JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Analyse des distributions par sondage

Journal de la société statistique de Paris, tome 111 (1970), p. 170-177

<http://www.numdam.org/item?id=JSFS_1970__111__170_0>

III

ANALYSE DES DISTRIBUTIONS PAR SONDAGE

On se propose seulement de présenter quelques expériences (faites en 1967) en vue d'explorer un domaine inconnu de la théorie des sondages.

La *stratification* est une opération qui remonte au moins au XIX^{e} siècle, avec les fameuses *urnes de Poisson.*

Poisson étudiait seulement la *proportion* de boules d'une certaine couleur obtenue en tirant :

— soit n fois dans deux urnes de N boules,

— soit 2 n fois dans l'urne unique résultant du mélange des précédentes.

Il s'agissait de tirages bernoulliens (avec remise, c'est-à-dire d'une boule à la fois en remettant chaque fois la boule dans son urne d'origine).

Poisson constatait que le fait de *mélanger* les urnes *accroît la variance* d'échantillonnage.

En l'état actuel de la théorie classique, le résultat de Poisson a été étendu à une variable numérique quelconque X, ce qui comprend le cas binaire X = 1 ou 0, formalisation des deux couleurs des boules. Les urnes ou *strates* considérées sont en nombre quelconque. Il n'est pas nécessaire de supposer qu'elles ont toutes même taille N et même fraction sondée n/N, mais, avec des tailles différentes N_h on supposera les fractions sondées égales : $n_h/N_h = f$. L'échantillon est alors dit « représentatif ».

Les tirages exhaustifs peuvent sans inconvénients être substitués aux tirages bernoulliens.

La variance d'échantillonnage concernée est celle de la *moyenne* échantillon qui, dans le cas binaire, se réduit à la *proportion* ou fréquence, échantillon.

Le cas des échantillons « non-représentatifs » a été très étudié; avec la répartition dite « optimale » (NEYMAN, 1934) [1], la variance est encore bien meilleure qu'avec l'échantillon « représentatif ». Mais si la répartition est anti-représentative à contresens, il est facile d'obtenir une très mauvaise variance.

L'échantillon « représentatif » est en somme une *assurance tous risques.* Mais il n'est question dans tout cela que d'estimation de *moyenne* ou de *fréquence;* d'où l'on déduit l'estimation de la *masse* ou du *total* d'une variable étudiée, de l'*effectif* ou la *taille* d'un secteur de la population étudiée.

On désigne ces théories sous le nom de sondages *énumératifs.*

Les plus anciens et classiques professeurs (DEMING, YATES, COCHRAN, ...) distinguent les sondages *énumératifs* des sondages *analytiques* [2].

A vrai dire, une enquête déterminée est toujours polyvalente. Ses buts sont en partie *analytiques*, en partie *énumératifs.*

La distinction n'est donc pas dans les *objectifs* de l'enquête, mais dans le critère qu'a choisi le statisticien organisateur du plan d'échantillonnage, en adoptant telle ou telle statification, telle ou telle fraction de sondage.

Ces choix, ces décisions sont (bien sûr) largement conditionnés par les nécessités pratiques, le manque d'informations *a priori*, le désir de faire vite et bon marché, le modèle commode offert par les enquêtes antérieures.

Mais, au moins en théorie, ces choix visent à réduire au maximum certaines variances d'échantillonnage; ils sont inspirés par les théories des sondages *énumératifs*.

Pendant longtemps, on n'avait jamais entendu parler d'une enquête qui aurait été *organisée dans le but essentiel* d'*analyser*, c'est-à-dire d'informer sur la répartition d'une variable sur les diverses unités de sondage, — la répartition d'un couple de variable, d'un triplet de variables, — les indépendances ou contingences diverses: corrélations, associations entre variables : toutes choses pourtant fort connues du statisticien.

Les sondages *analytiques* étaient donc une fausse fenêtre, une vue de professeur puisque le praticien n'en avait jamais vu aucun.

