REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Le sondage systématique au sens large

Revue de statistique appliquée, tome 15, n° 2 (1967), p. 5-18

<http://www.numdam.org/item?id=RSA_1967__15_2_5_0>

LE SONDAGE SYSTÉMATIQUE AU SENS LARGE

P. THIONET
Faculté des Sciences de Poitiers

1 - INTRODUCTION

Lors de sa réunion en juillet 1963 à Genève, le groupe de travail de l'O.N U. sur les sondages a reconnu qu'en pratique les mots "sondage systématique" pouvaient avoir un sens beaucoup plus large que celui qu'on lui donne dans les manuels. Il nous a paru ressortir des échanges de vue que c'était les réunions et intersections d'échantillons systématiques (au sens strict) qu'on désignait par échantillons systématiques (au sens large). Toutefois le groupe n'avait pas à écrire un manuel de sondage, il ne devait rien enseigner, ni exposer, mais seulement mentionner (dans bien des cas, suggérer) ; d'ailleurs en majorité les personnes présentes (évitant l'usage des mathématiques) semblaient peu disposées à clarifier ce point. Enfin en cherchant les conséquences extrêmes d'une telle définition on arriverait à cette conclusion que n'importe quel échantillon peut être considéré comme systématique (au sens large).

2 - NOTATIONS ET DEFINITIONS

On suppose que les unités de sondage doivent être tirées avec d'égales probabilités. Pour simplifier (bien que ce ne soit pas indispensable) on suppose N divisible par n ; on pose $N = nq$; $f = 1/q$ est la fraction sondée.

Définition : L'ordre d'énumération des unités étant supposé choisi, un échantillon systématique au sens strict est tel que les numéros des unités échantillon, soient a, $a + q, \ldots, a + (n - 1)\, q$, avec :

$$1 \leqslant a \leqslant q$$

Par exemple, si $f = 1/20$, l'échantillon formé des unités n° 3, 23, 43, 63,... est systématique.

Echantillon systématique (sens large). Le groupe de travail semble avoir admis que l'échantillon suivant serait également systématique : n° 3, 4, 43, 44, 83, 84, etc. ou encore : n° 3, 7, 43, 47, 83, 87, etc.

Il s'agit (comme on le voit) de la réunion de 2 échantillons distincts systématiques (au sens strict) de fraction sondée 1/40.

1ère définition.

Plus généralement, on peut constituer un échantillon systématique qui corresponde à la fraction de sondage f en réunissant plusieurs échan-

tillons systématiques dont la somme des fractions de sondage est f, à condition que leurs intersections 2 à 2 soient vides.

1er exemple : les unités de sondage :

n° 1, 2, 6, 10, 11, 14, 18, 21, 22, 26, 30, 31, 34, 38 ; etc. constituent un échantillon systématique avec f = 7/20 ; car on peut y reconnaître la réunion de 2 échantillons systématiques (sens strict)

	1	11	21	31	fraction $\frac{1}{10} = \frac{2}{20}$
et	2,6,10	14 18	22,26,30	34,38	fraction $\frac{1}{4} = \frac{5}{20}$

c'est-à-dire 7 échantillons systématiques au $1/20^{e}$ avec respectivement a = 1, 2, 6, 10, 11, 14, 18.

Toutefois ceci suppose que les intersections (deux à deux) des échantillons systématiques réunis soient vides : montrons le sur un exemple :

2ème exemple : la réunion des 2 échantillons systématiques suivants :

n° 1 11 21 31

n° 3 7 11 15 19 23 27 31

de fractions sondées 2/20 et 5/20 respectivement, constitue un échantillon systématique (au sens large), dont la fraction de sondage est $\frac{2+5-1}{20} = \frac{6}{20}$ et non $\frac{7}{20}$; à moins de convenir qu'on pondère par 2 les unités-intersections n° 11, 31, 51,...).

2ème définition du sondage systématique (sens large).

On définit d'abord le dessin (pattern) de l'échantillon, c'est-à-dire un sous-ensemble de s unités de sondage diversement disposées qui, reproduit μ fois systématiquement, va constituer l'échantillon de taille $n = s\mu$.

Dans un exemple donné ci-dessus, le dessin de l'échantillon consiste en les unités suivantes :

n° 1, 2, 6, 10, 11, 14, 18

donc s = 7, tandis que f = 7/20.

On voit que :

$$\mu = \frac{n}{s} = \frac{Nf}{s} \; ; \; \frac{\mu}{N} = \frac{f}{s} \quad \left(\text{ici} = \frac{1}{20}\right).$$

3 - PLANS DE SONDAGE SYSTEMATIQUES PROBABILISTES

On obtient un plan de sondage "probabiliste" en tirant au sort l'une des unités de l'échantillon systématique.

