A + \delta) - (\text{est. } \gamma)\ (\bar{z} - A + \delta')$$

On estime β, γ sur les unités échantillons $(x\ y\ z)_i$.

Les possibilités d'emploi pratique paraissent très douteuses, ce qui nous dispense d'examiner les difficultés, pouvant se présenter.

II - STRATIFICATION A POSTERIORI DE L'ECHANTILLON

Avec le vocabulaire en usage en France (à l'I.N.S.E.E.), on peut dire, qu'ici l'information :

$$\overline{Y} \leqslant A$$

va être exploitée en vue de redresser l'échantillon, supposé mal réparti entre les 2 classes :

$$(y \leqslant A) \text{ et } (y > A)$$

Nous allons trier les unités échantillons (i) sur la variable y, et

dénombrer $\begin{cases} n_1 \text{ unités échantillon vérifiant } y_i \leqslant A \text{ (classe 1)} \\ n_2 \quad - \quad - \quad - \quad y_i > A \text{ (classe 2)} \end{cases}$

$n = n_1 + n_2$

contre respectivement N_1 et N_2 dans la population. Désignons les moyennes de classe par $\overline{y}_1$ $\overline{y}_2$ $\overline{Y}_1$ $\overline{Y}_2$; il vient

$$\overline{y} = \frac{n_1}{n}\overline{y}_1 + \frac{n_2}{n}\overline{y}_2 > A$$

contre

$$\overline{Y} = \frac{N_1}{N}\overline{Y}_1 + \frac{N_2}{N}\overline{Y}_2 \leqslant A$$

Nous allons admettre qu'on aurait (si N_1 N_2 étaient connus) :

$$\frac{N_1}{N}\overline{y}_1 + \frac{N_2}{N}\overline{y}_2 \leqslant A$$

parce que l'échantillon est suffisamment grand pour que le point moyen observé $(\overline{y}_1\ \overline{y}_2)$ soit du bon côté de la droite $\frac{N_1}{N}n_1 + \frac{N_2}{N}n_2 = A$ (fig. 1).

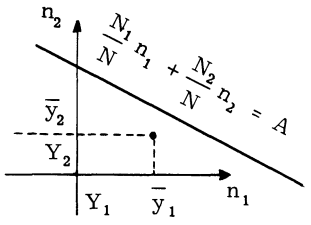

fig. 1

S'il n'en était pas ainsi, il conviendrait plutôt d'augmenter d'abord la taille n de l'échantillon. Nous allons admettre qu'on a :

$$E\overline{y}_1 = \overline{Y}_1, \quad E\overline{y}_2 = \overline{Y}_2$$

c'est-à-dire que les déformations subies par l'échantillon global n'em-

pêchent pas les deux sous-échantillons de classe d'être représentatifs ; de leurs classes respectives. L'influence de ces déformations consiste seulement en ceci : que les rapports $\frac{n_1}{n}$ et $\frac{N_1}{N}$ sont beaucoup plus éloignés l'un de l'autre que la loi binomiale des écarts ne le ferait attendre a priori.

Avoir $\boxed{\bar{y} > A}$ signifie donc que $\frac{n_1}{n}$ (poids affecté à $\bar{y}_1$) est trop petit, $\frac{n_2}{n}$ trop grand, d'où :