Si les années 50 ont été marquées par Dalenius et sa stratification optimale, on peut penser que les années 60 l'auront été par Sedransk et les quelques *problèmes de sondage analytique* qu'il a su isoler et traiter jusqu'au bout [3].

Parmi les problèmes non résolus de sondage, Dalenius a depuis longtemps signalé l'étude des médianes, qui est restée embryonnaire [4].

Plus généralement, la connaissance d'une distribution par les *statistiques d'ordre :* médiane, quartiles, déciles, plus grande et plus petite valeurs, étendue ou « range », etc., a fait l'objet d'une littérature orientée vers les distributions *continues*, dont la densité admet des dérivées d'ordres 1, 2 ... et non vers les populations *discrètes* des sondages.

Il ne s'agit d'ailleurs pas de s'en tenir au sondage *élémentaire* (exhaustif ou bernoullien) qui n'offre guère d'intérêt pratique, ni de mystère.

Le moindre échantillon de l'enquête la plus banale est *stratifié*.

On n'a en fait aucune idée du caractère *faste ou néfaste* d'une stratification inspirée par la *tradition énumérative*, quand le but premier d'une enquête est au contraire analytique.

Tel est le domaine où nous avions tenté une exploration. Souhaitons que quelqu'un la poursuive *avec des moyens adéquats*.

I. — EXPÉRIENCE CONCERNANT LA PLUS GRANDE VALEUR

Matériau : On s'intéresse à une seule variable; elle ne prend que des valeurs distinctes :

$$a > b > c > d > e > f > g > h > \ldots$$

Par exemple 2 strates, de N = 4 unités chacune : n unités tirées de chaque strate.
Stratification Parfaite (S. P. pour abréger) :

Ceci désigne un cas comme :

Strate 1 : a b c d
Strate 2 : e f g h

Stratification Imparfaite (S. I. pour abréger) :

Ceci désigne un cas comme :

Strate 1 : a b d f
Strate 2 : c e g h

C'est le cas réaliste de stratification (car « la perfection n'est pas de ce monde »).
Stratification inexistante ou Absence de stratification (A. S. pour abréger)

On considère comme également probables tous les échantillons de taille $2n$ tirés de la population unique de taille 2N (résultant de la réunion des 2 strates).

On obtiendrait les mêmes résultats avec l'ensemble des stratifications au hasard de cette population. Ici il en existe :

$$C_8^4/2 = 35 \text{ (y compris S. P. et S. I.)}$$

Choix d'un critère :

On écarte *a priori* la variance comme critère approprié à une telle étude.

On retient :

— La proportion de cas favorables (à telle ou telle situation);

— La somme des écarts (tous de même signe) ou *perte d'information*, pour l'étude de la *plus grande* (*petite*) *valeur* (cf. notre thèse, 1958).

Quand ce sera utile, on admettra que : $a - b = b - c = c - d = \ldots = \Delta$, en remarquant que les résultats alors obtenus n'ont guère de sens *qu'en moyenne.*

Première expérience : $n = 1$; Plus grande Valeur — 28 *Échantillons :*

S. P. :	*e*	*f*	*g*	*h*
a	*a*	*a*	*a*	*a*
b	*b*	*b*	*b*	*b*
c	*c*	*c*	*c*	*c*
d	*d*	*d*	*d*	*d*

S. I. :	*c*	*e*	*g*	*h*
a	*a*	*a*	*a*	*a*
b	*b*	*b*	*b*	*b*
d	*c*	*d*	*d*	*d*
f	*c*	*e*	*f*	*f*

A. S. : *a* :	*b*	*c*	*d*	*e*	*f*	*g*	*h*
b :	*c*	*d*	*e*	*f*	*g*	*h*	
c :	*d*	*e*	*f*	*g*	*h*		
d :	*e*	*f*	*g*	*h*			
e :	*f*	*g*	*h*				
f :	*g*	*h*					
g :	*h*						

Premier résultat (évident) : Dans 1/4 des échantillons, *a est le plus grand des 2* et ceci pour S. P., pour S. I. et pour A. S.