Base refermée sur elle-même. Il est commode d'imaginer que chaque unité de sondage constituant la base est affectée des numéros $i + \alpha N = I$

$$i = 1, 2, \ldots, N, \qquad i \equiv I \pmod{N}$$

A - Sondage systématique (sens strict)

On tire au sort 1 échantillon parmi un sous-ensemble de q échantillons également probables,

- soit qu'on tire au sort le numéro a de la première unité de l'échantillon (parmi 1, 2,..., q) ;

- soit qu'on tire au sort le numéro i (parmi 1, 2,..., N = na) d'une certaine unité de l'échantillon.

B - Sondage systématique (sens large)

On choisit arbitrairement l'une des unités du dessin comme début du dessin ; et on tire au sort l'un des numéros i, qu'on affecte à ce début du dessin : soit i. Ainsi le dessin est mis en place.

L'échantillon s'obtient alors en faisant subir au dessin un tour complet sur la base de sondage refermée, reproduisant (n/s) fois le dessin.

Ici l'échantillon se trouve tiré (avec probabilités égales) dans un sous-ensemble de qs échantillons (de taille n).

Nota : Le symbole E (espérance mathématique) qui sera employé plus loin signifie donc : moyenne de qs valeurs équiprobables (avec s = 1 dans le sondage systématique au sens strict).

Exemple : N = 100, n = 4

Dessin : formé de 2 unités (i, i + 2) ; s = 2.

On tire (disons) i = 75 d'où le dessin : 75,77.

L'échantillon est formé des unités de sondage n° 25-27 - 75-77.

Autrement dit : le dessin subit deux rotations (sur la base refermée) et revient à sa position initiale.

Remarque 1 : Le dessin d'un échantillon systématique (au sens strict) se réduit à une seule unité de sondage. Tout l'échantillon est engendré par cette unité, à qui l'on fait subir des rotations d'amplitude $q = \frac{N}{n}$ sur la base refermée sur elle-même.

Remarque 2 : Tout échantillon systématique au sens large est la réunion de s sous-échantillons systématiques (au sens strict) ayant pour taille $n/s = \mu$, chacun étant engendré par des rotations d'amplitude $N/\mu = sq$ sur la base refermée.

Exemple : 25-75 et 27-77 sont deux sous-échantillons systématiques (fraction sondée 1/50).

Remarque 3 : L'échantillon systématique au sens large est défini par la position d'un seul élément de chaque sous-échantillon, notamment par les plus petits numéros d'ordre (compris entre 0 et sq).

Soit par exemple 1. 2. 6. 10. 11. 14. 18 le dessin d'un sondage avec f = 7/20. On tire au sort la position de l'unité de début parmi les numéros i = 1. 2. ... 20. On tire donc cet échantillon parmi 20 échantillons également probables. Chaque unité de sondage peut appartenir à 7 de ces 20 échantillons.

4 - ESTIMATION PAR SONDAGE SYSTEMATIQUE (AU SENS LARGE

A - Estimation d'une moyenne

Soit $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ ... les moyennes des échantillons S_1, S_2, S_3 ... de la population U ; soit λ_1, λ_2, λ_3 ... des poids positifs quelconques non aléatoires, avec :

$$\sum_h \lambda_1 = 1$$

$$E\bar{x}_h = X \qquad E \sum_h \lambda h \, \bar{x}_h = \bar{X}$$

donc : si les $\bar{x}_h$ sont estimations sans biais de la moyenne $\bar{X}$ de U, toute combinaison linéaire $\Sigma \lambda_h \bar{x}_h$ est aussi estimation sans biais.

Corollaire : Si les S_h sont des échantillons systématiques (au sens strict), tout estimateur $\Sigma \lambda_h \bar{x}_h$ estime sans biais $\bar{X}$.

En particulier : Si les S_h ont deux à deux leurs intersections vides, la moyenne de l'échantillon systématique (au sens large) qu'ils constituent est l'estimation sans biais de la moyenne de la population.

B - Variance de l'Estimation

On a :

$$V\left(\sum_h \lambda_h \bar{x}_h\right) = \sum_h \lambda_h^2 \, V\bar{x}_h + 2 \sum_{hh'} \lambda_h \, \lambda_{h'} . \mathrm{Cov}\, \bar{x}_h \bar{x}_{h'}.$$

En particulier : Pour deux sous-échantillons systématiques S_1 et S_2 de même taille, on a :

$$V\left(\frac{\bar{x}_1 + \bar{x}_2}{2}\right) = \frac{1}{2} V\bar{x}_1(1 + \rho)$$

où ρ = Corrélation $(\bar{x}_1, \bar{x}_2)$.

B' - Estimation de la variance

C'est le point sur lequel "accroche" le sondage systématique (sens strict). Comme pour le sondage semi-systématique de Gautschi [1], on peut estimer $V\bar{x}_1$ par $(\bar{x}_1 - \bar{x}_2)^2$ C.