$$\frac{n_1}{n}\bar{Y}_1 + \frac{n_2}{n}\bar{Y}_2 > A$$

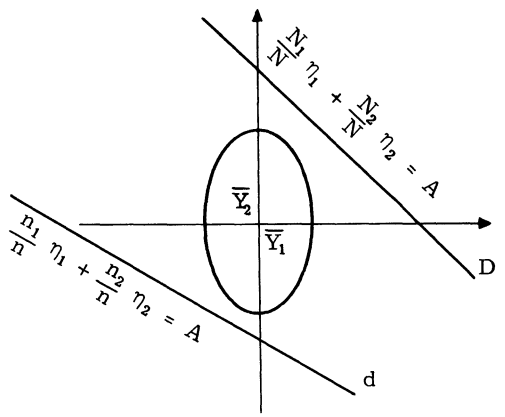

fig. 2

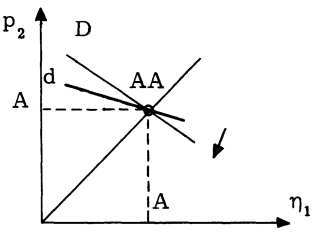

Fig. 3

La classe 1 est "sous-représentée" et la classe 2 "sur-représentée". Ce phénomène est bien connu des praticiens; par exemple les entreprises de petites tailles font défaut dans les échantillons ; de même les ménages ayant des revenus élevés (fig. 2).

Considérons la droite (d) d'équation parfaitement déterminée :

$$\frac{n_1}{n}\eta_1 + \frac{n_2}{n}\eta_2 = A$$

et la droite (D) de pente inconnue

$$\frac{N_1}{N}\eta_1 + \frac{N_2}{N}\eta_2 = A$$

Elles se coupent en un point tel que :

$$\frac{n_1}{n}\eta_1 + \frac{n_2}{n}\eta_2 = \frac{N_1}{N}\eta_1 + \frac{N_2}{N}\eta_2$$

c'est-à-dire sur la 1e bissectrice ($\eta_1 = \eta_2$), au point (A, A) (fig. 3).

On passera donc de (d) à (D) en faisant pivoter (d) autour du point (A, A). L'angle dont il faut faire tourner (d) doit être tel que le point ($\bar{y}_1$ $\bar{y}_2$), qui se trouve au-dessus de (d), soit au-dessous de (D) ou sur (D).

La rotation minimum fait donc passer (D) par le point (y_1, y_2).

L'équation de cette "droite D minimum," passant par (A, A) et $(\bar{y}_1, \bar{y}_2)$ est :

$$\nu_1 \eta_1 + \nu_2 \eta_2 = A$$

avec
$$\nu_1 A + \nu_2 A = A \Longleftrightarrow \nu_1 + \nu_2 = 1$$
$$\nu_1 \bar{y}_1 + \nu_2 \bar{y}_2 = A$$

ou
$$\frac{\bar{y}_2 - A}{\bar{y}_2 - \bar{y}_1}\eta_1 + \frac{A - \bar{y}_1}{\bar{y}_2 - \bar{y}_1}\eta_2 = A$$

L'ensemble des droites (D) vérifie :

$$\nu_1 \bar{y}_1 + \nu_2 \bar{y}_2 = A - \delta, \quad \delta \geqslant 0,$$

d'où :

$$\left(\frac{\bar{y}_2 - A + \delta}{\bar{y}_2 - \bar{y}_1}\right)\eta_1 + \left(\frac{A - \delta - \bar{y}_1}{\bar{y}_2 - \bar{y}_1}\right)\eta_2 = A$$

Conséquence : Formule d'estimation pour une variable non contrôlée :

$$x^* = \frac{\bar{y}_2 - A + \delta}{\bar{y}_2 - \bar{y}_1}\bar{x}_1 + \frac{A - \delta - \bar{y}_1}{\bar{y}_2 - \bar{y}_1}\bar{x}_2$$

Et bien entendu :

$$\text{est.}\left(\frac{N_1}{N}\right) = \frac{\bar{y}_2 - A + \delta}{\bar{y}_2 - \bar{y}_1}, \quad \text{est.}\left(\frac{N_2}{N}\right) = \frac{A - \delta - \bar{y}_1}{\bar{y}_2 - \bar{y}_1}$$

B - UTILISATION AU COURS DU TIRAGE

I - Echantillon "équilibré" (balanced sample")

1/ Procédons encore par analogie. La condition

$$\bar{Y} = A$$

peut être parfois utilisée pour obtenir un échantillon de (x y) , de moyennes

$$(\bar{x}, \bar{y} \neq A)$$

sous réserve que le nombre d'unités de sondage effectivement atteintes soit plus élevé que la taille n de l'échantillon ; il s'agit donc au fond d'une sorte de sondage à 2 phases, dont la 1e phase ne porte que sur la ou les variables de contrôle y, la 2e phase concernant les variables à l'étude x. (On sait que pareille technique est souvent anti-économique).

