REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

Sur les sondages avec probabilités inégales

Revue de statistique appliquée, tome 17, n° 4 (1969), p. 5-44

<http://www.numdam.org/item?id=RSA_1969__17_4_5_0>

SUR LES SONDAGES AVEC PROBABILITÉS INÉGALES

P. THIONET

INTRODUCTION

Parmi les divers sujets traités dans la statistique la plus traditionnelle figurent les méthodes dont les statisticiens préconisent l'emploi pour constituer des échantillons ; avec ces méthodes d'échantillonnage va bien entendu de pair l'utilisation des dits échantillons en vue d'en extraire des informations valables sur les populations sondées.

Un préjugé(1) - qu'il faut sans cesse combattre tant il est enraciné, - est que les "bons" échantillons seraient les échantillons "représentatifs" ; rien n'est moins vrai bien entendu. Toute personne désireuse d'étudier un échantillon des 38 000 communes de France devine bien qu'il ne faut pas le constituer avec 1 grande ville sur 100, 1 petite ville sur 100, 1 commune rurale sur 100. Sans parler des très grandes villes (dont on inclut la liste complète dans les échantillons) il est à peu près intuitif qu'un bon échantillon doit correspondre une proportion beaucoup plus élevée des villes que des bourgades, -et a fortiori que des villages. Bien entendu une théorie (un peu technique) justifie cette façon de faire.

C'est ce qui est obtenu en fait au moyen de tirages au sort de communes avec des probabilités proportionnelles à leur population.

---

(1) Il est vrai que les méthodes d'échantillonnage sont peu connues, non seulement des utilisateurs mais aussi de théoriciens de la statistique mathématique, qui souvent oublient leur existence.

Dans beaucoup de problèmes (mais pas ici) il est sage de supposer les échantillons constitués d'éléments obtenus par tirages au sort indépendants, afin d'assimiler les valeurs-échantillons à des valeurs indépendantes (en probabilité) d'une variable aléatoire donnée. Il est alors naturel de convenir d'appeler "échantillon" un ensemble de n valeurs indépendantes d'une variable aléatoire donnée. Il est en revanche fâcheux qu'on puisse enseigner la chose comme suit : "En statistique, un échantillon désigne le système de n valeurs prises par une variable aléatoire au cours de n tirages indépendants".

Autrement dit nous nous étonnons qu'on puisse présenter un sens restreint du mot échantillon comme étant son seul sens correct. On laisse ainsi supposer que ce qui est d'usage dans beaucoup de chapitres de statistique (pour des raisons de facilité) serait la règle absolue (donc que le reste est erreur, pour peu que l'élève ait l'esprit dogmatique).

Plus généralement, un "univers" étant partagé en "unités de sondage", on a besoin d'en extraire un échantillon par tirage au sort avec probabilités inégales. Les techniques courantes de tirage au sort ont ceci de fâcheux qu'elles n'excluent pas le cas où la même unité serait tirée 2 fois ou davantage ; et le fait se produit (bien sûr) d'autant plus souvent que les unités sont plus probables (plus grosses).

Les ressources du bon sens sont, en cette affaire, insuffisantes à elles seules.

a) Si l'on partage l'univers en strates et qu'on tire une unité de chaque strate, l'accident fâcheux ne peut plus se produire, mais on n'a plus assez de renseignements (avec un tel échantillon) pour évaluer les variances d'échantillonnage, c'est-à-dire mesurer les erreurs d'échantillonnage (tout "calcul d'erreur" suppose qu'on ait des informations sur les écarts entre données concernant les unités de sondage de la même strate).

Si donc une strate fournit n unités échantillon, il est souhaitable, pour des raisons de "calculs de variance", d'avoir n = 2 au moins (et souvent même : n = 3).

b) Si l'on a décidé de tirer n = 2 unités et que le sort fournisse 2 fois la même unité, le calcul d'erreurs ne sera donc pas possible pour les strates où cet accident s'est produit. En outre les échantillons de familles ou de personnes, qu'on tirerait ensuite des communes échantillons, seraient perturbés par cet accident. En effet : l'échantillonnage de communes est presque toujours le 1er degré d'un sondage à plusieurs degrés de logements ou d'individus.

S'il était prévu de tirer m personnes par commune échantillon, soit 2m au total, il faudra soit tirer 2m personnes de la seule commune obtenue, soit n'en tirer que m mais leur affecter un poids double (par exemple les représenter par 2 cartes perforées identiques dans le dépouillement).

c) Si l'on devait tirer n = 3 unités et qu'on ait obtenu les unités A B B au lieu de A B C, on peut envisager de faire pour B seule ce qui vient d'être dit au § b) ci-dessus ; les deux procédés seraient corrects pour le dépouillement d'une enquête. Le second coûterait moins cher (m enquêtes de moins à faire) mais bien entendu informerait moins bien que le premier.

En revanche les calculs d'erreur seraient perturbés ; et qu'ils aient été programmés pour un calculateur humain ou électronique, il faudrait totalement les reprendre.

d) Il semble donc assez tentant de s'arranger, d'une façon ou d'une autre, pour que l'accident redouté ne se produise pas. C'est même ce qui arrive presque fatalement si le tirage de l'échantillon est confié à un employé "intelligent". Or ces errements à la longue sont assez redoutables ; car (sans qu'on s'en doute) ils équivalent à modifier totalement les probabilités de tirage, donc ils conduisent à des échantillons systématiquement déformés .

Il est facile de voir que, dans une strate, l'accident devrait arriver surtout aux unités les plus probables, c'est-à-dire aux communes les plus peuplées. En se débrouillant pour qu'on ne les tire pas 2 fois, l'employé trop intelligent réduit en fait leur probabilité de tirage, au profit des communes moins peuplées ; on finit à la longue par s'en apercevoir mais trop tard.

Nous arrivons à cette conclusion qu'en cette affaire une méthode scientifique, même un peu compliquée, sera préférable à la "débrouillardise" d'un agent. D'ailleurs quand le tirage au sort est confié aux soins des machines, le même problème est posé à un échelon plus élevé ; et la "débrouillardise" du chef du bureau d'échantillonnage sera au moins aussi pernicieuse que celle de son employé.

Tous ceux qui (comme nous) ont eu à organiser des enquêtes régulières par sondage ont rencontré fatalement le problème ; beaucoup l'ont résolu par le mépris, estimant (avec quelques motifs) que l'erreur d'échantillonnage est bien moins sérieuse que les erreurs de relevé faites sur le terrain par les enquêteurs et les erreurs faites par les enquêtés (de bonne foi ou non). Nous allons voir en somme qu'il est possible d'éliminer (avec beaucoup de mathématique) un type d'erreurs assez négligeables :

La mathématique ne peut pas grand'chose au contraire contre les erreurs les plus massives. Ce fait est assez déprimant, il faut le reconnaître. Quoi qu'il en soit, nous essaierons donc ici de faire le point sur de nombreuses recherches mathématiques dont le but est, en somme, d'éliminer une petite erreur d'échantillonnage.

Comme nous le signalons dans notre Communication à la Société de Statistique de Paris du 16 Novembre 1966 [1] il y a une disproportion incontestable entre ce but assez mince et ces moyens fort puissants. Dans un "speech" qu'il prononçait récemment aux Etats-Unis le professeur COCHRAN, actuel président de l'Institut International de Statistique, n'hésitait pas à choisir le problème des sondages avec probabilités inégales comme le type même du cas où (comme il dit, avec humour) l'on fait "trop de statistique mathématique" [2](1). A son avis, il eut été raisonnable que 3 ou 4 personnes étudient cette question (dans différents pays) ; or on en a déjà dénombré (paraît-il) 34 ; et chaque fois qu'il reçoit un nouveau fascicule d'une publication statistique, il appréhende d'y trouver une 35ème méthode. Plutôt que de lui causer cette peine, nous nous efforcerons ici de mettre un peu d'ordre dans nos lectures, qui (de toute façon) n'ont pas un caractère exhaustif.

La première difficulté que nous rencontrerons est le choix d'un ordre à peu près cohérent pour présenter les différents aspects d'une question assez embrouillée.

1 - Les procédés d'échantillonnage

Ils sont relativement faciles à exposer et peuvent intéresser un public assez large ; nous commencerons donc par là.

Nous distinguerons :

- ceux qui dérivent de l'urne de Bernoulli,
- le sondage de Hajek (qui dérive des urnes de Poisson),
- le sondage systématique,
- puis le tirage d'échantillons tournants (problèmes de Fellegi).

Enfin (parce qu'ils sont récents) parmi bien d'autres, les procédés de Hanurav dont l'intérêt théorique dépasse la portée pratique.

---

(1) Espérons qu'aucun lecteur ne prendra cette boutade au pied de la lettre.

Rien ne permet de dire qu'on n'emploie pas d'autres procédés[1], ni qu'on n'en découvrira pas d'autres. En particulier l'existence d'enquêtes à buts multiples incite à rechercher des échantillons polyvalents, problème qu'on ne peut actuellement tenir ni pour résolu, ni pour insensé.

Le résultat de ces méthodes est finalement une certaine distribution de probabilités, sur l'espace-échantillon (comme disent les statisticiens anglo-saxons) ou l'ensemble fondamental (comme disent les probabilistes). En particulier la même distribution peut résulter de deux méthodes distinctes. Nous devons à Hajek d'avoir replacé l'espace-échantillon au premier plan, alors qu'on avait tendance à l'oublier au dépens des méthodes de tirage au sort.

2 - Les probabilités - Les estimateurs et les variances

Le choix des formules mathématiques convenant à chaque échantillonnage est un sujet important mais très technique, ce qui explique qu'il soit reporté à la 2ème partie.

On trouve en outre quelque avantage à rapprocher et comparer des formules les unes des autres. Mais bien entendu il est un peu artificiel de séparer méthodes et formules.

3 - Les choix à faire. Les "stratégies"

Le choix reste à faire non seulement entre tel ou tel estimateur, entre telle ou telle méthode d'échantillonnage, mais entre le tirage sans remise avec probabilités inégales et d'autres méthodes : sondage avec remise et probabilités inégale ou encore emploi d'estimateurs spéciaux combiné au tirage sans remise avec probabilités égales.

Ce choix pourrait faire l'objet d'une 3ème Partie, mais on a finalement préféré renvoyér son étude à quelque article ultérieur, qui sera plus difficile à lire (et à écrire).

1ère PARTIE - LES PROCEDES D'ECHANTILLONNAGE

1 - INFLUENCE DE L'OPERATEUR

Tout d'abord il convient de bien voir que le comportement de l'opérateur au cours des tirages au sort a énormément d'importance quand il travaille à obtenir n unités de sondages distinctes (n étant un nombre imposé).

Considérons le cas d'un opérateur qui tire son échantillon comme Bernoulli avec des probabilités inégales, sans se soucier d'abord de savoir s'il peut obtenir 2 fois la même unité de sondage. Ensuite il se soucie de l'ordre qu'il a reçu de ne pas apporter à son chef deux fois la même. (Par exemple 2 fois le nom d'une ville).

---

(1) Le texte révisé de la présente étude a été déposé à la Revue de Statistique Appliquée avant la sortie du numéro de mars 1969 du J.A.S.A. où se trouve un article très intéressant de Raymond JESSEN [24]. Nous regrettons très vivement de n'avoir pu en rien dire, ni en tenir compte, d'autant plus que JESSEN fut un des pionniers dans la pratique des sondages.

2 - TIRAGES BERNOULLIENS AVEC REJETS (OU "REJECTIFS")

Une façon de faire[1] (qui peut être fort laborieuse) consiste à recommencer tous les tirages à partir de zéro, s'il apparait qu'on a tiré 2 fois la même unité. C'est le sondage "rejectif" ou avec rejet. Bien entendu, on peut s'en apercevoir bien avant d'avoir les n unités de sondage, et il faut s'arrêter aussitôt ; mais ceci ne modifie par le sondage, si c'est manifestement un gain de temps et d'argent.

Supposons qu'on dispose de 4 unités A B C D, affectées de probabilités suivantes :

$$A(0,1) \ ; \ B(0,2) \ ; \ C(0,3) \ ; \ D(0,4).$$

On désire en tirer 2, sans avoir jamais la même. Supposons qu'on procède d'abord à 2 tirages au sort Bernoulliens. L'échantillon peut avoir l'une des 10 compositions suivantes, leurs probabilités sont écrites au regard.