Avec :

$$E(\bar{x}_1 - \bar{x}_2)^2 = 2 \, V\bar{x}_1(1 - \rho)$$

il vient :

$$\text{est. } V\left[\frac{\bar{x}_1 + \bar{x}_2}{2}\right] = \left[\frac{\bar{x}_1 - \bar{x}_2}{2}\right]^2 \cdot \frac{1 + \rho}{1 - \rho}$$

et on se retrouve devant le problème insoluble : estimer ρ, corrélation sériale :

Comme on tire 2 sous-échantillons à coup sûr distincts, il est assez vraisemblable que ρ est plutôt négative que positive. En confondant $(1+\rho)/(1-\rho)$ avec 1, il est donc possible qu'on estime la variance par excès, ce qui est préférable ; mais que dire de plus !

Remarque : Une valeur acceptable pour ρ en moyenne serait : $-(qs-1)^{-1}$. En effet, considérons les qs moyennes d'échantillons systématiques $\bar{x}_1$, $\bar{x}_2$,... et retranchons de chacune d'elles $\bar{X}$, moyenne générale. On a :

$$0 = (\bar{x}_1 - \bar{X}) + (\bar{x}_2 - \bar{X}) + \ldots$$

Elevons au carré les 2 nombres : il vient $(qs)^2$ termes au 2 membres, avec :

$$0 = qs\,\sigma^2 + qs(qs-1)\,\rho'\sigma^2$$

d'où

$$\rho' = -1/(qs-1);$$

mais ρ' n'est pas la corrélation sériale, c'est la corrélation entre les $\bar{x}_1$, $\bar{x}_2$ quel que soit leur ordre ; on peut y voir simplement une moyenne des corrélations sériales diverses obtenues en permutant $\bar{x}_1$, $\bar{x}_2$... de toutes les façons possibles.

Conclusion : Le sondage systématique au sens large ne se prête guère mieux aux calculs de variance d'échantillonnage que celui au sens strict.

Toutefois, il y a des raisons d'un tout autre ordre qui peuvent encourager son adoption.

5 - RAISONS POUR ADOPTER LE SONDAGE SYSTEMATIQUE AU SENS LARGE

L'extension du concept de sondage systématique proposé ici n'est pas nécessairement académique. Tout d'abord, il est parfaitement concevable que, par exemple, l'échantillon :

n° 1, 3, 11, 13, 21, 23,...

soit aussi précis que l'échantillon :

n° 1, 6, 11, 16, 21, 26,...

mais corresponde à un moindre coût d'enquête, les unités j et j + 2 étant voisines d'une de l'autre sur le terrain. C'est un argument de cette nature qui fait (on le sait) employer les sondages en grappes et les sondages à 2 degrés.

Nous allons voir qu'en outre l'échantillon systématique au sens large peut apporter plus d'informations sur la variance de la population que l'échantillon systématique au sens strict.

Pour cela, reprenons l'étude d'une population dont les unités sont rangées dans un ordre arbitraire mais donné.

6 - POPULATION ORDONNEE (RAPPEL)

La formule de la variance peut s'écrire :

$$1/ \qquad N^2\sigma^2 = \Sigma\Sigma(x_i - x_j)^2 \ , \ i < j$$

Classons les $(x_i - x_j)^2$ suivant la valeur de $\Delta = j - i$. Il y a :

$(N - 1)$ termes : $(x_i - x_{i+1})^2$

$(N - 2)$ - $(x_i - x_{i+2})^2$

---

1 - $(x_1 - x_N)^2$

au total $1 + 2 + \ldots + N - 1 = \frac{N}{2}(N - 1)$ différences carrées distinctes.

Posons :

$$\gamma_\Delta = \frac{\Sigma_i (x_i - x_{i+\Delta})^2}{(N - \Delta)}$$

$$\Longrightarrow \boxed{N^2\sigma^2 = (N - 1)\ \gamma_1 + (N - 2)\ \gamma_2 + (N - 3)\ \gamma_3 + \ldots + (x_N - x_1)^2}$$

$$2/ \quad 2N^2\sigma^2 = \Sigma\Sigma(x_i - x_j)^2 \qquad \forall i, j = 1, 2, \ldots N, (N+1), \ldots 2N$$

avec N termes

$$(x_i - x_{i+\Delta})^2 : i = 1, 2, \ldots, N \ ; \ \Delta \text{ fixé}$$

et ceci pour chacune des valeurs $\Delta = 1, 2, \ldots, N - 1$.

Au total $N(N - 1)$ termes.