Le procédé n'est correct (estimateur sans biais) qu'autant que l'échantillon final est bien tiré au sort, ce qui est le cas pour une certaine technique de sélection due à YATES.

2/ Par analogie, nous pouvons aborder par la même technique le cas où l'on sait seulement que :

$$\bar{y} \leqslant A$$

Un échantillon S de taille n nous fournit $\bar{y}$ supérieur à A.

Soit $(y_1\ y_2 \ldots y_{n-1}\ y_n)$ cet échantillon rangé dans un ordre quelconque mais préalablement donné ; on tire une $(n+1)^e$ unité et on considère l'échantillon (privé de y_1) :

$$(y_2\ y_3 \ldots y_n\ y_{n-1}) \text{ de moyenne } \bar{y}'.$$

Si $\bar{y}' \leqslant A$ on s'arrête.

Si $\bar{y}' > A$ on poursuit en éliminant y_2 et en tirant au sort y_{n+2}. Et on continue aussi longtemps qu'on n'a pas obtenu $\bar{y}^* \leqslant A$.

On finit donc par avoir un échantillon S^*, de taille n, tiré au sort et vérifiant $\bar{y}^* \leqslant A$. On soumet S^* à la 2e phase du sondage, qui consiste à observer les x_1 ; d'où $\bar{x}^*$ = est $\bar{X}$.

L'échantillon S^* est bien entendu meilleur que S en ce qui concerne les y ; et s'il y a une forte corrélation (linéaire) entre les x et les y, il est également meilleur pour les x. Cependant c'est un échantillon biaisé, on n'a pas :

$$E\,\bar{x} = \bar{X}$$

alors qu'avec S on aurait eu : $E\,\bar{x} = \bar{X}$

Ceci est d'autant plus à signaler que l'échantillon est au contraire, sans biais :

- rigoureusement si $\bar{y} = A$ (Yates) ;
- ou approximativement si $B \leqslant \bar{y} < A$ lorsque A et B sont 2 bornes voisines l'une de l'autre ;
- soit enfin si la corrélation entre x et y est nulle (cas dépouvu d'intérêt).

Pour le voir il suffit de porter $\bar{x}$ et $\bar{y}$ sur deux axes de coordonnées cartésiennes et d'admettre n assez grand. Le couple $(\bar{x}, \bar{y})$ est distribué suivant une loi de Gauss de centre $(\bar{X}, \bar{Y})$.

Considérons une quelconque ellipse de probabilité de cette loi :

Si $\bar{y}^* = A$ ou $B < \bar{y}^* < A$ (B, A très voisins) le point $(\bar{x}^*\ \bar{y}^*)$ est astreint à rester dans une bande (fig. 4) horizontale de l'ellipse, centrée en X ; par raison de symétrie, on a donc :

$$E\,\bar{x}^* = X$$

et la bande est moins large que l'ellipse, ce qui prouve que la loi de $\bar{x}^*$ est moins dispersée que celle de $\bar{x}$.

Mais si l'échantillon S^* est soumis à la seule restriction $\bar{y}^* < A$ le point $(\bar{x}^*, \bar{y}^*)$ est astreint à se trouver dans la portion d'ellipse (fig. 5), située en dessous de $\bar{y} = A$; le point $(\bar{x}\ \bar{y})$ pouvait se trouver dans toute l'ellipse.

La dissymétrie introduite implique un biais : $E\,\bar{x}^* < X$ biais d'autant plus important que A est plus voisin de $\bar{Y}$ et aussi que la corrélation entre $\bar{x}$ et $\bar{y}$ est plus grande.