	Probabilités		Probabilités
A A	0,01	A B ou B A	0,04
B B	0,04	A C ou C A	0,06
C C	0,09	A D ou D A	0,08
D D	0,16	B C ou C B	0,12
Total	0,30	B D ou D B	0,16
		C D ou D C	0,24
			0,70

Cet ensemble des 10 éléments X Y est appelé ensemble fondamental ou espace-échantillon.

Il arrivera donc en moyenne 30 fois sur 100 que les 2 unités tirées au sort seront identiques. Si l'on convient de tout recommencer alors, et ceci autant de fois qu'il faudra, l'espace échantillon est (comme on dit) tronqué de ses éléments AA, BB, CC, DD. Et les probabilités des autres éléments sont regonflées proportionnellement pour que leur total soit porté à 1,00. Il convient donc de les multiplier par 10/7.

Résultat

	Probabilités	
A B (ou B A)	0,057	Sondage "rejectif"
A C (ou C A)	0,086	
A D (ou D A)	0,114	
B C (ou C B)	0,171	
D B (ou B D)	0,229	
C D (ou D C	0,343	
	1,000	

---

(1) Due à Durbin (1953) [3] semble-t-il. Le mot "rejectif" est dû à Hajek (1964) [4].

3 - TIRAGES BERNOULLIENS SUCCESSIFS

Un tout autre comportement pour un opérateur (un peu paresseux) consistera manifestement à poursuivre les tirages Bernoulliens jusqu'à obtention de n unités distinctes, dut-on pour cela effectuer (n + 1), (n + 2), (n + 3)... tirages.

Ceci correspond à un type général de sondage appelé successif par Hajek [4].

Bien entendu les sondages Bernoulliens successifs[1] ne conduisent pas à la même distribution de probabilités que les sondages Bernoulliens avec rejet.

Montrons quelle serait la probabilité de AB ou BA, dans l'exemple précédent.

On peut obtenir A, x fois de suite, puis enfin B ; ou l'inverse. Voici les probabilités de tels évènements.

$$AAA\ ..\ AB\ :\ P = (0,1)(0,2) + (0,1)^2(0,2) + \ldots + (0,1)^x(0,2) + \ldots$$

$$= (0,02)\,[1 + (0,1) + (0,1)^2 + (0,1)^3 + \ldots]$$

$$= (0,02) / (1 - 0,1) = 0,02/0,9$$

$$BBB\ \ldots\ A\ :\ P = 0,02/(1 - 0,2) - 0,02/0,8$$

		au lieu de
Ensemble AB :	$0,02\left(\frac{10}{9} + \frac{10}{8}\right) = 0,047$	0,057
De même		
AC :	$0,03\left(\frac{10}{9} + \frac{10}{7}\right) = 0,076$	0,086
AD :	$0,04\left(\frac{10}{9} + \frac{10}{6}\right) = 0,111$	0,114
BC :	$0,06\left(\frac{10}{8} + \frac{10}{7}\right) = 0,161$	0,171
BD :	$0,08\left(\frac{10}{8} + \frac{10}{6}\right) = 0,233$	0,229
CD :	$0,12\left(\frac{10}{7} + \frac{10}{6}\right) = 0,372$	0,343
	1,000	1,000

Le lecteur qui serait satisfait de la concordance approximative des résultats est invité à ne pas poursuivre la lecture du présent article ; nous supposerons au contraire qu'une telle discordance éveille l'intérêt et pique la curiosité.

4 - TIRAGES BERNOULLIENS EN CASCADE[1]

On peut convenir de tirer d'abord une unité avec les probabilités : 0,1 - 0,2 - 0,3 - 0,4. Après quoi le second tirage sera fait en excluant l'unité déjà tirée. Appliquons le théorème des probabilités composées.

		Probabilité	Résultat	Probabilité
1e tiré A	2e tiré B	2/9	AB	2/90
proba 0,1	C	3/9	AC	3/90
	D	4/9	AD	4/90
1e tiré B	2e tiré A	1/8	BA	2/80
proba 0,2	C	3/8	BC	6/80
	D	4/8	BD	8/80
1e tiré C	2e tiré A	1/7	CA	3/70
proba 0,3	B	2/7	CB	6/70
	D	4/7	CD	12/70
1e tiré D	2e tiré A	1/6	DA	4/60
proba 0,4	B	2/6	DB	8/60
	C	3/6	DC	12/60

En regroupant AB ou BA :

$$\frac{2}{90} + \frac{2}{80} = 0,02 \left(\frac{10}{9} + \frac{10}{8}\right) = 0,047$$

Conclusion

On obtient ainsi les mêmes résultats qu'avec les sondages successifs, ce que rien le laissait (au fond) prévoir.

5 - UNE METHODE JAPONAISE

Une méthode (connue depuis une vingtaine d'années) attribuée souvent à Midzuno (1952) [6], consiste à tirer la première des n unités échantillons avec des probabilités inégales et la seconde avec d'égales probabilités.

Pour le même exemple, on obtiendrait les probabilités suivantes :

2e tirage / 1er tirage	A	B	C	D
A : 1/10	-	1/30	1/30	1/30
B : 2/10	2/30	-	2/30	2/30
C : 3/10	3/30	3/30	-	3/30
D : 4/10	4/30	4/30	4/30	-

---

(1) YATES et GRUNDY (1953) [5]

Récapitulation

	A B ou B A	A C ou C A	A D ou D A	B C ou C B	B D ou D B	C D ou D C	Total
Probabilité :	$\frac{1+2}{30}$	$\frac{1+3}{30}$	$\frac{1+4}{30}$	$\frac{2+3}{30}$	$\frac{2+4}{30}$	$\frac{3+4}{30}$	$\frac{30}{30}$
ou	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{5}{30}$	$\frac{6}{30}$	$\frac{7}{30}$	$\frac{30}{30}$
ou	0,100	0,133	0,167	0,167	0,200	0,233	1,000

Bien entendu les probabilités de tirage d'un couple donné d'unités sont (on le voit) totalement modifiées par rapport aux procédés précédents. Mais ce n'est pas là un défaut à proprement parler (voir § II-5-2).

6 - SONDAGE STRATIFIE DE COCHRAN [7]

Cochran (avec Hartley et Rao) a proposé une méthode qui paraissait très intéressante (mais s'est révélée, après étude, peu efficace).

1/ On partage la population de n strates, ici en 2 ; c'est-à-dire qu'on tire au sort entre :

I) A B, C D ; ou II) A C, B D ; ou III) A D, B C

2/ On tire au sort 1 unité de chaque strate, avec probabilités proportionnelles

I		II		III	
A	0,1	A	0,1	A	0,1
B	0,2	C	0,3	D	0,4
	0,3		0,4		0,5
C	0,3	B	0,2	B	0,2
D	0,4	D	0,4	C	0,3
	0,7		0,6		0,5

I		II		III	
Echantillons	Probabilités	Echantillons	Probabilités	Echantillons	Probabilités
A C	$\frac{1}{3}$ $\frac{3}{7}$	A B	$\frac{1}{4}$ $\frac{2}{6}$	A B	$\frac{1}{5}$ $\frac{2}{5}$
A D	$\frac{1}{3}$ $\frac{4}{7}$	A D	$\frac{1}{4}$ $\frac{4}{6}$	A C	$\frac{1}{5}$ $\frac{3}{5}$
B C	$\frac{2}{3}$ $\frac{3}{7}$	C B	$\frac{3}{4}$ $\frac{2}{6}$	D B	$\frac{4}{5}$ $\frac{2}{5}$
B D	$\frac{2}{3}$ $\frac{4}{7}$	C D	$\frac{3}{4}$ $\frac{4}{6}$	D C	$\frac{4}{5}$ $\frac{3}{5}$

Récapitulation

Probabilités

A B	$\frac{1}{3}\ \frac{1}{4}\ \frac{2}{6} + \frac{1}{3}\ \frac{1}{5}\ \frac{2}{5} = \frac{1}{36} + \frac{2}{75} = 0{,}028 + 0{,}027 = 0{,}055$
A C	$\frac{1}{3}\ \frac{1}{3}\ \frac{3}{7} + \frac{1}{3}\ \frac{1}{5}\ \frac{3}{5} = \frac{1}{21} + \frac{1}{25} = 0{,}047 + 0{,}040 = 0{,}087$
A D	$\frac{1}{3}\ \frac{1}{3}\ \frac{4}{7} + \frac{1}{3}\ \frac{1}{4}\ \frac{4}{6} = \frac{4}{63} + \frac{1}{18} = 0{,}064 + 0{,}055 = 0{,}119$
B C	$\frac{1}{3}\ \frac{2}{3}\ \frac{3}{7} + \frac{1}{3}\ \frac{3}{4}\ \frac{2}{6} = \frac{2}{21} + \frac{1}{12} = 0{,}095 + 0{,}083 = 0{,}178$
B D	$\frac{1}{3}\ \frac{2}{3}\ \frac{4}{7} + \frac{1}{3}\ \frac{4}{5}\ \frac{2}{5} = \frac{8}{68} + \frac{8}{75} = 0{,}127 + 0{,}107 = 0{,}234$
C D	$\frac{1}{3}\ \frac{3}{4}\ \frac{4}{6} + \frac{1}{3}\ \frac{4}{5}\ \frac{3}{5} = \frac{1}{6} + \frac{4}{25} = 0{,}167 + 0{,}160 = 0{,}327$
	1,000

Il y a une assez nette ressemblance entre ces résultats et ceux des sondages Bernoulliens avec rejet ou successifs.

7 - TIRAGES DE POISSON AVEC REJETS (OU REJECTIFS)

Passons aux sondages de Poisson avec rejets, dûs à Hajek [4]

On décide de procéder à 1 tirage oui-non pour chaque unité de la population, afin de décider si elle fera ou non partie de l'échantillon.

Mais on rejette l'échantillon et on recommence les opérations si le résultat obtenu n'est pas exactement n = 2 unités distinctes.

Ici encore, il n'est pas toujours nécessaire d'achever les N tirages pour savoir que n dépasse 2 ou n'atteindra pas 2, - donc pour rejeter l'échantillon ; et (quand c'est possible) on a intérêt, pour être plus vite renseigné, à procéder aux tirages dans l'ordre des p_i décroissants.

Exemple

Les probabilités initiales étant encore $p_i = (0,1 - 0,2 - 0,3 - 0,4)$ et le nombre d'unités à tirer étant n = 2, on choisit d'affecter à l'unité i la probabilité ($n\,p_i$) d'être tirée ; ceci suppose qu'un $n p_i$ ne dépasse 1. (Si on a $n p_i = 1$, on inclut l'unité i automatiquement dans l'échantillon).

On constate ainsi que l'échantillon doit être rejeté 57 fois sur 100. En divisant les probabilités des cas conservés par 0,4304, on trouve les résultats définitifs ci-dessous. On voit que les résultats sont totalement différents de ceux des méthodes Bernoulliennes.

Remarque

Hajek n'a pas envisagé de sondages de Poisson successifs (ce serait fort compliqué et les résultats dépendraient profondément de l'ordre dans lequel serait rangée la population pour procéder aux tirages).

Les résultats possibles (= les éléments de l'espace échantillon) sont alors :

A B C D	Ω	Probabilité	Sans rejet	Avec rejet
. . . .	Ø	$8.6.6.2.10^{-4}$		
+ . . .	A	$2.6.4.2.10^{-4}$		
. + . .	B	$8.4.4.2.10^{-4}$		
. . + .	C	$8.6.4.8.10^{-4}$		
. . . +	D	$8.6.4.8.10^{-4}$		Résultats définitifs
+ + . .	A B	$2.4.4.2.10^{-4}$	$= 64.10^{-4}$	0,0148
+ . + .	A C	$2.6.6.2.10^{-4}$	$= 144.10^{-4}$	0,0334
+ . . +	A D	$2.6.4.8.10^{-4}$	$= 384.10^{-4}$	0,0893
. + + .	B C	$8.4.6.2.10^{-4}$	$= 384.10^{-4}$	0,0893
. + . +	B D	$8.4.4.8.10^{-4}$	$= 1024.10^{-4}$	0,2379
. . + +	C D	$8.6.6.8.10^{-4}$	$= 2304.10^{-4}$	0,5353
+ + + .	A B C	$2.4.6.2.10^{-4}$		
+ + . +	A B D	$2.4.4.8.10^{-4}$		
+ . + +	A C D	$2.6.6.8.10^{-4}$		
. + + +	B C D	$8.4.6.8.10^{-4}$		
+ + + +	A B C D	$2.4.6.8.10^{-4}$		
			4304.10^{-4}	1,0000
+ = oui				
. = non				

(Ω_1 : A B, A C, A D, B C, B D, C D)

Dans la colonne Ω se lisent les $2^4 = 16$ éléments de l'espace-échantillon Ω. Le sous-espace Ω_1 (6 éléments) est celui qui est conservé par la procédure des rejets ; c'est Ω tronqué).