Posant : $$1 - \rho_\Delta = \sum_{i=1}^{N} \frac{[x_i - x_{i+\Delta}]^2}{2N\sigma^2} \ ,$$

on définit les corrélations sériales ρ_Δ, dont le graphe est le corrélogramme (avec $\rho_0 = 1$ bien entendu). On en déduit :

$$\rho_0 + \rho_1 + \rho_2 + \ldots + \rho_{N-1} \equiv 0$$

Lorsque N est extrêmement grand, on peut confondre le début des deux développements, c'est-à-dire négliger $(x_N - x_1)^2$ dans le développement de $\sum_{i=1}^{N} (x_i - x_{i+1})^2$ où x_{N+1} signifie x_1 ; négliger $(x_{N-1} - x_1)^2 + (x_N - x_2)^2$ dans celui de $\Sigma(x_i - x_{i+2})^2$ où x_{N+2} signifie x_2 ; etc.

On a refermé la base de sondage sur elle-même.

Quelle que soit la commodité du procédé, nous nous en tiendrons (dans ce qui suit) au premier des deux développements, celui de σ^2 en fonction de γ_1 γ_2 γ_3 etc.

7 - USAGE D'UN PROCESSUS STOCHASTIQUE

Gautschi [1] désigne Cochran [2] comme le promoteur des processus stochastiques en sondage (1946) ; l'idée a été reprise par Hajek beaucoup plus tard [3].

Assimilons les unités de sondage ordonnées $i = 1,2,\ldots N$ à des dates discrètes successives et considérons la valeur de X sur l'unité i (soit x_i) comme une variable aléatoire X_t dépendant plus ou moins des valeurs $x_1\ x_2\ \ldots\ x_{i-1}$ prises par X_t aux dates précédant la date i. On peut supposer que i est un processus stationnaire du 2ème ordre. (L'espérance mathématique, la variance, la covariance entre x_t et $x_{t-\Delta}$ ne dépendent pas de t).

L'idée est à retenir en pratique si N est très grand, et qu'on se désintéresse des toutes premières et des toutes dernières valeurs de x_i.

On ne doit cependant pas en exagérer la portée, jusqu'à supposer X_t un processus d'accroissements indépendants ou d'autres processus particuliers aux propriétés très particulières : ce serait peu réaliste.

Le corrélogramme donne $\rho_\Delta = E(x_t - \bar{X})\ (x_{t-\Delta} - \bar{X})/\sigma^2$ avec $Ex_t = \bar{X}$, $\mathcal{U}x_t = \sigma^2$. Par hypothèse, ρ_Δ ne dépend pas de t.

On peut lui substituer $E(x_t - x_{t-\Delta})^2 = 2\sigma^2(1 - \rho_\Delta) = \gamma_\Delta$. Cette grandeur varie entre 0 et $4\sigma^2$.

Dans les problèmes réels de sondage, si Δ grandit, il arrive souvent que ρ_Δ devienne nulle : c'est-à-dire que x_t corresponde à n'importe quelle valeur de $x_{t-\Delta}$; alors que pour de faibles valeurs de Δ la corrélation entre x_t et $x_{t-\Delta}$ sera très nette, soit positive, soit négative. Il existe des cas particuliers où, au contraire, ρ_Δ est proche de + 1, ou de - 1, pour de grandes valeurs de Δ. Nous en donnerons plus loin des exemples.

Pour l'étude des cas pratiques, il faut revenir à N fini ; nous abandonnerons les corrélations sériales ρ_Δ et adopterons les écarts carrés sérials γ_Δ. Nous assimilerons :

$$\gamma_\Delta = \frac{\Sigma_i (x_i - x_{i+\Delta})^2}{(N - \Delta)} \qquad \text{avec} \qquad E(x_t - x_{t-\Delta})^2$$

et nous nous limiterons à $\Delta \leqslant N - 1$.

Au total :

$$E(x_i - x_j)^2 = \frac{\sum\sum_{i<j} (x_i - x_j)^2}{[N(N-1)/2]}$$

$$= \frac{N^2\sigma^2}{[N(N-1)/2]} = \frac{2N\sigma^2}{(N-1)}$$

8 - EXEMPLES

Exemple A : Supposons la base de sondage constituée par une liste de personnes composant les ménages (cas de la Schedule du recensement U.S.A.).

Les noms des mères de famille occupent des rangs dont l'écart moyen est (disons) $\Delta = 4$. Soit x_{ij} une variable égale à 1 ou 0, suivant que l'unité (ij) est ou non une mère de famille. On voit que γ_Δ varie énormément avec Δ quand Δ est petit : γ_1 est certainement peu inférieur à 1. γ_3 γ_4 et γ_5 sont certainement peu supérieurs à 0.

Mais, pour les grandes valeurs de Δ, on peut penser que γ_Δ reste de l'ordre de $\frac{6}{16} \cdot 1 + \frac{10}{16} : 0 = \frac{6}{16}$.

	x_j =	0	1	
	Probabilité	3/4	1/4	
Probabilité $(x_i = 0) = 3/4$		0	1	Valeurs de $(x_i - x_j)^2$
$(x_i = 1) = 1/4$		1	0	

Nota : La vraie valeur de σ^2 est pq, c'est-à-dire $(1/4).(3/4) = 3/16$.