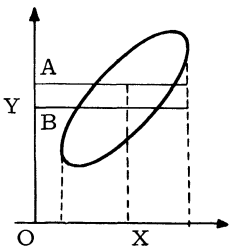

fig. 4

Dans le cas où cette corrélation est nulle (fig. 6) l'ellipse a ses axes parallèles aux axes de coordonnées et il n'y a plus de biais.

fig. 5 et 6

4/ Si l'on veut bien accepter l'idée que la présence d'un biais est en pratique un défaut moins grave pour un estimateur qu'on ne se l'imagine en lisant les manuels, - on peut essayer de se faire une idée de la variance $V\bar{x}^*$ ou de tout autre critère de qualité. On observe que :

- Dans le cas de la figure 6 : si A coïncide avec $\overline{Y}$ l'effet est nul, avec $A > \overline{Y}$ on supprime des cas de faible probabilité autour du centre mais non aux extrêmités de l'intervalle de variation de $\bar{x}^*$, d'où

$$V\bar{x}^* > V\bar{x}$$

- Dans le cas de la figure 5 (avec forte corrélation) le contrôle sur $\bar{y}$ ramène vers le centre les cas aberrants ($\bar{x}$ trop grand) et par suite réduit la variance

$$V\bar{x}^* < V\bar{x}$$

II - Utilisation de l'information $\overline{Y} \leqslant A$ au cours du tirage (suite)

1/ Peu avant son décès M. Vincent FONSAGRIVE nous avait fait part de sa conviction que l'information supplémentaire ($\overline{Y} \leqslant A$) serait susceptible de conduire à une meilleure utilisation des ressources dans un sondage stratifié. Soit une population (constituée de deux strates) ; $\overline{Y}_1$, $\overline{Y}_2$ les moyennes de strate, et $(N_1 \overline{Y}_1 + N_2 \overline{Y}_2)/N = Y$ la moyenne générale. Pour simplifier supposons :

$$N_1 = N_2$$

donc :

$$\frac{\overline{Y}_1 + \overline{Y}_2}{2} = \overline{Y}$$

(avec $N_1 \neq N_2$ on aurait des moyennes pondérées qui alourdissent les calculs sans altérer le résultat ; bien entendu on suppose connus N_1 et N_2). Supposons qu'on sache a priori que :

$$\overline{Y} \leqslant A$$

c'est-à-dire :

$$\overline{Y}_1 + \overline{Y}_2 \leqslant 2\ A$$

Soit $\overline{y}_1$ $\overline{y}_2$ les moyennes échantillons.

Supposons les tailles des échantillons assez grandes pour que

$\overline{y}_1$ $\overline{y}_2$ suivent chacune une loi de Laplace-Gauss

$\overline{y}_1 = \mathcal{N}(\overline{Y}_1\ ,\ \sigma_1/\sqrt{n_1})$

$\overline{y}_2 = \mathcal{N}(\overline{Y}_2\ ,\ \sigma_2/\sqrt{n_2})$

Le point $(\overline{y}_1\ ,\ \overline{y}_2)$ suit une loi de Gauss à 2 variables indépendantes si les tirages d'échantillon dans les 2 strates sont indépendants l'un de l'autre. Alors on a :

$$V\overline{y} = V\left(\frac{\overline{y}_1 + \overline{y}_2}{2}\right) = \frac{1}{4}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}\right)$$

Dans le plan rapporté aux axes rectangulaires $(\eta_1\ \eta_2)$, l'ellipse de probabilités est centrée en $(\overline{Y}_1\ \overline{Y}_2)$ et droite (fig. 7).