8 - TIRAGES SYSTEMATIQUES

Il a été beaucoup question des sondages systématiques ces dernières années (de 1962 à 1966), encore que l'invention en revienne à GOODMAN et KISH (1950) [8].

Les articles récents sont de Hartley, Rao, Connor [9, 10, 11, 12].

Reprenons l'exemple de la population (A B C D) avec probabilités individuelles (0,1 - 0,2 - 0,3 - 0,4).

Il existe autant de sondages systématiques que de rangements distincts des 4 lettres.

Par rangements, il faut entendre permutations circulaires compte non tenu du sens dans lequel elles sont écrites. Soit (n - 1)/2 = 3 dans le cas présent ; ces 3 rangements sont :

I A B C D (D C B A, A D C B, etc...)

II A C B D (D B C A, A D B C, etc...)

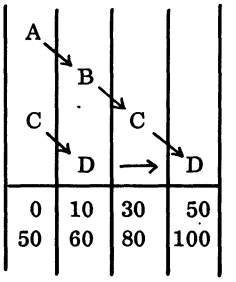
III A B D C (C D B A, A C D E, etc...)

En faisant la moyenne de ces 3 cas, considérés comme équiprobables, on obtient ce qu'on appelle le sondage systématique "randomisé" (randomized systematic sampling).

Cas I

	Probabilités	Cumulées	
A	0,1	0,10	0,0 à 0,10
B	0,2	0,30	0,10 à 0,30
C	0,3	0,60	0,30 à 0,60
D	0,4	1,00	0,60 à 1,00
			Positions des deux index

Echantillons



Ainsi : 3 échantillons seulement sont possible A C probabilité 0,20
B C " 0,40
C D " 0,40

Explication

Un index est supposé descendre la colonne des probabilités cumulées de 0 à 0,50 puis à 1,00 et un second index simultanément à la même vitesse, de 0,50 à 1,00, puis de 0 à 0,50. Le premier index allant de 0 à 0,10 désigne A, alors que le second descend à 0,50 à 0,60 désignant C.

Il se produit un double décrochage :

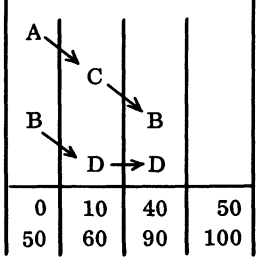
Quand l'index va de 0,10 à 0,30, il désigne B (au lieu de A). Alors le 2ème index va de 0,60 à 0,80 c'est-à-dire désigne D (au lieu de C).

Enfin l'index va de 0,30 à 0,50 désignant C (au lieu de B), tandis que le 2ème index désigne toujours D (de 0,80 à 1,00).

Cas II

	Probabilités	Cumulées
A	0,1	0 à 10
C	0,3	10 à 40
B	0,2	40 à 60
D	0,4	60 à 100

Echantillon



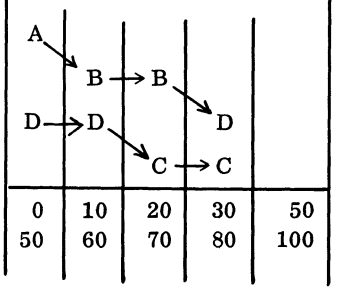
Résultats

3 échantillons	A B	Probabilités 0,20
	C D	" 0,60
	B D	" 0,20

Cas III

	Probabilités	Cumulées
A	0,1	0 à 10
B	0,2	10 à 30
D	0,4	30 à 70
C	0,3	70 à 100

Echantillon



0	10	20	30	50
50	60	70	80	100

Résultats

4 échantillons	A D	Probabilités 0,20
	B D	" 0,20
	B C	" 0,20
	D C	" 0,40

Récapitulation

	A B	A C	A D	B C	D B	C D	
Systématique I	-	0,2	-	-	0,4	0,4	1,00
II	0,2	-	-	-	0,2	0,6	1,00
III	-	-	0,2	0,2	0,2	0,4	1,00
Total	0,2	0,2	0,2	0,2	0,8	1,4	3,00
Systématique "randomisé" (moyenne)	0,067	0,067	0,067	0,067	0,266	0,466	1,00
Comparaison avec sondages successifs	0,047	0,076	0,111	0,161	0,233	0,371	1,00

On voit que : les déformations par rapport aux cas précédents sont considérables.

Remarques

1/ Le sondage systématique fournit des unités-échantillons distinctes sans qu'on ait à intervenir : c'est le procédé le plus simple, le plus naturel, pour ce faire.

2/ Le seul cas d'impossibilité serait celui où la population (ou plutôt la strate) comprendrait une très grosse unité, affectée d'une probabilité p supérieure à 1/n ; alors il conviendrait de retirer cette unité de sa strate

et d'en faire une strate à elle seule. En somme c'est qu'on aurait commis une erreur de stratification avant de tirer l'échantillon (erreur réparable).

3/ Mais le sondage systématique "randomisé" est un mythe ; ou bien les unités de sondage sont rangées dans un ordre alphabétique, topographique, administratif, etc.... ou bien on a essayé de les ranger dans un ordre privilégié, par exemple par rang de tailles croissantes (ou décroissantes). Jamais l'ordre de rangement n'est tiré au sort, dans la pratique.

4/ Quand les probabilités de tirage sont égales entre elles, le sondage systématique "randomisé" équivaut mathématiquement au sondage bernoullien sans remise : mêmes probabilités d'inclusion de 1, 2, 3, ... unités. En revanche nous voyons sur l'exemple précédent que, si les probabilités sont inégales, il n'y a plus identité mais simple ressemblance entre ces procédés de tirage d'un échantillon.

5/ Lorsqu'on calcule les variances d'échantillonnage (cf. II § 10 ci-après), on retrouve le même phénomène : les termes les plus importants des variances concordent, mais il existe aussi des termes différents.

Quant à la variance d'un échantillonnage systématique donné (: non "randomisé") on éprouve la sérieuses difficultés à s'en faire une idée, ce qui est d'ailleurs déjà un peu le cas quand les unités sont équiprobables (cf. [12]).

9 - LE PROBLEME DES ECHANTILLONS TOURNANTS

La construction d'échantillons en vue d'enquêtes permanentes ou périodiques pose des problèmes (à la fois théoriques et pratiques) qui sortent du cadre de la présente étude et mériteraient un article spécial. Disons seulement qu'il n'est ni possible ni souhaitable de conserver longtemps le même échantillon.

Au lieu de le renouveler en totalité (chaque fois ou à certaines dates), on améliore la comparabilité des résultats dans le temps (et les conditions matérielles d'enquête sur le terrain) en renouvelant progressivement (et par fraction) le dit échantillon.

En fait il s'agit généralement d'échantillons à 2 degrés, composés de logements sis dans certaines communes. Si le renouvellement progressif a lieu au niveau des logements, il n'en vient pas moins un moment où toutes les familles d'une même commune ont été soumises (f fois) à l'enquête et où il est indispensable de changer de commune.

Si une strate est représentée dans l'échantillon par 2 communes, on préfère ne pas les changer toutes deux à la fois (si possible). Ici entre le jeu le fait que les communes de tailles inégales sont tirées au sort avec d'inégales probabilités.

Bien entendu ces tailles sont en outre variables dans le temps, surtout dans certains pays neufs où la population est très mobile. On s'efforce de tenir à jour les listes des localités et leur population la dernière en date. Nous n'en tiendrons guère compte ici.

Un fait important (à notre avis) est que (en supposant les enquêtes également espacées dans le temps), il convient que les communes figurent dans

l'échantillon pendant un laps de temps proportionnel à leur taille (supposée invariable) : c'est une sorte de représentativité de l'échantillon au cours du temps.

En effet, s'il était procédé à un très grand nombre de tirages au sort (disons un chaque jour) et qu'on change d'échantillon chaque fois, chaque commune (ou unité) ferait partie de l'échantillon pendant un laps de temps qui au total serait à peu près proportionnel à sa taille. Il semble utile que ce phénomène soit (artificiellement) conservé quand on change rarement d'échantillon, en fixant à chaque unité un temps de présence dans l'échantillon qui soit proportionnel à la taille.

Dès lors, il est commode d'adopter le mode de tirage systématique grâce auquel en outre aucune duplication n'est possible. L'ordre de rangement des unités (de la strate) étant fixé, on fera défiler les unités à travers l'échantillon à vitesse constante, à la manière d'un film de cinéma. Après un passage complet, on pourra adopter un nouvel ordre de rangement.

Exemple

Unité :	A	B	C	D	Total	T = 100 mois d'enquête
Proba :	0,1	0,2	0,3	0,4	1,00	

On se reportera au § 8 ci-dessus. On voit que 3 cas sont possibles :

Cas I Ordre A B C D On partira de la date t = 10 : débuter avec B et D (ou 60) (finir avec AC)

Cas II Ordre A C B D On partira de la date t = 10 : débuter avec C et D (ou 60) (finir avec AB)

Cas III Ordre A B D C On est en présence d'un ruban sans fin.

Le cas III diffère donc totalement des cas I et II ; on ne peut y entrer à aucun moment sans commencer (pour une première phase) par écourter le temps de séjour d'une des unités. En contrepartie, à aucun moment il n'y aura de renouvellement des 2 unités simultanément.

Il s'est produit simplement ceci :

$$\text{Cas I et II} \begin{cases} \text{B et C} & 0,2 + 0,3 = 0,5 \\ \text{A et D} & 0,1 + 0,4 = 0,5 \end{cases}$$

On pouvait partager la population en 2 moitiés de même taille ; et l'ordre choisi permettait de partager la liste d'unités de la même façon (cas I et II). Ce phénomène peut se reproduire pour bien d'autres listes (encore que ce soit en quelque sorte un évènement de probabilité 0).

10 - LES ECHANTILLONS TOURNANTS DE FELLEGI [13] (1963)

FELLEGI a étudié (dans sa thèse, 1961) divers schémas d'échantillonnage, dont certains sont tournants ; nous ignorons s'ils ont plus ou moins droit-de-cité au Dominion Bureau of Statistics du Canada, où il travaille. Son article est assez long. Son apport essentiel semble avoir été mis en lumière par BREWER [14] (1967), statisticien Australien : il s'agit d'une pos-

sibilité de calculer des probabilités de travail dont on se sert pour tous les tirages au sort (d'une certaine unité de sondage) se succédant dans le temps. Elles sont choisies de façon à assurer à chaque unité de sondage et pour chaque tirage une probabilité fixée à l'avance d'être prise dans l'échantillon (autrement dit : une probabilité d'inclusion π_i/n au $t^{ième}$ tirage, (t = 1, 2,... n), indépendante de t).

Nous examinerons plus loin cette question (au II § 7).

Pour le moment, disons que Fellegi n'emploie pas le sondage systématique, mais une méthode du type en cascade du I § 4. Il ne paraît pas avoir rencontré de cas où les 2 unités échantillons quittaient ensemble l'échantillon (cas I et II du 9 ci-dessus).

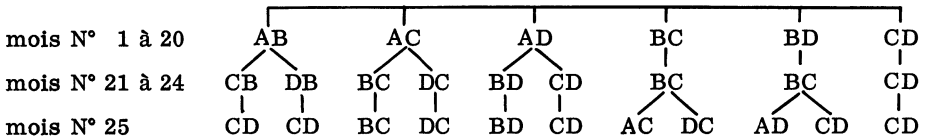
Indiquons l'un de ses exemples :

Une strate comprend 4 unités A B C D, dont les probabilités d'inclusion devraient être comme suit :

$$A(0,20) - B(0,24) - C(0,26) - D(0,30)$$

Donc : Durées de présence : 20 mois ; 24 mois ; 26 mois ; 30 mois (pour 100 mois d'enquête).

On programme tout ce qu'il peut se passer au moins jusqu'au 25ème mois inclus. Enumérons tous les échantillons possibles (sous forme d'un arbre) :



Il existe donc 11 cas possibles. On peut facilement calculer leurs probabilités respectives, en supposant les tirages effectués (à la date 20 et à la date 24) avec des probabilités proportionnelles à (0,20 - 0,24 - 0,26 - 0,30) entre choix permis.

Dès lors il nous a semblé qu'on n'avait plus qu'à tirer au sort l'une des 11 branches de l'arbre et qu'on obtenait ainsi un plan de sondage pour 26 mois (au moins).