C'est ensuite que les différences apparaissent :

b est cru le plus grand : dans 4 cas sur 16, pour S. P. et pour S. I.
dans seulement 6 cas sur 28 pour A. S.

etc. D'où le tableau :

Plus grande valeur :	*a*	*b*	*c*	*d*	*e*	*f*	*g*	Total
S. P.	4	4	4	4				16 cas
S. I.	4	4	2	3	1	2		16
A. S.	7	6	5	4	3	2	1	28

Mesurant en écarts l'erreur d'estimation de *Max x*, on obtient ainsi :

S. P. $[4(a-b) + 4(a-c) + 4(a-d)]/16 = 24\,\Delta/16 = 3\,\Delta/2$

S. I. $[4(a-b) + 2(a-c) + 3(a-d) + (a-e) + 2(a-f)]/16 = 31\,\Delta/16 \neq 2\,\Delta$

A. S. $[6(a-b) + 5(a-c) + 4(a-d) + 5(a-e) + 2(a-f) + (a-g)]/28 = 56\,\Delta/28 = 2\,\Delta$

Conclusion : *S. I. est presque aussi mauvaise que A. S. pour estimer le maximum.*

Encore a-t-on étudié une stratification imparfaite mais *favorable.* Envisageons S'. I. définie comme suit :

Strate 1 :	*a*		*c*	*d*		*f*	*g*
Strate 2 :		*b*		*d*	*e*		*h*

Les 16 échantillons correspondants se répartissent comme suit :

Maximum =	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	Total
	4	3	3	2	2	1	1	0	16

et nous mesurons l'erreur moyenne d'estimation (de *a*) par : $34\,\Delta/16 > 2\,\Delta$

De sorte que *S'. I. est plus mauvaise que A. S.*

Résultat : En matière de recherche du maximum (ou du minimum) la stratification avec fractions de sondage égales est *souvent nocive.*

Dénombrons les cas où elle est nocive

Écrivons les 35 stratifications possibles et, pour chacune, les 16 maxima, ce qui est assez lourd. Le seul moyen de contrôle est la moyenne des 35 erreurs moyennes, qu'on retrouve bien égale à $2\,\Delta$.

Synthèse des résultats :

Erreur moyenne : (mesurée en $\Delta/16$)	24	27	29	30	31	32	33	34
distribution des 35 cas	1	1	2	4	1	6	12	8

Ainsi il existe 20 stratifications (sur 35) qui sont pires que l'absence de stratification; 6 autres équivalent à la non-stratification. Outre S. P. et S. I., il existe 7 cas de stratification imparfaite mais efficace : c'est peu.

Détail des résultats (1re expérience)

Strate 1	Strate 2	Maximum (nb de fois) *a b c d e f g*	Erreur moyenne en $\Delta/16$	Strate 1	Strate 2	Maximum (nb de fois) *a b c d e f g*	Erreur moyenne en $\Delta/16$
a b c d	*e f g h*	4 4 4 4	24	*a c e f*		4 3 3 2 2 1 1	34
a b c e	*d f g h*	4 4 4 1 3	27	*a d e f*		4 3 3 2 2 2	33
a b c f	*d e g h*	4 4 4 1 1 2	29	*a d e g*		4 3 3 2 2 1 1	34
a b c g	*d e f h*	4 4 4 1 1 1 1	30	*a d e h*		4 3 3 2 2 1 1	34
a b c h		4 4 4 1 1 1 1	30	*a b f g*		4 3 3 2 2 1 1	33
a b d e		4 4 2 3 3	29	*a b f h*		4 4 2 2 2 1 1	33
a b d f		4 4 2 3 1 2	31	*a b g h*		4 4 2 2 2 2	32
a b d g		4 4 2 3 1 1 1	32	*a c f g*		4 3 3 2 2 1 1	34
a b d h		4 4 2 3 1 1 1	32	*a c f h*		4 3 3 2 2 1 1	34
a c d e		4 3 3 3 3	30	*a c g h*		4 3 3 2 2 2	33
a c d f		4 3 3 3 1 2	32	*a d f g*		4 3 3 2 2 1	34
a c d g		4 3 3 3 1 1 1	33	*a d f h*		4 3 3 2 2 1 1	34
a c d h		4 3 3 3 1 1 1	33	*a d g h*		4 3 3 2 2 2	33
a b e f		4 4 2 2 2 2	32	*a e f g*		4 3 3 3 1 1 1	33
a b e g		4 4 2 2 2 1 1	33	*a e f h*		4 3 3 3 1 1 1	33
a b e h		4 4 2 2 2 1 1	33	*a e g h*		4 3 3 3 1 2	32
a c e f		4 3 3 2 2 2	33	*a f g h*		4 3 3 3 3	30
a c e g		4 3 3 2 2 1 1	34				