La formule $E(x_i - x_j)^2 = 2N\sigma^2/(N-1)$ (pour N grand) donne :

$$E(x_i - x_j)^2 = 2\,\sigma^2 = 6/16$$

C'est dire que les valeurs γ_Δ pour Δ petit n'ont guère d'influence et que $2\sigma^2$ n'est autre que la limite de $\varepsilon(x_i - x_{i-\Delta})^2$ pour Δ infini.

Exemple B : x_i est fonction uniforme (non croissante) de i. Alors γ_Δ est fonction uniforme (croissante) de Δ (figure 1).

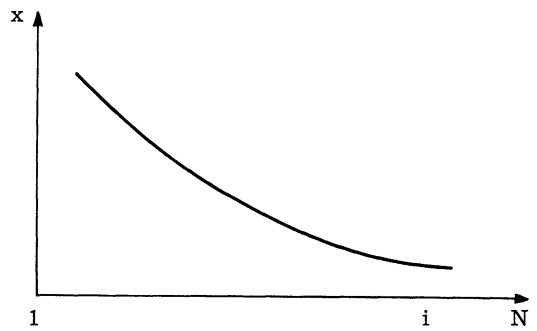


Figure 1 - Exemple B.

Exemple C : x_i est fonction périodique de i, de période T. Alors γ_Δ est nul pour $\Delta = T$ et multiples de T. En revanche γ_Δ est maximum pour $\Delta = \frac{T}{2}, \frac{3T}{2}$ etc.

Remarque : Les exemples ABC sont caricaturaux ; mais en supposant la variable X somme de deux composantes dont une est décrite en A B ou C et l'autre est aléatoire de moyenne nulle, on obtient des cas typiques bien connus illustrant l'intérêt (A) du sondage en petites grappes ; (B) du sondage stratifié ou systématique ; (C) du sondage en grandes grappes (ou les dangers du sondage systématique).

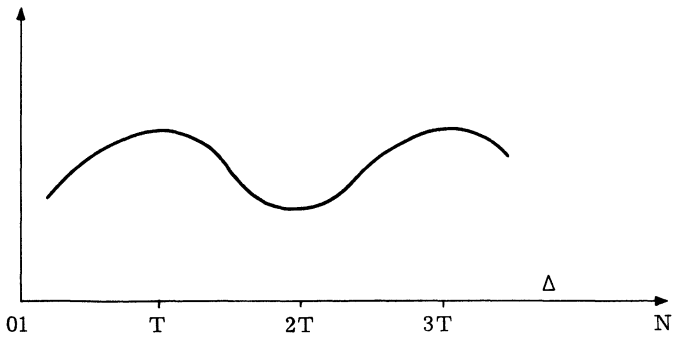


Figure 2 - Exemple C.

9 - ECHANTILLONNAGE SYSTEMATIQUE (AU SENS STRICT)

Supposons maintenant que, seul de toute cette population un échantillon systématique banal de raison q est connu. Alors on peut appliquer à sa variance s^2 la formule ayant servi pour σ^2 :

$$n^2 s^2 = S_i S_j (x_i - x_j)^2$$

et l'on a successivement :

$$[i - j] = q, \text{ ou } 2q, \text{ ou } 3q, \ldots$$

Il existe $(n - 1)$ termes $(x_i - x_{i+q})^2$, $i = a, a + q, a + 2q, \ldots$

$$(n-2) \text{ termes } (x_i - x_{i+2q})^2 \ldots\ldots$$

$$\ldots\ldots\ldots\ldots$$

$$1 \quad \text{terme } (x_a - x_{a+(n-1)q})^2$$

Posons :

$$g_\Delta = E(x_i - x_{i+\Delta})^2$$

il vient :

$$n^2 s^2 = (n - 1)\, g_q + (n - 2)\, g_{2h} + (n - 3)\, g_{3h} + \ldots$$

Par ailleurs, g_s est (bien entendu) estimateur sans biais de γ_Δ (quel que soit Δ).

Problème : Peut-on estimer σ^2 à l'aide de g_a, g_{2h}, g_{3h}, etc. ?

Si la population $x_1, x_2, \ldots x_N$ est mise sous forme d'un tableau à double entrée :

$$\begin{array}{llll} x_1 & x_{1+q} & x_{1+2q} & \ldots\ldots \\ x_2 & x_{2+q} & x_{2+2q} & \ldots\ldots \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ x_q & x_{2q} & x_{3q} & \ldots\ldots \quad x_N \end{array}$$

Si l'on pose $\sigma^2 = \sigma_e^2 + \sigma_i^2$

où σ_e^2 est la variance entre moyennes de ligne et σ_i^2 la variance moyenne à l'intérieur des lignes ; -on a pour la moyenne de l'échantillon systématique :

$$V\bar{x}_s = \sigma_e^2 = \sigma^2 - \sigma_i^2$$

(en effet, $\bar{x}_s$ est tout simplement la moyenne d'une ligne (a) du tableau).