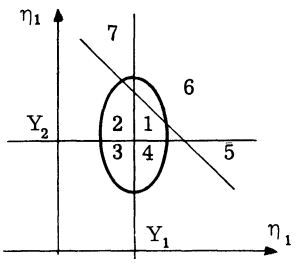

fig. 7

Elle a pour équation :

$$\frac{\eta_1^2}{\sigma_1^2/n_1} + \frac{\eta_2^2}{\sigma_2^2/n_2} = \lambda^2$$

Sa forme dépend du choix de n_1 et n_2 (en supposant σ_1 σ_2 connus). Ce plan est partagé en outre en 7 régions par les droites d'équations :

$$\eta_1 + \eta_2 = 2\ A,\ \eta_1 = \overline{Y}_1,\ \eta_2 = \overline{Y}_2$$

2/ La théorie de la répartition optimale de l'échantillon (Neyman-Yates) détermine la forme "optimale" de l'ellipse ; car parmi une certaine classe de distributions (telles que $n_1 + n_2$ = constante), elle désigne la distribution pour laquelle on a :

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \text{ minimum}$$

3/ On peut alors imaginer qu'on pourra, dans certains cas, sans augmentation du coût d'enquête, procéder à l'enquête par tranches progressives, en suivant l'évolution de $\bar{y}_1$ et $\bar{y}_2$ à mesure que l'échantillon grandit. Comme $\bar{Y}_1$ et $\bar{Y}_2$ sont inconnues, on ignore dans laquelle des 7 régions se trouve précisément le point $(\bar{y}_1, \bar{y}_2)$ intermédiaire, mais comme A est connu, on sait si ce point $(\bar{y}_1\ \bar{y}_2)$ est au-dessus ou au-dessous de la droite $\bar{y}_1 + \bar{y}_2 = 2\,A$, s'il appartient à (5, 6 ou 7), ou bien à (1 2 3 4).

Dès lors on peut imaginer que le statisticien essaie d'intervenir pour améliorer son sondage en dirigeant son point $(\bar{y}_1\ \bar{y}_2)$ intermédiaire, de préférence vers les régions (1 2 3 4) avant que son échantillon n'ait atteint la taille ou le coût imposé par le problème.

A priori il paraît impossible qu'on obtienne jamais une répartition optimale de l'échantillon, meilleure que celle que donne la théorie de Neyman-Yates. Mais ceci s'explique pourtant : si nous parvenons à intervenir c'est que nos tirages dans la strate 2 ne sont plus indépendants du résultat des tirages dans la strate 1 ; et par conséquent la formule

$$V\bar{y} = \frac{1}{4}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \geqslant \frac{1}{4}\,\frac{(\sigma_1 + \sigma_2)^2}{n}$$

doit céder la place à :

$$V\bar{y} = \frac{1}{4}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) + \frac{1}{2}\,\frac{\sigma_1\,\sigma_2}{\sqrt{n_1\,n_2}}\,\rho(\bar{y}_1,\ \bar{y}_2)$$

Dans la mesure où nous serons capable de rendre ρ négatif, nous pourrons faire descendre $V\bar{y}$ au-dessous de $(\sigma_1 + \sigma_2)^2/n$, son minimum au sens de Neyman.

Sur la figure 8 l'ellipse de répartition de la loi de Gauss suivie par $(\bar{y}_1\ \bar{y}_2)$ n'est plus droite mais oblique (allongée du Nord-Ouest vers

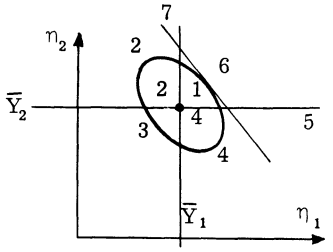

fig. 8

le Sud-Est) de façon à éviter le plus possible les régions 5-6-7.

4/ Les difficultés qui se présentent alors nous ont longtemps arrêté.

- D'une part il est clair que l'estimation de $\bar{Y}$ sera biaisée si aucune mesure n'est prise pour diminuer les chances de tomber dans la région 3, symétrique des régions 6 et 1.