AUTRE PROBLÈME : *Détection de la strate renfermant le maximum*

Profitons du travail déjà fait pour explorer le problème de la *détection* du vrai maximum.

En « épuisant l'urne », on serait certain de découvrir quel est le plus grand des 2N éléments; bien entendu on veut éviter d'en arriver là, et on prend certains risques.

Dans l'exemple précédent, *a* (l'élément maximal) est toujours dans la strate 1.

Combien de fois se tromperait-on si l'on admettait que cet élément se trouve là où l'on a tiré le plus grand des 2 éléments-échantillons?

Les erreurs proviennent de ce que la composition de la strate 1 n'est pas

a b c d (S. P.)

Strate 1	Proportion d'erreurs	Strate 1	Proportion d'erreurs	Strate 1	Proportion d'erreurs
$a\ b\ c\ d$	0	$a\ b\ e_2\ f_2$	4/16	$a\ d_2\ f_2\ g_2$	8/16
$a\ b\ c\ e_1$	1/16	$a\ b\ e_2\ g_2$	5/16	$a\ d_2\ f_2\ h_4$	9/16
$a\ b\ c\ f_2$	2/16	$a\ b\ e_2\ h_4$	6/16	$a\ d_2\ g_4\ h_4$	10/16
$a\ b\ c\ g_2$	3/16	$a\ c_1\ e_2\ f_2$	5/16	$a\ e_2\ f_2\ g_2$	9/16
$a\ b\ c\ h_4$	4/16	$a\ c_1\ e_2\ g_2$	6/16	$a\ e_2\ f_2\ h_4$	10/16
$a\ b\ d_1\ e_1$	2/16	$a\ c_1\ e_2\ h_4$	7/16	$a\ e_2\ g_4\ h_4$	11/16
$a\ b\ d_1\ f_2$	3/16	$a\ d_2\ e_2\ f_2$	6/16	$a\ f_4\ g_4\ h_4$	12/16
$a\ b\ d_1\ g_2$	4/16	$a\ d_2\ e_2\ g_2$	7/16		
$a\ b\ d_1\ h_4$	5/16	$a\ d_2\ e_2\ h_4$	8/16	*Récapitulation*	
$a\ c_1\ d_1\ e_1$	3/16	$a\ b\ f_2\ g_2$	6/16	1	0/16
$a\ c_1\ d_1\ f_2$	4/16	$a\ b\ f_2\ h_4$	7/16	1	1/16
$a\ c_1\ d_1\ g_2$	5/16	$a\ b_1\ g_4\ h_4$	8/16	2	2/16
$a\ c_1\ d_1\ h_4$	6/16	$a\ c\ g_2\ g_2$	7/16	24 3	3/16
		$a\ c\ f_2\ h_4$	8/16	4	4/16
		$a\ c_1\ g_4\ h_4$	9/16	4	5/16
				5	6/16
				4	7/16
				4	8/16
				3	9/16
				11 2	10/16
				1	11/16
				1	12/16
				35	

On peut donc dire que, avec pareille stratégie, il existe environ un tiers des stratifications pour lesquelles on se tromperait au moins la moitié du temps.

Globalement, pour l'ensemble des stratifications, la moyenne (de cette distribution symétrique) est $6/16 = 3/8$ d'erreurs.

Il revient au même de prendre au hasard 5 des 8 éléments : il y a 3 chances sur 8 que cet échantillon ne renferme pas l'élément *a*.