D'autre part, on sait que $V\bar{x}_s$ peut s'écrire :

$$V\bar{x}_s = \frac{\sigma^2}{n} [1 + (n - 1)\ \bar{\rho}]$$

$\bar{\rho}$ désignant la corrélation "intraclasse", c'est-à-dire intra-ligne ; et on sait que $\bar{\rho}$ en pratique est positif ou négatif mais voisin de zéro. Ainsi σ_e^2 est-il généralement beaucoup plus petit que σ^2 ; et $\sigma^2 = \sigma_e^2 + \sigma_i^2$ est un peu plus grand que σ_i^2.

On sait aussi que s^2 (variance des éléments de l'échantillon systématique) est la variance de la ligne (a) du tableau et par suite estime (sans biais) σ_i^2.

1ère conclusion : Estimer σ^2 par s^2 introduit un biais par défaut (assez léger en pratique).

2ème conclusion : Pour estimer sans biais σ^2, il suffit d'avoir une estimation sans biais de γ_1, γ_2, γ_3, etc. ; mais il est toujours impossible de les obtenir par sondage systématique simple.

C'est le mérite du sondage systématique au sens large de fournir, au moins dans certains cas, des estimateurs sans biais de γ_1, γ_2, γ_3, ... γ_Δ quel que soit Δ, donc un estimateur sans biais de σ^2 :

$$\sigma^2 = \frac{N - 1}{N^2} \gamma + \frac{N - 2}{N^2} \gamma_2 + \frac{N - 3}{N^2} \gamma_3 + \ldots$$

$$\text{est } \sigma^2 = \frac{N - 1}{N^2} (\text{est } \gamma_1) + \frac{N - 2}{N^2} (\text{est } \gamma_2) + \frac{N - 3}{N} (\text{est } \gamma_3) + \ldots$$

10 - ECHANTILLONNAGE SYSTEMATIQUE AU SENS LARGE

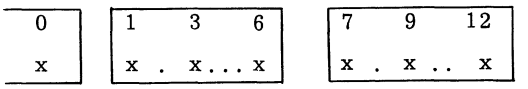
Nous allons montrer sur des exemples comment le dessin bien choisi d'un sondage systématique (au sens large) permet d'estimer les écarts carrés moyens entre unités de sondage γ_Δ dont les numéros d'ordre diffèrent d'une quantité donnée Δ.

On reconstituera σ^2 en pondérant convenablement les estimations des γ_Δ.

A - Fraction de sondage : $\frac{1}{2}$

Le sondage systématique strict nous renseigne sur les écarts de rangs $\Delta = 2, 4, 6, \ldots$ mais non sur les écarts impairs $\Delta = 1, 3, 5, \ldots$ Et il arrive souvent que $E(x_i - x_{i-1})^2$ soit très différente de $E(x_i - x_{i-2})^2$.

Au contraire, le dessin suivant nous informe complètement,



etc.

les écarts $(x_6 - x_7)^2$ sont du type $\Delta = 1$

$(x_1 - x_3)^2$ - $\Delta = 2$

$(x_3 - x_6)^2$, $(x_6 - x_9)^2$ - $\Delta = 3$

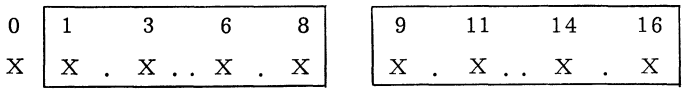
$(x_3 - x_7)^2$ - $\Delta = 4$

$(x_1 - x_6)^2$ - $\Delta = 5$

$(x_1 - x_7)^2$ - $\Delta = 6$

et bien entendu nous avons aussi les écarts correspondants à $\Delta \geqslant 7$.

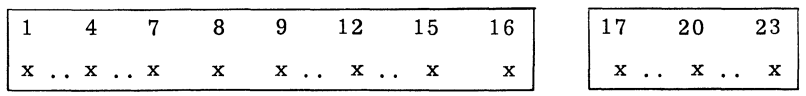
B - Autre dessin pour f = 1/2



$\Delta = 1$	$(x_9 - x_8)^2$
$\Delta = 2$	$(x_3 - x_1)^2$ et $(x_8 - x_6)^2$
$\Delta = 3$	$(x_3 - x_6)^2$ et $(x_9 - x_6)^2$
$\Delta = 4$	manque
$\Delta = 5$	$(x_6 - x_1)^2$, $(x_8 - x_3)^2$, $(x_{11} - x_6)^2$
$\Delta = 6$	$(x_9 - x_3)^2$

L'absence de $\Delta = 4$ n'est pas grave si l'on fait une hypothèse de variation régulière de $(x_i - x_{i-\Delta})^2$ avec Δ ; et ce dessin a l'avantage de la simplicité.