11 - LES PROCEDES DE HANURAV (et autres)

Il n'y a aucune raison pour que les procédés décrits ci-dessus soient les seuls possibles. Il est à tout instant possible qu'un mathématicien imagine un nouveau procédé de tirage d'échantillon, jouissant de propriétés (théoriques) intéressantes ; il n'est pas du tout certain qu'un praticien adopte jamais ce procédé.

Il peut arriver d'ailleurs que le procédé implique un accroissement du coût des opérations sur le terrain (voir notre exemple dans [1] page 18 .

Mais le plus vraisemblable est que le procédé demande un certain volume de calculs supplémentaires et ait un aspect déroutant pour le praticien.

HANURAV [15], 1967, a décrit quelques nouveaux procédés de tirage de 2 unités de sondage sans remise.

L'un a des mérites certains, quant à la précision des sondages. Les 2 autres ont l'intérêt de procurer facilement des probabilités d'inclusion données d'avance. Malheureusement ils n'ont un aspect simple que si les 2 unités de sondage les plus grandes ont même probabilité. Comme il n'y a aucune raison pour qu'il en soit ainsi, les procédés s'adaptent à la situation réelle par un artifice assez compliqué.

Acceptons donc l'hypothèse : $p_{n-1} = p_n$ (les unités u_1, u_2, ..., u_n étant rangées dans l'ordre des probabilités croissantes).

Schéma A : Hanurav adopte le procédé Bernoullien "rejectif" (cf. ci-dessus § 2) ; mais, à chaque reprise des tirages, il élève les probabilités de tirage p_i au carré. Très rapidement, les 2 seules unités ayant des chances réelles d'être tirées seront u_{n-1} et u_n. Il est remarquable que ce procédé fournisse finalement :

1/ Un échantillon ayant les probabilités d'inclusion $2p_1$, $2p_2$, ... $2p_n$ (ce que nous établirons en Annexe) ;

2/ Une variance d'échantillonnage très faible ;

3/ Une estimation de variance jamais négative,
c'est-à-dire un ensemble de performances rarement réunies.

Schéma A' : On tire la 1ère unité parmi u_1, u_2, ..., u_{n-1} avec des probabilités d_1, d_2, ..., d_{n-1} ; soit u_i l'unité tirée.

On tire la 2ème unité parmi u_{i+1} u_{i+2} ..., u_n avec les probabilités égales entre elles.

Les probabilités d_i se calculent en fonction des π_i par des formules très simples en $\delta_i = \pi_i - \pi_{i-1}$ qui d'ailleurs sont fausses dans [15](1).

Schéma A" : On tire les 2 unités comme ci-dessus, mais la 2ème avec des probabilités inégales proportionnelles aux π_i.

Les nouvelles probabilités d_i adéquates se déduisent des π_i de proche en proche.

Nous avons décrit ces schémas parce qu'on vient de publier l'article de Hanurav. D'autres comme celui de STEVENS [16] (1958) étaient beaucoup plus simples mais sont un peu oubliés. On devait (pour Stevens) altérer légèrement les probabilités π_i de façon à constituer la population de paires d'unités de même taille ; dès lors si une unité était par hasard tirée 2 fois, on n'avait qu'à lui associer son double (de même taille). (Les calculs de variance en résultant étaient assez compliqués).

Bien d'autres procédés ont été décrits, notamment dans Sankhya et dans des publications qu'on ne lit guère en France, sans oublier les communications aux divers congrès de statistique. Il n'est aucunement question de prétendre qu'ils ont moins d'intérêt que les précédents. (Voir notamment PATHAK [17] (1964)).

---

(1) Sur ce point et quelques autres, le rectificatif est intervenu [25] postérieurement au dépôt de la présente étude.

On se sentirait davantage concerné, cependant, si l'un de ces procédés répondait (même mal) à un problème essentiel (comme c'est le cas pour les échantillons tournants). Comme type de problème rarement abordé, citons (un peu au hasard) les échantillons à plusieurs fins.

12 - L'ECHANTILLON A FINS MULTIPLES

Il semble que le sondage de Poisson-Hajek convienne pour former des échantillons destinés à 2 enquêtes différentes, (1) et (2).

Soit π_i et τ_i les probabilités d'inclusion pour la même unité i.

Une classe très étroite d'unités de sondage est celle où $\pi_h = \tau_h + \varepsilon$; les unités tirées tirées de cette sous-strate seront bivalentes (classe III).

Distinguons plutôt les classes I : $\pi_i > \tau_i$; et II : $\pi_i < \tau_i$.

Classe I

Nous jouons le sort de l'unité i, à la loterie $(\pi_i,\ 1 - \pi_i)$; si l'unité i est retenue (pour l'enquête 1), nous jouons son sort une 2ème fois à la loterie :

$$\left(\frac{\tau_i}{\pi_i},\ 1 - \frac{\tau_i}{\pi_i}\right)$$

Ainsi nous aurons obtenu :

n unités pour les enquêtes 1 et 2,

et

m unités pour la seule enquête 1.

Classe II

Nous jouons le sort des unités i', à la loterie $(\tau'_i,\ 1 - \tau'_i)$; puis pour celles retenues, à la loterie

$$\left(\frac{\pi'_i}{\tau'_i},\ 1 - \frac{\pi'_i}{\tau'_i}\right)$$

D'où

n' unités pour les enquêtes 1 et 2

et

m' unités pour la seule enquête 2.

Classe III

Nous jouerons le sort des unités, à la loterie $(\pi_h,\ 1 - \pi_h)$; soit ν le nombre des unités tirées (pour les enquêtes 1 et 2).

Finalement l'échantillon comprendra :

$\nu + n + n'$ unités bivalentes,

m unités pour (1) seule,

m' unités pour (2) seule.

Taille probable de l'échantillon : Il vient

$E\ \nu = \Sigma\ \pi_h$	
$E(m + n) = \Sigma\ \pi_i$	$E\ n\ = \Sigma\ \tau_i$
$E(m' + n' = \Sigma\ \tau_i,$	$E\ n' = \Sigma\ \pi_i,$

$$E(\nu + n + n' + m + m') = \Sigma\ \sup\ (\pi_j, \tau_j)$$

(l'indice j regroupe h, i, i').

Dans l'esprit du § 7, on pourrait (par des sondages rejectifs) vouloir obtenir un échantillon ayant exactement la composition théorique :

$$\nu^\circ = E\ \nu\ ;\ n^\circ = E\ n\ ;\ n'^\circ = E\ n'\ ;\ m^\circ = E\ m\ ;\ m'^\circ = E\ m'$$

Il serait souhaitable que cette méthode fut expérimentée et étudiée.

Il parait vraisemblable, cependant, qu'elle ne soit pas pratique. Il est alors tentant de vouloir procéder comme suit.

1/ Un tirage bernoullien opéré sur la base des probabilités sup (π_i, π_i) fournira un échantillon de taille N. Pour éviter les duplications, on procédera en fait à un tirage systématique.

2/ Des tirages de Poisson effectués ensuite dans l'échantillon N (considéré comme une population) permettront de dire pour chaque unité i si l'on y effectuera les 2 enquêtes, ou l'enquête (1) seule, ou l'enquête (2) seule.

Tout l'appareil mathématique relatif à ce procédé d'échantillonnage (qui combine les schémas de Bernoulli et de Poisson) devrait être réinventé.

2ème PARTIE - FORMULATION MATHEMATIQUE

1 - PRINCIPES GENERAUX

Le but premier d'un échantillonnage est de permettre certaines estimations sur échantillons.

1.1 - Le plus souvent, le tirage de l'échantillon est suivi d'opérations de mesures, d'observations ou d'évaluation, donnant les valeurs x_i prises par une certaine variable x sur les unités échantillon u_i. A partir de cet échantillon (x_i) se calcule une estimation $\widehat{X}$ du total X des x_i (de la population échantillonnée). Dans le cas présent on doit aussi connaître les valeurs y_j (et leur total Y) d'une variable y appelée taille ; y_j est la "taille" de l'unité u_j (échantillon ou non).

Souvent les tailles sont "normées", c'est-à-dire ramenées à un total Y = 1 (il s'agit alors d'une distribution de probabilités p).

1.2 - Le tirage de l'échantillon avec des probabilités proportionnelles aux y_j se justifie dans la mesure où l'on a des raisons de croire x plus ou moins proportionnel à y.

Les rapports $r_i = x_i/y_i$ doivent alors être peu dispersés autour de $R = X/Y$; de telle sorte que leur moyenne (échantillon)

$$\frac{1}{n}\sum_{(n)}\frac{x_i}{y_i} = \sum_{(n)}\frac{x_i}{ny_i}$$

soit proche de X/Y, même si n est petit (disons n = 2).

1.3 - Les tirages avec remise sont pleinement justifiés si la population comprend un nombre d'unités N suffisamment grand, alors qu'aucune unité n'a une taille beaucoup plus grande que la moyenne : le risque de tirer plusieurs fois la même unité U_i dans de telles conditions est très faible.

1.4 - En revanche, ce risque n'est plus du tout négligeable (et les tirages sans remise s'imposent) dans le cas d'une population (ou strate) comprenant quelques très grosses unités, et si au total N n'est pas beaucoup plus grand que n. Toutefois aucune unité u_i ne devrait être grande au point qu'on ait :

$$\frac{y_i}{Y} \geq \frac{n}{N} \quad \left(\text{ou } y_i \geq \frac{n}{N}\text{, si } Y = 1\right)$$

Ce serait le signe d'une mauvaise définition des populations (ou strate) et des unités de sondages.

1.5 - Les tirages sans remise font intervenir simultanément la probabilité d'inclusion d'une unité u_j dans l'échantillon (soit π_j) et la probabilité de tirage de cette unité u_j (soit p_j).

En vue d'analyser les liens entre π_j et p_j, revenons d'abord sur les tirages classiques (avec remise), où la probabilité d'inclusion sera désignée par P_j :

Dans le cas de n tirages bernoulliens (avec remise), on aurait :

(Probabilité d'exclusion :) $1 - P_j = (1 - p_j)^n$;

donc, pour n = 2 :

$$P_j = 2p_j - p_j^2$$

Comme $\Sigma p_j = 1$, il vient

$$\Sigma P_j = 2 - \Sigma p_j^2 \leq 2 - \frac{1}{N}$$

Le nombre probable d'unités distinctes tirées est ΣP_j (le nombre lui-même étant un aléatoire prenant la valeur 2 ou 1).

La probabilité d'inclusion d'un couple d'unités (u_j, u_k), d'un triplet d'unités, etc... se définit aussi. On a ainsi, quand n est égal à 2 :

$$P_{jk} = 2p_j p_k$$

Quand on passe aux tirages sans remise, on réaliserait un certain idéal de simplicité si l'on pouvait obtenir :

$$\text{Probabilité d'inclusion de (j)} \quad : \pi_j = 2p_j$$

$$\text{" \quad " \quad de (jk)} \quad : \pi_{jk} = 2p_j p_k = \frac{\pi_j \pi_k}{2}$$

On s'aperçoit vite que c'est parfairement utopique. Que peut-on dire qui soit général ?

1.6 - Procédés successifs

D'abord, parler de la probabilité de tirage de u_j n'a guère de sens :

Si l'on procéde à des tirages successifs, il existe seulement une probabilité $p_j(t)$ de tirer j au $t^{ième}$ tirage : (t = 1, 2, ..., n).

Définissons $\zeta(j, t) = 1$ en cas de tirage et 0 en cas de non-tirage (: indicateur).

Que les tirages soient indépendants ou non, on posera :

$$E \sum_j \zeta(j, t) = \sum_j E\, \zeta(j, t) = \sum_j p_j(t) = n(t) \quad \Big| \quad \sum_t n(t) = n$$

$$E \sum_t \zeta(j, t) = \sum_t E\, (j, t) = \sum_t p_j(t) = \pi_j \quad \Big| \quad = \sum_j \pi_j$$

Lors de tirages bernoulliens avec ou sans remise, $n(t) = 1$; et $\sum_j \zeta(j, t) = 1$, $\forall$ t.

Avec les tirages sans remise, $\sum_t \zeta(t, t) = 1$ ou 0 seulement.

Alors π_j est une probabilité (tandis que, pour les tirages avec remise, π_j est seulement une espérance mathématique) :

π_j est la probabilité d'inclusion de u_j (sans duplication possible).

On a $\sum_j \pi_j = n$ bien entendu.