C'est ce qu'on ferait d'ailleurs si, après avoir tiré l'élément de chaque strate, ou urne, on décidait de vider l'une des 2 urnes, — bien entendu celle qui a donné le plus fort résultat à l'essai.

Ainsi cette stratégie ne serait *en rien fondée*, à moins d'avoir des raisons d'admettre une *homogénéité certaine* des strates ou urnes.

II. — EXPÉRIENCES CONCERNANT LA MÉDIANE OU L'INTERVALLE MÉDIAN

Expérience 1 bis

Nous reprenons les mêmes échantillons, mais nous nous intéressons à l'intervalle médian : (d, e).

Nous nous demandons si notre échantillon (stratifié) x, y forme un intervalle qui recouvre plus ou moins le précédent.

Ceci conduit à *classer* les échantillons comme suit :

Échantillons (par type)	Cas principaux					
	S P		S I	A S		
de		1	1		1	
ce, df		2	0		2	
cf	3	1	3	3	1	
be, dg		2			2	
ae, bf, cg, dh		4	2		4	
af, bg, ch		3	1		3	
ag, bh		2	2		2	
ah		1	1		1	
		16	10		16	
Autres échantillons		0	6		12	
			16		28	

Si l'on ne considère comme *bons échantillons* que *de*, *ce*, *df*, on a donc une probabilité de 3/16 avec S. P., 1/16 seulement avec S. I., *3/28 avec A .S.*

On peut user de critères fort différents. Par exemple les proportions 6/16 pour S. I. et 12/28 pour A. S. sont les fréquences des types d'échantillon qu'ignore une stratification parfaite.

A cet égard, on voit que *S. I. serait moins bon que A. S.*

Qu'en est-il pour les 35 stratifications possibles?

Mise à part la stratification parfaite S. P., il n'existe que 2 types de stratification :

le type (3 1, 1 3) et le type (2 2, 2 2)

Avec respectivement : 6 et 8 échantillons absents avec S. P.

Ainsi *toutes les stratifications autres que S. P.* imposent une proportion d'échantillons *aberrants* (: ne recouvrant pas *d-e*) supérieure à ce que donnerait l'absence de toute stratification.

Ce paradoxe se justifie par l'égalité :

$$(35 \times 16)^{-1} (0 \times 1 + 6 \times 16 + 8 \times 18) = 240/560 = 12/28$$

étant entendu (cf. 1re expérience) qu'on dénombre *16 cas* du type (31, 13)

soit 12, de *abce* à *acdh*, puis 4, de *aefg* à *afgh;*

contre 18 cas du type (2 2, 2 2) : de *abef* à *adgh.*

Conclusion : Nous venons de découvrir un problème concernant l'intervalle *médian* où il est tout à fait *déconseillé de stratifier* (vu qu'on n'a qu'une chance sur 35 d'atteindre le seul cas favorable).

Expérience 2 :

Nous modifions légèrement notre dispositif expérimental en faisant $n = 2$ au lieu de $n = 1$ dans chaque strate.

Alors chaque stratification donne naissance à $C_2^2 = 36$ échantillons, cependant que le cas A. S. comporte $C_8^4 = 70$ échantillons.

à savoir : les 36 échantillons du cas S. P. (chevauchant sur *d-e*) et *34 autres.*

Si l'on examine le cas S. I., le partage des échantillons se fait à raison de

18 échantillons appartenant à chacune des 2 classes;

et à nouveau 18/36 = 1/2 est plus grand que 34/70 (fort peu d'ailleurs).

Échantillons (par type)	Cas principaux		
	S P	S I	A S
par nature de l'intervalle médian			
de	9	5	9
ce *df*	12	4	12
cf	4	4	4
be *dg*	6	4	6
bf *cg*	4	0	4
bg	1	1	1
	36	18	36
Autres échantillons	0	18	8 34
		36	70

Détail de l'expérience :

S. P.							S. I.						
	ef	*eg*	*eh*	*fg*	*fh*	*gh*		*ce*	*cg*	*ch*	*eg*	*eh*	*gh*
ab	*be*						*ab*	*bc*					
ac							*ad*						
ad							*af*						
bc							*bd*						
bd							*bf*						
cd							*df*						

On pourrait étudier les 35 stratifications possibles (cf. Tableau 1re Expérience).