- D'autre part (et c'est très grave) nous n'avons pas rencontré d'exemple concret où l'on sache comment diriger le cheminement du point $(\bar{y}_1, \bar{y}_2)$; nous n'avons (rappelons-le) le droit de faire qu'une chose : tirer des unités au sort dans chaque strate mais pas forcément dans les proportions qu'indique la théorie de Neyman-Yates.

Pour faire baisser $\bar{y}_1 + \bar{y}_2$ (sachant $\bar{Y}_1 < \bar{Y}_2$ par exemple), vaut-il mieux tirer les 2 prochaines unités de la strate 1, ou de la strate 2, ou des 2 strates ? Cela doit dépendre de toutes les informations que nous avons recueillies au cours des étapes précédentes de l'échantillonnage ; mais c'est en soi un problème théorique à étudier à part. Sinon nous en extrayons l'estimation de σ_1 et σ_2 et la répartition de n entre n_1 et n_2 à la façon de Neyman.

Il n'y a aucune raison pour qu'on doive prélever dans la strate 1 pour faire diminuer $\bar{y}_1 + \bar{y}_2$ ou dans la strate 2 pour faire augmenter cette somme.

Il n'en reste pas moins que l'idée de M. FONSAGRIVE était fort raisonnable ; parce que toute la théorie que nous évoquons est asymptotique, - elle peut être en défaut dès qu'on a affaire à des petits échantillons.

Supposons les distributions très dissymétriques dans les 2 strates et, pour simplifier, admettons qu'on ait : $\bar{y}_1 < A$ dans la strate 1,

$\bar{y}_1 > A$ dans la strate 2.

Nous constatons que (disons) $\bar{y}_1 + \bar{y}_2$ est trop grand. Avec des distributions dissymétriques dans chaque strate, médiane et moyenne diffèrent ; et un écart a plus de chances d'être entre A et la moyenne de strate que de s'écarter. Ainsi une donnée supplémentaire y aurait une probabilité supérieure à 1/2 de faire baisser $\bar{y}$ si on la prélève dans la strate 2, ou de faire monter $\bar{y}$ si on la prélève dans la strate 1, du moment qu'on suppose la moyenne et la médiane disposées comme dans une demi-loi de Laplace-Gauss, et que les échantillons ne sont pas très grands. Il faut donc prélever dans la strate 2.

Ceci montre bien combien il convient d'être prudent avec les théories asymptotiques, - qui conduisent à rejeter des conceptions valables avec certains petits échantillons, comme celle de M. FONSAGRIVE.

C - UTILISATION DE L'INFORMATION $\bar{Y} \leqslant A$ DANS UN PLAN DE SONDAGE

I - Sondage à deux phases

Un sondage à 2 phases pourrait, dans certains cas, convenir pour le présent problème. Cela suppose que l'acte de juger si y_i est supérieur ou inférieur à A soit beaucoup moins onéreux que l'enquête proprement dite fournissant les valeurs $(x_i\ y_i\ z_i\ ..)$ prises par $(x\ y\ z\ ..)$ sur (i).

Nous évaluerons donc, dans une phase de l'enquête, les proportions P et Q des unités appartenant aux 2 classes $(y_i \leqslant A,\ y_i > A)$. Après quoi on peut espérer l'estimateur

$$x^* = p\,\bar{x}_1 + q\,\bar{x}_2, \qquad \text{où} \quad \begin{cases} p = \text{est. } P \\ q = \text{est. } Q \end{cases}$$

meilleur que :

$$\bar{x} = \frac{n_1}{n}\bar{x}_1 + \frac{n_2}{n}\bar{x}_2$$

Problème : Trouver la répartition optimale des ressources entre les phases et strates.