De même : soit l'indicateur $\zeta(j, k, t) = \begin{cases} 1 \text{ en cas de tirage du couple (j k)} \\ 0 \text{ en cas de non-tirage " "} \end{cases}$

Supposons n = 2, t = 1, 2. La probabilité d'inclusion de (j k) est :

$$\pi_{jk} = E \sum_t \zeta(j, k, t) = E\, \zeta(j, k, 2) \text{ ; car } \zeta(j, k, 1) = 0$$

Formules générales

$$\pi_{jj} = 0 \quad ; \quad \pi_{kj} = \pi_{jk}$$

$$\pi_j = \sum_k \pi_{jk} \quad ; \quad \pi_k = \sum_j \pi_{jk}$$

Mais il n'existe pas de formule générale donnant π_{jk} en fonction de π_j et π_k.

1.7 - Procédé de Hajek-Poisson

a) sans restriction

Soit P_j la probabilité de tirage de l'unité j, qui est aussi sa probabilité d'inclusion.

Le nombre n d'unités tirées est aléatoire, mais en moyenne on a

$$E\ n = \sum_1^N P_j$$

C'est pourquoi, si l'on désire avoir E n = 2 (comme dans le cas de Bernoulli) on est amené à adopter $2p_j$ comme probabilité de tirage Poissonnien (si p_j désigne la probabilité de tirage Bernoullien, avec $\Sigma\, p_j = 1$).

On a :

$$P_{jk} = P_j\, P_k = 4\, p_j\, p_k$$

On note que l'inclusion de (u_i) et celle de (u_j) dans l'échantillon S sont des évènements indépendants.

b) Toutefois, en rejetant tous les échantillons qui n'ont pas pour taille n = 2, on perturbe ces relations : on n'a pas $\pi_j = p_j$, et pas davantage $\pi_{jk} = \pi_j\, \pi_k$.

On a seulement

$$\pi_{jk} \begin{cases} \div\ P_{jk} & \text{sur} \quad \Omega_1 \qquad \text{(espace échantillon tronqué)} \\ = 0 & \text{sur} \quad \Omega - \Omega_1 \end{cases}$$

et

$$\pi_j = \sum_k \pi_{jk}$$

1.8 - Autres procédés

Dans le cas de Midzuno (§ 5) : $p_j\,(1) = p_j\ p_j(2) = (N-1)^{-1}$, $p_j\,(3) = (N-2)^{-1}$ etc... Dans les autres procédés courants, il n'est pas question de tirages successifs. Certains procédés peu courants sont examinés au § 6.

2 - PROBABILITES D'INCLUSION SUIVANT LE PROCEDE DE TIRAGE SANS REMISE EMPLOYE

Nous avons repris l'exemple des 4 unités de sondage (A B C D), avec n = 2. Les probabilités p_i qui leur sont associées, sont en fait nos $p_j(1)$, probabilités de tirage au 1er tirage.

$$p_j = (0,1 - 0,2 - 0,3 - 0,4)$$

Rassemblons dans un même tableau les probabilités d'inclusion de A, B, C, D.

On les obtient (à partir des résultats concernant les couples) comme suit :

Exemple : Bernoullien rejectif

A B	0,057	A B	0,057	A C	0,086	A D	0,114
A C	0,086	B C	0,171	B C	0,171	B D	0,229
A D	0,114	B D	0,229	C D	0,343	C D	0,343
π_A	0,257	π_B	0,457	π_C	0,600	π_D	0,686

Procédés	unité j =	A	B	C	D	Total
	n p_j (1) =	0,200	0,400	0,600	0,800	2,000
Bernoullien : rejectif		0,257	0,457	0,600	0,686	2,000
successif		0,234	0,441	0,609	0,716	2,000
Midzuno		0,400	0,467	0,533	0,600	2,000
Hartley-Rao-Cochran	I	0,333	0,667	0,429	0,571	2,000
	II	0,250	0,333	0,750	0,667	2,000
	III	0,200	0,400	0,600	0,800	2,000
Moyenne		0,261	0,467	0,592	0,680	2,000
Poisson-Hajek		0,137	0,342	0,658	0,863	2,000
Systématique	I	0,2	0,4	0,6	0,8	2,000
	II	0,2	0,4	0,6	0,8	2,000
	III	0,2	0,4	0,6	0,8	2,000
Systématique "randomisé"		0,2	0,4	0,6	0,8	2,000

Les formules mathématiques correspondantes sont, pour n = 2 :

Procédé : Bernoullien rejectif $\pi_j = 2p_j\,(1 - p_j)\,(1 - \sum_i p_i^2)^{-1}$ (découle de π_{jk})

successif $\pi_j = p_j\,(1 + S) - p_j^2\,(1 - p_j)^{-1}$ (découle de π_{jk})

avec $S = \sum_i p_i\,(1 - p_i)^{-1}$

Procédé Midzuno $\pi_j = [(N-2)\,p_j + 1](N-1)^{-1}$

$= p_j + (1 - p_j)\,\frac{1}{N-1}$ (direct)

Hartley-Rao-Cochran $\pi_j = p_j \left(\sum_{I_1} p_i\right)^{-1}$,

$i \in I_1$: indices de la 1ère strate (direct)

Si l'unité j est dans la 1ère strate

Poisson-Hajek ("rejectif")	$\pi_j = 2q_j\left(\sum_i q_i\right)$	découle de π_{jk} : $r_j = 2p_j(1 - 2p_j)^{-1}$ $R = \sum r_i$ $q_j = r_j(R - r_j)$
Systématique	$\pi_j = 2p_j$	direct

3 - INCLUSION D'UN COUPLE (j, k) DANS L'ECHANTILLON

La probabilité d'inclusion du couple (j, k) peut aussi s'obtenir dans chaque cas (toujours moyennant n = 2).

Bernoullien rejectif $\pi_{jk} = 2\, p_j p_k \left(1 - \sum p_i^2\right)^{-1}$ (direct)

successif $\pi_{jk} = p_j p_k \left[(1 - p_j)^{-1} + (1 - p_k)^{-1}\right]$ (direct)

Midzuno $\pi_{jk} = (p_j + p_k)\,(N - 1)^{-1}$ (direct)

Hartley-Rao-Cochran $\pi_{jk} = 0$ si j et k sont dans la même strate (direct)

$$= p_j p_k \left(\sum_{I_1} p_i\right)^{-1} \left(\sum_{I_2} p_i\right)^{-1},\ j \in I_1,\ k \in I_2$$

Poisson-Hajek (rejectif") $\pi_{jk} = r_j r_k \left(\sum_{\Omega_1} r_h r_i\right)^{-1}$, avec $r_i = \dfrac{2p_i}{1 - 2p_i}$ (direct)

Ω_1 désigne le sous-ensemble de Ω pour lequel 2 unités ont été effectivement tirées.

Systématique : $\pi_{jk} = 0$ en général $\leqslant \min(\pi_j, \pi_k)$ — varie suivant la place de k par rapport à j dans le rangement des unités adopté.

4 - FORMULES GENERALES DE : π_j π_{jk} π_{jkl} (n quelconque)

On pourra se reporter utilement à l'article de POMPILJ (en italien) [18], (1961).

On a :

$$\pi_j = \frac{n}{N} p_j \frac{m^{[j]}_{\{n-1\}}}{m_{\{n\}}}$$

$$\pi_{jk} = \frac{n(n-1)}{N(N-1)} p_j p_k \frac{m^{[jk]}_{\{n-2\}}}{m_{\{n\}}}$$

où : $m_{\{n\}}$ est le moment combinatoire (terminologie de GINI) d'ordre n de $p_1, p_2, \ldots p_N$, c'est-à-dire la moyenne de tous les produits possibles $p_{i_1} p_{i_2} \cdots p_{i_n}$.

Autrement dit, $m_{\{n\}}$ est le "polykey" : $< 11 \ldots 1 > = < 1^n > = k_{1^n}$

De même : $m^{[j]}_{\{n-1\}}$, $m^{[jk]}_{\{n-2\}}$, désignent une moyenne analogue des produits de (n - 1) (ou n - 2) termes p_i prélevés dans l'ensemble des p_α amputé d'abord de l'élément p_j, puis de l'élément p_k, etc....

1/ Cas particulier

$p_\alpha = \frac{1}{N}$, on retrouve bien $\pi_j = \frac{n}{N}$, $\pi_{jk} = \frac{n(N-1)}{N(N-1)}$, $\pi_{jkl} = \frac{n(n-1)}{N(N-1)} \frac{(n-2)}{(N-2)}$

car

$$m_{\{n\}} = N^{-n}$$

$$m^{[j]}_{\{n-1\}} = N^{-n+1} \quad \text{etc.....}$$

2/ Cas particulier : n = 2

$$m_{\{n\}} = \frac{\Sigma' p_i \, p_k}{N(N-1)} = \frac{1 - \Sigma p_i^2}{N(N-1)}$$

$$m^{[j]}_{\{n-1\}} = \frac{1 - p_j}{N - 1} \quad ; \quad m^{[jk]}_{\{n-2\}} = 1$$

$$\Longrightarrow \quad \pi_j = \frac{2p_j (1 - p_j)}{N(N-1)} \cdot \left[\frac{1 - \Sigma p_i^2}{N(N-1)} \right]^{-1}$$

$$= 2p_j (1 - p_j) \, [1 - \Sigma p_i^2]^{-1}$$

$$\pi_{jk} = \frac{2}{N(N-1)} p_j \, p_k \cdot \left[\frac{1 - \Sigma p_i^2}{N(N-1)} \right]^{-1}$$

$$= 2p_j \, p_k \, [1 - \Sigma p_i^2]^{-1}$$

On reconnait les expressions de π_j, π_{jk}, dans le cas bernoullien rejectif, le seul que Pompilj ait envisagé (sous le nom d'extraction en blocs).

5 - PROBLEME : OBTENIR UN PROCEDE DE TIRAGE AYANT DES PROBABILITES D'INCLUSION DONNEES.

5.1 - Les formules exprimant π_j en fonction de p_i, données aux § 2 et 4, sont rarement faciles à inverser. Or on a besoin en fait des p_i en fonction des π_j. Le but poursuivi est, en effet, (voir II 1.3) de pouvoir poser

$$\pi_j = \frac{2y_j}{Y} \quad \left(= \frac{ny_j}{Y} \right)$$

les tailles, y_j, étant connues et à peu près proportionnelles aux x_j inconnus.

En posant : $p_j = y_j / Y$, donc $p_j = \pi_j / 2$, on atteint le but cherché avec les procédés de sondage systématique. Avec les autres procédés on s'égare.

Ceux de Hajek et Cochran semblent inextricables. Dans le cas de Midzuno, l'inversion est très commode ; c'est la raison d'être de ce procédé.

5.2 - Cas MIDZUNO

$$\pi_j = \frac{N-2}{N-1} p_j + \frac{1}{N-1} \quad \text{(pour } n = 2\text{)}$$

Cette formule équivaut à :

$$p_j = \frac{N-1}{N-2} \pi_j - \frac{1}{N-2}$$

Comme p_j ni π_j ne peuvent être négatifs, ceci implique : $\pi_j \geq (N-1)^{-1}$. On ne peut donc (et c'était facile à prévoir) obtenir par ce procédé des probabilités π_j très petites ; on ne peut donc l'employer (disons) avec la "population" constituée par les communes rurales françaises où figurent sur la même liste des communes de tailles très variables ; il faudrait au préalable procéder à certains regroupements de communes, trop petites pour mériter au 2ème coup une probabilité $(N-1)^{-1}$.

Ce fut là notre objection essentielle contre cette méthode [19] (1954)

Remarque : Le phénomène ne fait que s'aggraver quand n augmente.

Exemple :

n = 3 $$\pi_j = p_j + (1 - p_j)\left[\frac{1}{N-1} + \left(1 - \frac{1}{N-1}\right)\frac{1}{N-2}\right]$$

c'est-à-dire

$$\pi_j = \frac{N-3}{N-1} p_j + \frac{2}{N-1}$$

$$p_j = \frac{N-1}{N-3} \pi_j - \frac{2}{N-3} \qquad \boxed{\pi_j \geq \frac{2}{N-1}}$$

autrement dit

$$\frac{\pi_j}{3} \geq \frac{2/3}{N-1} \text{ remplace } \frac{\pi_j}{2} \geq \frac{1/2}{N-1}$$

5.3 - Cas Bernoullien "successif"

Yates et Grundy [5] ont proposé (dès 1953) de résoudre par itérations le système d'équations en p_j.

$$\pi_j = p_j \left[\frac{1 + S - p_j}{(1 - p_j)}\right]$$

où

$$S = \frac{\sum_i p_i}{(1 - p_j)}$$

Si N est un peu grand, ce n'est guère praticable sans un ordinateur. Encore s'agit-il du cas n = 2. Les formules pour n = 3 et plus sont inextricables.