Expérience 3 :

On considère 6 nombres décroissants (strictement) répartis en 3 strates de 2 soit :

a b c d e f, Intervalle médian (*c*, *d*) ou *c-d*.

On tire $n = 1$ unité par strate; l'échantillon *xyz* a une médiane *y*.
A chaque stratification correspondent $2 \times 2 \times 2 = 8$ échantillons.
Nombre des stratifications : $5 \times 3 = 15$

(en effet, on peut mettre *a* avec l'une des 5 autres lettres, puis former 2 paquets de 2 lettres avec les 4 autres de 3 façons seulement.)

L'énumération des cas est donc facile :

Stratification parfaite : *ab cd ef*: Alors la médiane *y* est *c* ou *d*.
C'est aussi le cas pour *ae cd bf* dans 4 cas sur 8
af cd be dans cette même proportion.
Plus généralement, seront dits *favorables* les échantillons de la forme :
acz, *bcz* ($z = d, e, f$); *xde*, *xdf*, ($x = a, b, c$) : 12 cas
Cette définition est plus large que les 8 cas de la seule S. P.
Cas de A. S. Sur C_6^3 échantillons (soit 20), 12 seraient *favorables*.
Les cas S. I. Type : *ce, df, ab ;* 6 échantillons favorables sur 8.
ou *ca, db, ef ;*
En revanche : Type : *ca, de, bf :* 4 échantillons favorables sur 8.

Récapitulation :

15 { 1 stratification à 8/8 favorables ; *ab, cd, ef*
4 stratifications à 6/8 favorables ; *ca, db, ef* ; *ce, df, ab*
10 stratifications à 4/8 favorables *ae, cd, bf* ; *ca, de, bf* ; *cb, df, ae* ; *cb, de, af* ; *ca, df, be* ; *cb, da, ef* ; *cf, de, ab* ; *af, cd, be* ; *ce, da, bf* ; *cf, db, ae* ; *ce, db, af* ; *cf, da, be*

Moyenne : $72/15 \times 8 = 0{,}6 = 12/20$

Il ne semble pas qu'il serait très difficile de passer au cas de *m* strates renfermant chacune 2 nombres, puis de faire tendre *m* vers l'infini.

CONCLUSION

On peut retenir de ce travail exploratoire que le préjugé favorable dont bénéficie (depuis Poisson) l'échantillonnage dit représentatif n'est absolument pas fondé quand on s'intéresse à autre chose que la *moyenne de population.*

Toute stratification qu'on n'espère pas « presque parfaite », parce que ne reposant pas sur des informations supplémentaires très sérieuses, paraît présenter de sérieux risques, — du moins avec fractions de sondage égales.

Il conviendrait à présent d'explorer d'autres voies (notamment le sondage à 2 phases), nos conclusions présentes étant des plus négatives.

Il est à noter que, pour d'autres sondages analytiques, Sedransk s'est occupé du sondage à *plusieurs phases*, — procédé général pour disposer, à la 2e phase, d'informations supplémentaires *ad hoc.*

P. THIONET

BIBLIOGRAPHIE

[1] J. NEYMAN, Two different aspects of the representative method, *Jour. Royal Stat. Soc.*, 1934.

[2] W. G. COCHRAN, Sampling Techniques, 1961 (pp. 106, 108) par exemple.

[3] J. SEDRANSK, A double sampling scheme for analytic surveys, *J. A. S. A.*, *60* (1965), 985.

Idem, Designing some multifactor analytical studies, *ibidem*, 62 (1967), 1121.

G. BOOTH & J. SEDRANSK, Planning some two-factor comparative surveys, *ibidem*, 64, june 1969, 560-73.

[4] T. DALENIUS, Potential Research Objects in Sample Survey Theory and Methods, Colloquium on Applications of Mathematics in Economics, Budapest, 1963.