Phase 1 : Tirer au sort m unités de la population et estimer

par p la proportion P
$q = 1 - p$ - Q

Phase 2 : Tirer n_1 unités de la strate $y \leqslant A$ (par exemple parmi les p m ci-dessus) et mesurer leurs $x_i\ y_i\ z_i$ (cette fois avec soin) et leurs moyennes $\bar{x}_1\ \bar{y}_1\ \bar{z}_1$

$$\text{Posons : } u = A - \bar{y}_1 = \frac{1}{n_1} \sum_{(1)} (A - y_i)$$

Procéder de même pour n_2 unités de l'autre strate ; soit $\bar{y}_2$ la moyenne des y.

$$\text{Posons : } v = \bar{y}_2 - A = \frac{1}{n_2} \sum_{(2)} (y_i - A)$$

$$\text{Soit : } y^* = p\,\bar{y}_1 + q\,\bar{y}_2$$

$$y^* < A \Longleftrightarrow p\,u > q\,v \Longleftrightarrow pu - qv > 0$$

Calcul de variance

Pour que (pu - qv) ait la variance minimum, on doit calculer

$$V(pu - qv) = V(pu) + V(qv) - 2\,\mathrm{Cov}\,(pu, qv)$$

Pour simplifier supposons que les tirages de n_1 et de n_2 unités des 2 strates sont indépendants entre eux et indépendants du premier tirage de m unités ; ce n'est qu'approximativement correct. Pour un couple donné $\bar{y}_1\ \bar{y}_2$ il vient :

$$V(pu) = u^2\,Vp = u^2 \frac{PQ}{m}$$

$$V(qv) = v^2\,Vq = v^2 \frac{PQ}{m}$$

$$\mathrm{Cov}(pu, qv) = uv\,\mathrm{Cov}(p, q) = -uv\,\frac{PQ}{m}$$

$$V(pu - qv) = \frac{PQ}{m}(u^2 + v^2 + 2uv) = \frac{PQ}{m}(u + v)^2$$

$$= \frac{PQ}{m}(\bar{y}_1 - \bar{y}_2)^2$$

Pour des couples aléatoires $\bar{y}_1\ \bar{y}_2$ (indépendants entre eux et indépendants de p), il vient :

$$V(pu - qv) = \frac{PQ}{m}(V\bar{y}_1 + V\bar{y}_2)$$

$$= \frac{PQ}{m}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Coût : Supposons les ressources disponibles égales à C, la fonction de coût étant linéaire

$$m\ c_o + n_1\ c_1 + n_2\ c_2 = C$$

Optimum : Le minimum de V lié par C correspond à :

$$\frac{\frac{1}{m}\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{c_o} = \frac{\frac{\sigma_1^2}{n_1^2}}{c_1} = \frac{\frac{\sigma_2^2}{n_2^2}}{c_2}$$

ou :

$$\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{m\ c_o} = \frac{\frac{\sigma_1^2}{n_1}}{n_1 c_1} = \frac{\frac{\sigma_2^2}{n_2}}{n_2 c_2} = \frac{2\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)}{C}$$

d'où : $\begin{cases} m\ c_o = \frac{1}{2}\ C, \quad m = C/2\ c_o \\ \frac{n_1^2 c_1}{\sigma_1^2} = \frac{n_2^2 c_2}{\sigma_2^2} = \frac{1}{2}\ C \end{cases}$ $\qquad \begin{cases} n_1 = \sigma_1 \sqrt{C/2\ c_1} \\ n_2 = \sigma_2 \sqrt{C/2\ c_2} \end{cases}$

la moitié des ressources devra être consacrée à l'estimation de p, l'autre sera répartie suivant les normes classiques (proportionnellement à σ et à $c^{-\frac{1}{2}}$).

Critique :

Le calcul précédent (très classique) répond-il bien aux besoins ? On a cherché à avoir une variable (pu - qv) = ξ aussi peu dispersée que possible autour de sa moyenne. On désirait en fait que (pu - qv) ait une probabilité maximum d'être une variable positive, une probabilité minimum d'être négative (on sait que son espérance mathématique est fixe, positive, inconnue). On constate que les 2 points de vue coïncident si ξ a une distribution symétrique.