5.4 - Cas Bernoullien "rejectif"

Le système d'équations s'écrit :

$$p_j^2 - p_j + \frac{\pi_j}{2}(1 - \Sigma p_j^2) = 0 \quad ; \quad j = 1, 2, \ldots, N$$

on peut donc le récrire

$$\begin{cases} (1) & z_j^2 - z_j + \frac{\pi_j}{2}\lambda = 0 \quad ; \quad j = 1, 2, \ldots N \\ (2) & z_j = p_j \\ (3) & \Sigma p_j^2 = 1 - \lambda \\ (4) & \Sigma p_j = 1 \end{cases}$$

En général[1] les équations du 2ème degré (1) ont 2 racines de moyenne 1/2, comprises entre 0 et 1.

On essaiera de partir de $z_j^\circ = \pi_j/2$ et de résoudre par itération.

Les difficultés sont analogues à celles de 5.3.

6 - SUITE DU PROBLEME : PROCEDES DE HANURAV [15] (cf I § 11 ci-dessus)

Si l'on emploie le schéma A, on réalise automatiquement les probabilités d'inclusion désirées : Nous l'établirons en Annexe. Examinons les autres schémas.

6.1 - Schéma A'

Données $\quad \pi_1 \leqslant \pi_2 \leqslant \ldots\ldots\ldots \leqslant \pi_{N-1} = \pi_N$

$$\delta_1 = \pi_1 \quad ; \quad \delta_j = \pi_j - \pi_{j-1} \; ; \; \delta_N = 0$$

Les équations à inverser sont :

$$\begin{array}{l|l} \pi_1 = d_1 & \delta_1 = d_1 \\ \pi_2 = \frac{d_1}{N-1} + d_2 & \delta_2 = \left(\frac{1}{N-1} - 1\right) d_1 + d_2 \\ \pi_3 = \frac{d_1}{N-1} + \frac{d_2}{N-2} + d_3 & \delta_3 = \left(\frac{1}{N-2} - 1\right) d_2 + d_3 \\ \ldots\ldots\ldots & \ldots\ldots\ldots \\ \pi_{N-1} = \frac{d_1}{N-1} + \frac{d_2}{N-2} + \ldots + d_{N-1} & \delta_{N-1} = \left(\frac{1}{2} - 1\right) d_{N-2} + d_{N-1} \end{array} \quad \text{d'où}$$

---

(1) On peut s'en tenir en cas de $\pi_j \leqslant \frac{N}{2(N-1)}$ pour simplifier la discussion.

d'où

$$\begin{cases} d_1 = \delta_1 \\ d_2 = \left(\frac{N-2}{N-1}\right) d_1 + \delta_2 \\ d_3 = \left(\frac{N-3}{N-2}\right) d_2 + \delta_3 \\ \cdots\cdots\cdots\cdots \end{cases} \qquad \begin{cases} d_1 = \delta_1 \\ d_2 = \frac{N-2}{N-1}\,\delta_1 + \delta_2 \\ d_3 = \frac{N-3}{N-1}\,\delta_2 + \delta_3 \\ \cdots\cdots\cdots\cdots \end{cases}$$

Telle est l'expression cherchée des d_1, d_2, ... en fonction des δ_1, δ_2, ...

Remarque : lorsque δ_N n'est pas nul, on remplace π_N par $(\pi_N - \delta_N)$ et π_j par $\frac{\pi_j}{(1 - \delta_N)}$.

En outre (et pour compenser) on ajoute (dans une proportion convenable) des échantillons du type (j, N) : j = 1, 2, ..., N - 1, le tout conforme à l'intuition , mais absolument rigoureux.

6.2 - Schéma A"

Toujours par calcul direct, on obtient :

$$\begin{cases} 2p_1 = \pi_1 = d_1 \\ 2p_2 = \pi_2 = d_1 \frac{p_2}{1 - p_1} + d_2 \\ 2p_3 = \pi_3 = d_1 \frac{p_3}{1 - p_1} + d_2\, 1\, \frac{p_3}{-p_1 - p_2} + d_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots \end{cases}$$

d'où :

$$\begin{cases} d_1 = p_1 . 2 \\ d_2 = p_2\left(2 - \frac{d_1}{1 - p_1}\right) \\ d_3 = p_3\left(2 - \frac{d_1}{1 - p_1} - \frac{d_2}{1 - p_2 - p_3}\right) \end{cases}$$

Ceci donne, de proche en proche, $d_1(p_1)$, $d_2(p_1, p_2)$, $d_3(p_1, p_2, p_3)$.....

7 - LE PROCEDE FELLEGI (TIRAGES BERNOULLIENS SUCCESSIFS)

L'unité j peut n'être pas du tout tirée (probabilité $1 - \pi_j$), ou être tirée au 1er tirage (probabilité $p_j(1)$), ou au 2ème (probabilité $p_j(2)$) etc... ou au nième ($p_j(n)$). On a (voir ci-dessus II 1.6) :

Probabilité d'inclusion de l'unité j =

$$= \pi_j = \sum_1^n p_j(t) \quad \text{(il s'agit d'évènements incompatibles).}$$

On décide que les $p_j(t)$ seront égaux à p_j quel que soit t.

On calcule alors des probabilités de travail : $P_j(1)$, $P_j(2)$, ... $P_j(n)$ telles que le tirage n° t est effectué sur la file des $P_j(t)$ où sont barrées les lignes des unités déjà sorties. Il vient pour le 1er tirage :

$$P_j(1) = p_j = \pi_j / n$$

Puis les $P_j(2)$, $P_j(3)$.... sont définies par un système d'équations.

Par exemple, soit n = 2 : au 2ème tirage on posera $P_j(2) = \hat{P}_j$. On a :

$$\boxed{p_j = \sum_i \frac{\hat{p}_j}{1 - \hat{p}_j} p_i}$$

$i \neq j$ (i désignant l'unité sortie au 1er tirage)

ou

$$p_j = \left[\sum_1^n \frac{p_i}{1 - \hat{P}_i} - \frac{p_j}{1 - \hat{P}_i} \right] \hat{p}_j$$

d'où

$$P_j^2 - \hat{P}_j + p_j \frac{N - 1}{N R} = 0$$

ou

(F) :
$$\hat{P}_j = \frac{1}{2} - \sqrt{\frac{1}{4} - p_j \frac{N - 1}{N R}} \quad \text{avec} \quad R = \frac{1}{N} \sum \frac{p_i}{\hat{P}_i}$$

On commence avec $R_o = 1$, d'où P_j°, d'où R_1 ; d'où P_j^1, etc... (résolution par itérations).

Application numérique

N = 4 $\qquad p_j = (0,1 - 0,2 - 0,3 - 0,4)$

Le calcul est en défaut pour $p_4 = 0,4$; on remplace P_4° par $1 - P_1^\circ - P_2^\circ - P_3^\circ$.

On obtient :

$$P_j^\circ = (0,082 - 0,184 - 0,342 - 0,392)$$

On calcule

$$R_1 = \frac{1}{4}(1,22 + 0,92 + 0,88 + 1,02) = 1,01$$

Il est donc inutile de commencer les itérations.

Remarque

BREWER [14] a réussi à montrer que le système d'équations (F) avait une solution unique.

FELLEGI [13] a étudié aussi le cas de n > 2, notamment sur des exemples. Il ne s'agit plus de n tirages à la même date mais à des dates échelonnées (échantillons tournants) ; à toute époque 2 unités seulement sont en service à la fois et les probabilités d'inclusion sont $2p_i$; elles ne varient pas quand on change l'une des 2 unités de sondage.

8 - LES DIVERSES CLASSES D'ESTIMATION

Si x_1, x_2, ..., x_n désignent les données-échantillon, HORVITZ et THOMPSON dans leur travail de précurseurs [20] (1952) avaient distingué 3 classes d'estimateurs possibles à priori :

1/ Ceux de la forme $S\alpha_{(t)} X_{(t)}$, où t désigne l'ordre de tirage des unités échantillons, les poids $\alpha_{(t)}$ dépendant de l'ordre de tirage ;

2/ Ceux de la forme $S\beta_i X_i$, où les poids β_i dépendent de l'unité i tirée ;

3/ Ceux de la forme γSX, où le coefficient γ tient compte de l'ordre dans lequel les unités échantillon sont sorties.

Il est apparu depuis (et notamment dans les travaux de KOOP [21]) que les 3 classes d'estimateurs ci-dessus n'épuisaient pas du tout les possibilités réelles, l'emploi d'estimateurs mixtes combinant les caractères de α, β, γ étant encore possible.

Toutefois on ne peut définir ces classes que si les unités échantillons proviennent de tirages successifs : Bernoulliens, de Midzuno, ou autres. Cela n'a aucun sens ni pour les sondages systématiques, ni pour celui de Cochran-Hartley-Rao, ni encore pour les sondages "rejectifs" (de Poisson-Hajek ou de Bernoulli).

9 - LA CLASSE (1) D'ESTIMATEURS $S\alpha_{(t)} X_{(t)}$

Avec les estimateurs de la classe (1), on notera :

- d'une part que la formule d'estimation traduit la dissymétrie qui existe dans l'apparition des unités de sondage par tirages consécutifs : $X_{(1)}$, $X_{(2)}$, ..., $X_{(n)}$;

- d'autre part que les coefficients α_t ne dépendent pas de l'unité obtenue ; autrement dit $X_{(t)} = X_i$, $\alpha_{(t)}$ est le même quel que soit i.

Cette seconde condition est en fait très forte. On ne semblait pas s'être beaucoup intéressé à de tels estimateurs depuis Horvitz et Thompson. Mais un court papier de PRABHU AJGAONKAR [22] vient de faire faire un pas décisif à la question.

Il résulte de ce travail que les sondages bernoulliens successifs, avec n = 2, ne peuvent avoir de tels estimateurs qui soient sans biais.

Font exception les sondages de Midzuno (et d'ailleurs avec n quelconque).

Autrement dit : il suffit que les tirages à partir du second soient faits avec d'égales probabilités pour qu'on ait de tels estimateurs sans biais ; il existe alors d'ailleurs une infinité de ces estimateurs.

Le plus simple est :

$$t_1 = X_{(1)} + \frac{N - 1}{n - 1} [X(2) + X(3) + \ldots]$$

autrement dit

$$\alpha_{(1)} = 1$$

$$\alpha_{(2)} = \alpha_{(3)} = ... + = \frac{N - 1}{n - 1}$$

Il se trouve d'ailleurs que c'est (dans sa classe (1)) l'estimateur de variance minimum.

Cas de n = 2 : $t_1 = X_{(1)} + (N - 1) X_{(2)}$. Supposons $X_{(1)} = X_i$ et $X_{(2)} = X_j$, $j \neq i$ effectivement :

$$\left.\begin{aligned} E\, X_{(1)} &= \Sigma\ p_i\ x_i \\ E\, X_{(2)} &= \sum_i p_i \sum_{j \neq i} \frac{1}{N-1} x_j = \frac{1}{N-1} \sum p_i (X - x_i) \end{aligned}\right\} \text{ avec } \Sigma\ p_i = 1$$

d'où

$$E(t_1) = \Sigma\ p_i x_i + X - \Sigma\ p_i x_i = X \quad \text{(estimateurs sans biais)}$$

Remarque : Variance de l'estimateur

L'auteur donne la formule

$$V(t_1) = \frac{(N - n)}{n - 1} \left[\sum_i (1 - p_i^2)\ x_i - \sum_{i \neq j} \frac{(1 - p_i - p_j)}{N - 2}\ x_i x_j \right]$$

Il est possible d'estimer la variance seulement pour $n \geq 3$ (le second terme du crochet représentant une covariance qui se confond avec une variance si n = 2.

10 - LA CLASSE (2) D'ESTIMATEURS : $S\beta_i X_i$

10.1 - La classe (2) est de beaucoup la plus employée ; les n unités-échantillons y jouent des rôles symétriques, sans tenir compte de l'ordre dans lequel on les a obtenues. Le poids β_i affecté à l'unité (i) est habituellement $(\pi_i)^{-1}$, de façon à fournir avec $S\beta_i X_i$ un estimateur sans biais du total des X des x de la population.