Voici un dessin qui procure toutes les valeurs de Δ (mais est assez compliqué) :



$\Delta = 1$	$(x_8 - x_7)^2$, $(x_9 - x_8)^2$, $(x_{16} - x_{15})^2$, $(x_{17} - x_{16})^2$
$\Delta = 2$	$(x_7 - x_9)^2$, $(x_{17} - x_{15})^2$
$\Delta = 3$	$(x_4 - x_1)^2$, $(x_7 - x_4)^2$, $(x_{12} - x_9)^2$, $(x_{15} - x_{12})^2$
$\Delta = 4$	$(x_8 - x_1)^2$, $(x_{16} - x_{12})^2$, $(x_{20} - x_{16})^2$
$\Delta = 5$	$(x_9 - x_4)$, $(x_{17} - x_{12})^2$

$\Delta = 6$	$(x_7 - x_1)^2$, $(x_{15} - x_9)^2$
$\Delta = 7$	$(x_8 - x_1)^2$, $(x_{15} - x_8)^2$, $(x_{16} - x_9)^2$, $(x_{23} - x_{16})^2$
$\Delta = 8$	$(x_9 - x_1)^2$, $(x_{15} - x_7)^2$, $(x_{16} - x_8)^2$, $(x_{20} - x_{12})^2$, $(x_{23} - x_{15})^2$
$\Delta = 9$	$(x_{16} - x_7)^2$, $(x_{17} - x_8)^2$

Malheureusement le cas $f = 1/2$ présente peu d'intérêt pratique. Essayons de réduire cette fraction de sondage.

C - Fraction de sondage : 1/3

Le sondage systématique (strict) nous renseigne seulement si les écarts de rangs sont $\Delta = 3 ; 6 ; 9 ; \ldots$

Le dessin suivant nous informe mieux, mais encore incomplètement.

```
0   | 1   3   6      12 |  | 13     15   18      24 |  etc.
x   | x...x...x..... x  |  |  x  .  x  ..  x..... x |
```

On trouve $(x_{13} - x_{12})^2$ du type $\Delta = 1$

$(x_3 - x_1)^2$ - - $\Delta = 2$

$(x_6 - x_3)^2$ et $(x_{15} - x_{12})^2$ - - $\Delta = 3$

$(x_6 - x_1)^2$ - - $\Delta = 5$

$(x_{12} - x_6)^2$ - - $\Delta = 6$

puis bien entendu $\Delta = 7, 9, 10, 11, 12$, etc.

On ne trouve pas d'écart du type $\Delta = 4$, $\Delta = 8$.

Le dessin ci-dessous a les défauts complémentaires :

```
| 0 |  |    2   3   6      12 |
| x |  | .  x   x  .. x..... x |
```

il ne nous informe pas sur les écarts du type $\Delta = 5$ ou $\Delta = 7$.

En combinant bout à bout les 2 dessins, on serait complètement informé mais au prix d'une complication indéniable. En pratique on peut admettre au contraire que le cas $\Delta = 4$ se comporte comme une moyenne des cas $\Delta = 3$ et $\Delta = 5$ (et $\Delta = 8$ comme une moyenne de $\Delta = 9$ et 7).

Autre dessin, pour f = 5/24

On va mettre bout à bout des intervalles de longueur Δ :

$$\Delta = 1, 2, 3, 6 \text{ et } 12$$

autrement dit :

```
i = | 1. 3 .. 6 ..... 12 .......... 24 |
```

On trouve ainsi des différences carrées correspondant en outre à

$$\Delta = 5,\ 9,\ 11 \text{ (et au-delà de 12).}$$

A mesure qu'on cherche à avoir f plus petit, les lacunes dans les valeurs de Δ se multiplient ; et il paraît vraisemblable qu'on ne peut pousser très loin le procédé sans un dessin excessivement compliqué.

Nous n'affirmerons pas enfin qu'on puisse estimer γ_Δ, $\forall \Delta$: la question est vraisemblablement liée à des propriétés arithmétiques de f ou $q = 1/f$.

11 - RETOUR A L'ECHANTILLON SYSTEMATIQUE (au sens strict)

Il ressort du § 10 que l'échantillon systématique (au sens large) est commode pour $q = 2$ ($f = 1/2$) et de moins en moins commode quand q grandit.

En pratique, on a beaucoup plus souvent affaire à $q = 20$ ou 100 qu'à $q = 2$ ou 3. Examinons les conséquences pratiques que peut avoir l'estimation de σ^2 par sondage systématique au sens strict.