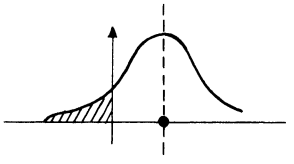

fig. 9

II - Sondage en 2 étapes (de STEIN)

On peut croire que l'information ($\bar{y} \leqslant A$) serait utilisable avec un sondage en plusieurs étapes analogue à celui (assez connu) dû à STEIN : Rappelons de quoi il s'agit : dans une première étape, un échantillon (disons de 300 unités) permet de se faire une bonne idée de σ et de $\overline{Y}$; après quoi l'échantillon total sert à estimer $\overline{Y}$ avec une précision exigée d'avance, à supposer que la 1ère étape n'ait pas suffi à elle seule à atteindre cette précision ; la taille de l'échantillon total est aléatoire, elle dépend de ce qu'a donné le premier tirage (de 300 unités).

Dans le même esprit on peut vouloir tirer un 1er échantillon (n_1) ; qui suffira d'ailleurs s'il donne $\bar{y}_1 \leqslant A$; mais qui, lorsqu'il fournit

$$\bar{y}_1 > A$$

permet d'évaluer σ et par conséquent de calculer quelle taille ($n_1 + n_2$) est nécessaire pour obtenir $\bar{y} < A$;

avec $$\bar{y} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2$$

Soit P la probabilité qu'on s'arrête à la 1ère étape. L'espérance mathématique de la taille (aléatoire) de l'échantillon est :

$$Pn_1 + (1 - P)\ (n_1 + n_2) = \nu$$

La variance est : dans le 1er cas :

$$V\bar{y}_1 = \frac{\sigma^2}{n_1}$$

dans le 2e cas :

$$V\bar{y}_2 = \frac{\sigma^2}{n_1 + n_2}$$

L'espérance mathématique de la variance est donc :

$$\mathcal{E}V = P\ V\bar{y}_1 + (1 - P)\ V\bar{y}_2 = \sigma^2\left(\frac{P}{n_1} + \frac{1 - P}{n_1 + n_2}\right)$$

On constate alors que, pour une valeur ν donnée de la taille moyenne, la variance $\mathcal{E}V$ est <u>minimum quelque soit P</u> pour :

$$n = 0$$

En effet posons : $n_1 + n_2 = \lambda$, $\nu = P\ n_1 + (1 - P)\ \lambda$

$$\sigma^{-2}\ \mathcal{E}V = \frac{P}{m_1} + \frac{1 - P}{\lambda}$$

La proportionnalité des dérivées partielles en n_1 et λ donne :

$$\frac{1}{n_1^2} = \frac{1}{\lambda^2} \Longrightarrow \lambda = n_1$$

<u>Conclusion</u> : La variance moyenne $\mathcal{E}V$ ne peut qu'augmenter si on répartit le sondage en deux étapes.

Cet accroissement de variance est en somme le prix à payer pour avoir l'information σ et par conséquent contrôler le risque d'avoir $\bar{y} > A$.

<u>Observation finale</u> (grand échantillon)

Le couple des moyennes ($\bar{y}_1$ $\bar{y}_2$) défini tant au AII (stratification a posteriori) qu'au CII suit une loi de Gauss - l'ellipse de probabilité étant droite.

L'estimateur $y^* = \nu_1 \bar{y}_1 + \nu_2 \bar{y}_2$, ν_1 paramètre quelconque

$$\nu_2 = 1 - \nu_1, \quad 0 \leqslant \nu_h \leqslant 1$$

est une variable de Laplace Gauss, estimateur sans biais de $\bar{Y}$.

On peut appliquer à y^* tout ce qui a été dit au §I à propos de $\bar{x}$.

Par suite, le choix de δ au §AII ci-dessus n'est pas arbitraire ; il dépend de la probabilité P_o qu'on peut accepter : Une fois δ choisi, P_o représente le risque (accepté) d'avoir $x > A$.