La formule SX_i/π_i est valable aussi bien pour l'échantillonnage sans remise que pour l'échantillonnage avec remise ; dans le 1er cas π_i est la probabilité d'inclusion de l'unité i dans l'échantillon ; dans le 2ème cas π_i est np_i, p_i étant la probabilité de tirage (au coup).

Démonstration dans le cas de tirages sans remise, avec n = 2.

$$S \frac{x_i}{\pi_i} = \frac{x_h}{\pi_h} + \frac{x_k}{\pi_k} \text{ si (h) et (k) sont les deux unités tirées.}$$

$$E\, S \frac{x_i}{\pi_i} = \sum_{hk} \pi_{hk} \left(\frac{x_h}{\pi_h} + \frac{x_k}{\pi_k} \right) = \sum_h \frac{\sum_k \pi_{hk}}{\pi_h} x_h = \sum_h x_h \ . \quad \text{(c q f d)}$$

Le point délicat est que $\sum_{hk}$ porte sur $\frac{N(N-1)}{2}$ couples h k, on a :

$$\sum_{hk} = \frac{1}{2} \sum_h \sum_k = \frac{1}{2} \left(\sum_h x_h + \sum_k x_k \right) = \sum_h x_h$$

(car bien entendu $\pi_{hh} = \pi_{kk} = 0$).

Les valeurs de π_i diffèrent avec la méthode d'échantillonnage. Le tableau du § II. 2 permet de voir quelles différences réelles se cachent sous une formule unique d'estimateur.

Trouvé par Horvitz et Thompson [20] (1952), cet estimateur jouit de certains caractères d'optimalité ; dans la classe des estimateurs linéaires homogènes sans biais, il est admissible, - c'est-à-dire qu'il n'existe aucun autre estimateur ayant une variance uniformément inférieure à la sienne [23] (cf. aussi [15], p. 376).

10.2 - Variance de l'estimateur

La formule de variance de l'estimateur $X = S\ x_i/\pi_i$, où π_i est la probabilité d'inclusion de l'unité i dans l'échantillon, a été donnée dès 1952 par Horvitz et Thompson [20].

$$V(\hat{X}) = \sum_N \frac{x_j^2}{\pi_j} + \sum^{\neq} \frac{\pi_{jk}}{\pi_j \pi_k} x_j x_k - X^2$$

ou

$$V(\hat{X}) = \sum_N x_j^2 \frac{1 - \pi_j}{\pi_j} + \sum^{\neq} x_j x_k \frac{\pi_{jk} - \pi_j \pi_k}{\pi_j \pi_k}$$

On l'établit directement sans difficulté.

Ici les π_{jk} sont les probabilités d'inclusion des unités j et k dans l'échantillon. Elles diffèrent suivant la méthode d'échantillonnage adoptée (cf. II § 3 ci-dessus). Il y a en outre une formule pour n = 2 et d'autres pour n > 2.

Toutefois c'est dans l'estimation de la variance $V(\hat{X})$ qu'on observe le plus de changements. La formule d'estimation proposée également par Horvitz et Thompson était défectueuse, en ce sens que, tout en étant sans biais, elle était affectée d'une telle variabilité qu'elle donnait couramment des estimations négatives ; ceci permet de penser qu'en contre-partie les valeurs positives obtenues étaient souvent trop grandes.

L'estimation de Yates et Grundy [5] (1953) généralement acceptée est

$$V(\hat{X}) = \frac{\pi_j \pi_k - \pi_{jk}}{\pi_{jk}} \left(\frac{x_j}{\pi_j} - \frac{x_k}{\pi_k} \right)^2$$

Pour qu'elle soit négative, il faudrait qu'on ait $\pi_{jk} > \pi_j \pi_k$.

Or il existe une inéquation due à Narain (1951) (tirages bernoulliens successifs), à savoir :

$$\pi_{jk} < 2 \left(\frac{n - 1}{n} \right) \pi_j \pi_k$$

qui, pour n = 2, se réduit à :

$$\pi_{jk} \leq \pi_j \pi_k$$

On peut donc affirmer que $V(\hat{X})$ est positif pour n = 2 avec ce type de tirage.

Pour n > 2, le raisonnement ne s'étend pas, mais l'inégalité reste de l'ordre des choses très vraisemblables.

10.3 - Comparaison des variances entre elles

RAO [10] (1963), a comparé (en faisant certaines approximations) les valeurs de $V(\hat{X})$ pour les trois procédés :

(1) bernoullien successif
(2) bernoullien rejectif
(3) systématique

Pour cela il lui fallait la valeur de π_{jk} dans chacun des trois cas. Pour les cas (1) et (2), voir ci-dessus II § 3. Pour le cas (3), l'expression approchée de π_{jk} avait été donnée par HARTLEY et RAO [9] (1962).

Des développements limités de ces 3 expressions en fonction de π_j et π_k sont alors obtenus, en supposant que p_i et $\pi_i/2$ soient voisins ($\forall_i$).

On constate que le terme principal de π_{jk} est la même dans les 3 cas, à savoir :

$$\frac{1}{2}\pi_j\pi_k + \frac{1}{4}(\pi_j^2\pi_k + \pi_j\pi_k^2) - \frac{1}{8}\pi_j\pi_k\sum_1^N \pi_i^2$$

mais les 3 expressions diffèrent par des termes complémentaires variés.

Passant de là à la variance estimée, on obtient le terme principal commun :

$$V(\hat{X}) \sim \left[1 - (\pi_j + \pi_k) + \frac{1}{2}\sum_1^N \pi_i^2\right]\left(\frac{x_j}{\pi_j} - \frac{x_k}{\pi_k}\right)^2 = W$$

A titre de curiosité, voici les termes complémentaires qu'il conviendrait d'ajouter au crochet figurant dans $V(\hat{Y})$.

partie commune	$-\frac{1}{2}(\pi_j^2 + \pi_k^2) + \frac{1}{2}\sum_1^N \pi_i^2$
tirages successifs	$+\frac{5}{8}\pi_j\pi_k - \frac{3}{32}\left(\sum_1^N \pi_i^2\right)^2 - \frac{1}{16}(\pi_j + \pi_k)\sum_1^N \pi_i^2$
rejectifs	$+\frac{1}{2}\pi_j\pi_i - \frac{1}{8}\left(\sum_1^N \pi_i^2\right)^2 + 0$
systématiques	$-\frac{1}{4}\left(\sum_1^N \pi_i^2\right)^2 + \frac{1}{4}(\pi_j + \pi_k)\sum_1^N \pi_i^2$

Ajoutons que RAO donne aussi des formules pour n > 2 (cas que nous n'avons guère envisagé ici). Ces calculs très laborieux nous apprennent en somme que les 3 modes de tirage ont pratiquement la même variance W.

D'autres calculs [7] montrent que la variance obtenue, avec le procédé de Cochran-Hartley-Rao (cf. I, 6) ne descend jamais en-dessous (est généralement au-dessus) de W.

En ce qui concerne le sondage de Poisson-Hajek (cf. I, 7) il existe [4] (1964), une théorie asymptotique, prouvant l'existence d'une loi de Gauss limite pour $\hat{X}$; mais on n'a pas procédé encore, semble-t-il, à une confrontation des π_{jk} ni des variances. Hajek ne semble d'ailleurs pas pleinement d'accord avec la façon dont Rao a lui-même procédé.

Nous n'examinerons pas ici les variances des estimateurs obtenus par d'autres procédés, tels ceux de HANURAV, quel qu'en soit l'intérêt.

11 - LA CLASSE (3) D'ESTIMATEURS : γSx_t

11.1 - Cas des sondages bernoulliens successifs (n = 2)

La probabilité de tirer (i) puis (j) est : $p_i(1 - p_i)^{-1} p_j \neq p_j(1 - p_j)^{-1} p_j$.

On aura donc 2 estimateurs sans biais de X :

$t = \lambda(1 - p_i) \frac{x_i + x_j}{p_i p_j}$, si (i) est sortie avant (j) ; $\lambda(1 - p_j) \frac{x_i + x_j}{p_i p_j}$ si c'est le contraire.

On a :

$$E\,t = \lambda \sum_{i,j} (x_i + x_j) + \lambda \sum_{i,j} (x_j + x_i) = 2\,\lambda N(N - 1) \left(\frac{\sum_1^N x_k}{N}\right)$$

L'absence de biais implique donc $\frac{1}{\lambda} = 2\,(N - 1)$

Exemple

$$p_i = (0{,}1 - 0{,}2 - 0{,}3 - 0{,}4) \; ; \; N = 4, \; \lambda = \frac{1}{6}$$

Valeurs de $\frac{1 - p_i}{p_i p_j}$		j = 1	2	3	4
t = 1	i = 1	-	45	30	22,5
	2	40	-	13,3	10
	3	23,3	11,7	-	5,8
	4	15	7,5	5	

11.2 - Cas des sondages de Midzuno (n = 2)

La probabilité de $(x_i + x_j)$ est $p_i/(N - 1)$ ou $p/(N - 1)$ suivant qu'on a tiré d'abord (i) ou bien (j). Le facteur λ est encore $1/2(N - 1)$. Ainsi il

vient $t = \frac{1}{2p_i}(x + x)$ quand on convient que i a été tiré d'abord

Exemple

	p_i	$\frac{1}{2p_i}$
i = 1	$\frac{1}{10}$	5
2	$\frac{2}{10}$	2,5
3	$\frac{3}{10}$	1,67
4	$\frac{4}{10}$	1,25

12 - ESTIMATEURS N'APPARTENANT PAS AUX CLASSES (1), (2), (3)

Il est facile de comprendre que les classes (1), (2), (3) ne recouvrent pas toutes les catégories d'estimateurs possibles, ni même tous ceux qui sont utiles.

Montrons comment constituer une classe (1 bis), par exemple :

$$S\alpha(t) \cdot x(t), \quad \text{où} \quad : \quad x(t) = x_i \Rightarrow \alpha = \alpha(t, i)$$

Il suffit de noter que, si l'on prenait :

$\alpha(t, i)$ = [Probabilité d'inclusion de l'unité i au $t^{ème}$ tirage]$^{-1}$, les $\alpha(1,.)$ x(1) ; $\alpha(2),.)$ x (2) ; .... ; $\alpha(n,.)$ x (n), seraient n estimateurs sans biais de X. Toute combinaison linéaire de ces éléments est dans le même cas ; d'où le type d'estimateurs suivant (à poids ω arbitraires) :

$$\tilde{X} = \sum_t \alpha(t,.)\ x(t)\ \omega_t \ , \text{ avec } \sum \omega_t = 1$$

Exemple : Reprenons l'exemple déjà utilisé si souvent

Population : (A B C D). Probabilités p_i = (0,1 - 0,2 - 0,3 - 0,4).

Procédons à n = 2 tirages successifs (sans remise).

1ère unité 2ème unité :	A	B	C	D	Total p_i
A	-	2/90	3/90	4/90	9/90 = 0,1
B	2/80	-	6/80	8/80	16/80 = 0,2
C	3/70	6/70	-	12/70	21/70 = 0,3
D	4/60	8/60	12/60	-	24/60 = 0,4
autrement dit	0,025	0,022	0,033	0,045	
	0,043	0,086	0,075	0,100	
	0,067	0,133	0,200	0,171	Total
$\frac{1}{\alpha_{(2)}}$ = Prob $[x_{(2)} = x_j] = q_j$	0,135	0,421	0,308	0,316	1,000

Estimateur Type 1 bis :

$$\alpha_{(1)}\, x_{(1)} + \alpha_{(2)}\, x_{(2)} = \boxed{\frac{x_i}{p_i}\,\rho + \frac{x_j}{q_j}\,(1 - \rho)} = \tilde{\bar{X}}$$

où ρ est un paramètre arbitraire compris entre 0 et 1.

Il est bien clair que la valeur "optimale" de ρ dépendrait des x_i.

Remarque

L'estimation de V(X) sur échantillon n'est pas disponible ; car on n'est pas en mesure d'estimer la covariance de x_i et x_j avec n = 2.

CONCLUSION

Ainsi s'achève cet examen (malgré tout très superficiel) des méthodes de tirage au sort d'échantillons, avec probabilités inégales et sans remise.

Signalons que la méthode la plus commode (tirages systématiques) est aussi la plus pauvre au point de vue théorique : les articles tout récents de CONNOR [12] et HARTLEY [11] permettent à peine de pénétrer dans les mystères des calculs de variance s'y rapportant. Il n'est pas tellement certain que les résultats asymptotiques (pour N extrêmement grand) concernant l'équivalence de 3 méthodes à cet égard (Bernoulli rejectif, Bernoulli successif, systématique) aient encore une certaine portée quand la population n'est pas très grande.