Cas de l'exemple A :

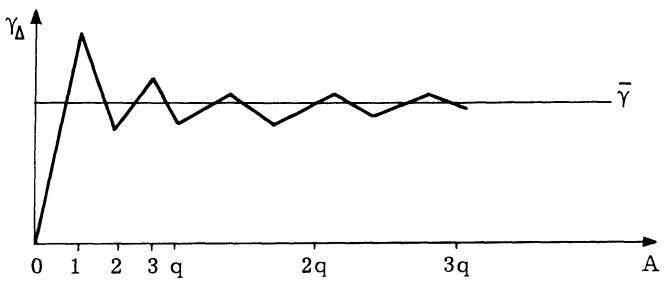


Figure 3.

Il est raisonnable alors (et dans bien d'autres cas) de penser que : si q est assez grand, γ_q, γ_{2q}, γ_{3q} ... diffèrent peu de :

$$\bar{\gamma} = \varepsilon(x_i - x_j)^2 = 2\,\frac{N\sigma^2}{N-1} \quad \text{(figure 3)}$$

$$\varepsilon(n^2 s^2) = \varepsilon\,[(n-1)\,g_q + (n-2)\,g_{2q} + (n-3)\,g_{3q} + \ldots]$$

d'où :

$$q^2\varepsilon(n^2 s^2) = (n-1)\,\gamma_q + (n-2)\,\gamma_{2q} + (n-3)\,\gamma_{3q} + \ldots$$

ou

$$\varepsilon(N^2 s^2) = q\,[(N-q)\,\gamma_q + (N-2q)\,\gamma_{2q} + (N-3q)\,\gamma_{3q} + \ldots]$$

comparé à :

$$N^2\sigma^2 = [(N-1)\gamma_1 + (N-2)\gamma_2 + \ldots + (N-q)\gamma_q] + [(N-q-1)\gamma_{q+1} + \ldots + (N-2q)\gamma_{2q}] + \ldots$$

Même si tous les γ étaient égaux entre eux, on aurait donc $\sigma^2 > \varepsilon s^2$. De façon plus précise,

$$n^2 \varepsilon s^2 = \frac{n(n-1)}{2} \bar{\gamma} \text{ , contre } N^2\sigma^2 = \frac{N(N-1)}{2} \bar{\gamma} \text{ ,}$$

donc :

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

L'égalité des γ n'est pas réaliste (bien entendu) et n'est supposée que pour obtenir un estimateur de σ^2.

Cas de l'exemple B : γ_Δ croît avec Δ.

Pour simplifier, supposons que $(N - \Delta)\ \gamma_\Delta$ croisse avec Δ.

On aurait donc

$$(n-1)\ \gamma_1 + (n-2)\ \gamma_2 + \ldots + (N-q)\ \gamma_q < q(N-q)\ \gamma_q = q^2(n-1)\gamma_q$$

$$(n-q-1\ \gamma_{q+1} + \ldots\ldots\ldots\ (N-2q)\ \gamma_{2q} < q(N-2q)\gamma_{2q} = q^2(n-2)\gamma_{2q}$$

$$\ldots\ldots\ldots\ldots$$

D'où il suivrait que $N^2\sigma^2 < N^2\varepsilon s^2$

ou $$\sigma^2 < \varepsilon s^2$$

Ceci est évidemment faux puisque $\sigma^2 = \sigma_1^2 + \sigma_e^2$, avec $\sigma_e^2 = \varepsilon s^2$.

C'est que, quand Δ grandit, $N - \Delta$ tend vers 0 et γ_Δ tend à plafonner, ce qui est contraire à l'hypothèse : $(N - \Delta)\gamma_\Delta$ croissant. Sans cette hypothèse grossière, on montrerait seulement que s^2 est abusivement grand comme évaluation de σ^2.

Cas de l'exemple C : γ_Δ a des oscillations périodiques non amorties (de période T). Si q coïncide avec T, 2T, les γ_q, γ_{2q}, ... seront tous très petits et s^2 sera abusivement petit ; ce sera l'inverse si q coïncide avec T/2, 3T/2, etc.

Conclusion : L'étude des exemples ci-dessus confirme l'intérêt que peut présenter la connaissance d'un échantillon systématique au sens large alors même que f est petit ; car cet échantillon fournira du graphe de γ_Δ un nombre de points beaucoup plus élevé que l'échantillon au sens strict ; on aura donc une connaissance au moins qualitative de la forme de ce graphe et des biais à attendre.

BIBLIOGRAPHIE

[1] GAUTSCHI W. - Some remarks on systematic sampling - A.M.S. 28, June 1957, p. 385/394.

[2] COCHRAN N.G. - Relative accuracy of systematic and stratified random samples for a certain class of populations - A.M.S. 17, (1946), p. 164/177.

[3] HAJEK J. - Some contributions to theory of probability sampling. Bull. Inst. Int. de Stat. 36 (1958) 127/134.