La méthode de Poisson-Hajek occupe toujours une place à part ; elle ne semble pas praticable sous sa forme rejective, quand on doit désigner parmi des bennes qui défilent devant vous celles qui constitueront l'échantillon (l'un des cas où la méthode ordinaire de Poisson-Hajek est commode). Il n'est pourtant pas impossible qu'on lui trouve une véritable utilisation ; l'essentiel est peut-être, pour l'instant, qu'on apprenne aux praticiens (par exemple ceux des études de marché) qu'on ne peut impunément tirer des échantillons à la Poisson et les traiter comme des échantillons de Bernoulli (ce qui semble assez fréquent).

La nécessité de choisir entre rejectif et successif, pour les tirages (bernoulliens) de villes (dans leurs strates) ne devrait pas davantage échapper aux organisateurs d'enquêtes sociales et économiques.

Ces aspects pratiques sont évidemment peu de chose à côté des efforts d'ordre mathématique faits par une véritable troupe de chercheurs - efforts dont rien ne permet de prévoir la fin.

ANNEXE

SUR LE PROCEDE DE SONDAGE A DE HANURAV[1]

1/ Soit p la probabilité de tirer l'unité u au 1er tour (de 2 tirages), $\Sigma p = 1$.

Soit q, r, ... les probabilités de la même unité u aux 2ème, 3ème... tours.

Soit u et u' 2 unités tirées au sort : si elles sont distinctes, les tirages au sort s'arrêtent.

La probabilité de tirer u et u' au 1er tour est $2\,pp'$

Mais celle de tirer 2 unités identiques est $\Sigma p^2 = P$.

De même au 2ème tour, s'il y en a un :

probabilité qu'il y ait un 2ème tour : P

probabilité d'y tirer u et u' : $2qq'$

On pose :

$$\Sigma q^2 = Q,\ \Sigma r^2 = R \ldots\ldots$$

La probabilité d'inclusion de u (soit π) est alors :

au 1er tour $2p\ \Sigma p' = 2p\ (1 - p) = 2p - 2p^2$

au 2ème tour $P\ 2q\ \Sigma q' = P\ 2q\ (1 - q) = 2q\ P - 2q^2P$

au 3ème tour $P\ Q\ 2r\ \Sigma\ r' = P\ Q\ 2r\ (1 - r) = 2r\ P\ Q - 2r^2\ PQ$

Regroupons les termes, il vient :

$$\pi = 2p - 2(p^2 - qP) - 2P\,(q^2 - rQ) - \text{etc...}$$

Donc il suffit d'avoir

$$q = p^2/P = p^2/\Sigma p^2$$
$$r = q^2/Q = q^2/\Sigma q^2$$
$$\ldots\ldots\ldots$$

pour qu'il ne reste que

$$\boxed{\pi = 2p}$$

Résultat 1

Si la probabilité de tirer u au 1er tour est p et au 2ème tour est proportionnelle à p^2, - au 3ème tour est proportionnelle à p^4, etc..., alors la probabilité d'inclusion de u dans l'échantillon est $\pi = 2p$.

---

(1) Tous les résultats donnés sont dans l'article de HANURAV, on a seulement cherché à en clarifier les démonstrations, aux dépens de la rigueur bien entendu.

2/ Si la suite ordonnée $p_1 \leq p_2 \leq \ldots\ldots \leq p_{N-1} \leq p_N$ des probabilités $p(u_i)$ ne se termine pas par $p_{N-1} = p_N$, on risque de ne jamais tirer qu'une seule unité : l'unité la plus probable (i = N).

Plus précisément, quelle est la probabilité que la séquence des tours (à 2 tirages) ne se termine jamais ? C'est le produit.

$$P\ Q\ R\ \ldots\ldots\ldots$$

(probabilité qu'il y ait un 2ème, un 3ème, un 4ème, ... tours)

On est donc ramené à étudier si le produit infini P Q R ..... converge vers 0 ou non : dans le 1er cas, l'évènement de probabilité 0 constitue un risque négligeable, dans le 2ème cas, la méthode de sondage peut être mise en défaut (et il faut la modifier).

Dire que P Q R... tend vers 0, c'est dire que la série positive

$$(-\ \text{Log}\ P) + (-\text{Log}\ Q) + (-\ \text{Log}\ R) + \ldots\ldots \text{ diverge (vers } +\infty)$$

Dans l'autre hypothèse, c'est que cette série converge.

Cette étude est assez délicate. Tenons-nous-en aux arguments de bon sens.

1/ S'il existe $p_N = a$ supérieure à toutes les autres, la suite P, Q, R,... est

$$P = \Sigma p^2 = a^2 + \ldots\ldots = a^2(1 + \varepsilon)$$

$$q = \frac{p^2}{P} \qquad Q = \Sigma q^2 = b^2 + \ldots\ldots = b^2(1 + \varepsilon') \qquad \text{avec} \quad b^2 = \frac{a^4}{p^2} = \frac{a^4}{(a^2 + \ldots)^2}$$

$$r = \frac{q^2}{Q} \qquad R = \Sigma r^2 = c^2 + \ldots\ldots = c^2(1 + \varepsilon'') \qquad c^2 = \frac{b^4}{Q^2} = \frac{b^4}{(b^2 + \ldots)^2}$$

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

Tous les a^2 se mettent en facteur ; les b^2, c^2, ... s'éliminent. Il vient :

$$P = a^2(1 + \varepsilon)$$

$$Q = (1 + \varepsilon')\ (1 + \varepsilon)^{-2}$$

$$R = (1 + \varepsilon'')\ (1 + \varepsilon')^{-2} \ldots\ldots$$

$$\Longrightarrow P\ Q\ R\ \ldots\ldots = a^2(1 + \varepsilon)^{-1}\ (1 + \varepsilon')^{-1}\ (1 + \varepsilon'')^{-1} \ldots\ldots$$

$$\Longrightarrow -\ \text{Log}\ (P\ Q\ R \ldots\ldots) \sim \text{Log}(a^{-2}) + (\varepsilon + \varepsilon' + \varepsilon'' + \ldots\ldots)$$

Nous voyons donc que la probabilité P Q R ... n'a aucune raison de tendre vers 0. Car ceci supposerait que la série

$$\varepsilon + \varepsilon' + \varepsilon'' + \ldots\ldots \text{ diverge,}$$

alors que c'est une somme de séries, chacune de la forme :

$$\rho + \rho^2 + \rho^4 + \rho^8 \ldots\ldots, \quad 0 < \rho < 1$$

on peut donc admettre que :

Résultat 2

Il y a une probabilité non nulle que le procédé (des élévations au carré successives des probabilités) ne fournisse jamais 2 unités distinctes si la population comprend une unité plus probable que toute autre.

3/ Si l'on suppose $p_{N-} = p_N$, il est évident au contraire qu'on obtiendra le couple d'unités (N - 1) et (N) à la limite. Il suffit cette fois de poser $p_{N-1} + p_N = 2p_N = a$.

Alors lim(P Q R....) est la probabilité d'obtenir le couple (N - 1, N).

On peut aussi le voir directement : le terme général T du produit infini sera de la forme :

$$\frac{2a^{2\nu} + \ldots.}{(2a^{\nu} + \ldots.)^2} = \frac{1}{2}\,\frac{a^{2\nu}}{a^{2\nu}}\,\frac{1 + \eta'}{(1 + \eta)^2} \longrightarrow \frac{1}{2}$$

Ainsi le produit infini se comporte comme $\left(\frac{1}{2}\right)^t$, et tend vers 0.

Nota : Quand on a $p_{N-2} = p_{N-1} = p_N$, il n'est plus possible d'énoncer la composition limite exacte de l'échantillon.

4/ Modification du procédé si $p_N = p_{N-1} + \delta$, $\delta > 0$.

Il est évident qu'on peut amputer l'unité (N) de son excédent de probabilité δ (ce qui conduit à diviser par $(1 - \delta)$ toutes les probabilités pour élever à 1 leur total). On peut alors appliquer le procédé. Mais il faut procéder à une "mixture", c'est-à-dire laisser une chance $(1 - \delta)$ à ce type d'échantillon et une chance δ à un échantillon comprenant nécessairement l'unité (N), plus une autre unité (j), $j < N$.

REFERENCES BIBLIOGRAPHIQUES

[1] THIONET (P) - Théorie des sondages : quelques problèmes récents. Journal de la Société de Statistique de Paris 108 (1967-1) p. 9/30.

[2] William G. COCHRAN awarded 1967 Wilks Memorial Medal. Acceptance Reply of Prof. COCHRAN, The American Statistician, dec. 1967, p. 1/2.

[3] DURBIN (J) - Some results in sampling theory when the units are selected with unequal probabilities. J.R.S.S. B 15-2 (1953) p. 262/269.

[4] HAJEK (J) - Asymptotic theory of rejective sampling with varying probabilities from a finite population. A M S 35 (dec 1964) p. 1491/1523

[5] YATES (F) & GRUNDY (P.M.) - Selection without replacement from within strata with probability proportional to size, J R S S B 15-2 (1953) p. 253/261.

[6] MIDZUNO (H) - On the sampling system with probability proportionate to sum of sizes. Ann. Inst. Statist. Math. Tokyo 3 (1152) p. 99/107.

[7] HARTLEY (H.O.), RAO (J.N.K.) COCHRAN (W.G.) - On a simple procedure of unequel probability sampling without replacement, J.R.S.S B 24-2 (1962) p. 482/491.

[8] GOODMAN (R) & KISH (L) - Controlled selection - A technique in probability sampling, J.A.S.A. 45 (1950) p. 350/372.

[9] HARTLEY (H.O.) & RAO (J.N.K.) - Sampling with unequal probabilities and without replacement, A.M.S 33 (June 1962) 350/374.

[10] RAO (J.N.K.) - On three procedures of unequal probability sampling without replacement, J.A.S.A. 58 (March 1963) p. 202/215.

[11] HARTLEY (H.O.) - Systematic sampling with unequal probability and without replacement, J.A.S.A. 61 (Sept. 1966) p. 739/748.

[12] CONNOR (W.S.) - An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement, J.A.S.A. 61 (June 1966) p. 380/390.

[13] FELLEGI (I.P.) - Sampling with varying probabilites without replacement : rotating and non rotating samples, J.A.S.A. 58 (1963) p. 183/201.

[14] BREWER (K.R.W.) - A note on Fellegi's method of sampling without replacement with probability proportional to size, J.A.S.A. 62 (1967) p. 79/85.

[15] HANURAV (T.V.) - Optimum utilization of auxiliary information : p.s. sampling of two units from a stratum, J.R.S.S. B 29-2 (1967) p. 374/391.

[16] STEVENS (W.L.) - Sampling without replacement with probability proportionate to size, J.R.S.S. B 20-2 (1958) p. 393/397.

[17] PATHAK (P.K.) - On sampling schemes providing unbiased ratio estimators, A.M.S. 35-1 (1964) p. 222/231.

[18] POMPILJ (G.) - Estrazioni in blocco di unita con probabilita differenti, Sec. serie Publ. Inst. Calcolo delle Probal dell Un. di Roma, 22 (1961), 5 pages.

[19] THIONET (P.) - A propos d'un problème de concours : la méthode de sondage de Midzuno. Bulletin d'information de l'I.N.S.E.E. (1968) N° 2, p. 7/20.

[20] HORVITZ (D.G.) & THOMPSON (D.J.) - A generalization of sampling without replacement from a finite universe, J.A.S.A. 47 (1952) p. 662/685.

[21] KOOP (J.C.) - Contribution to the general theory of sampling, etc.... Document de l'Institut de Statistique de l'Université de Caroline du Nord N° 296 (1961).

Cité par :

PRABHU AJGAONKAR (S.G.) - On a class of linear estimators in sampling with varying probabilities without replacement, J.A.S.A. 60 (1965) p. 637/642.

[22] PRABHU AJGAONKAR (S.G.) - On Horvitz and Thompson's T class of linear estimators, A.M.S. 38 (Dec. 1967) p. 1882/1886.

[23] ROY (J.) & CHAKRAVARTI (I.M.) - Estimating the mean of a finite population A.M.S. 31 (1960) p. 392/398.

GODAMBE (V.P.) - An admissible estimate for any sampling design Sankhya 02 (1960) p. 285/288.

Additif

[24] JESSEN R.) - Some methods of probability non-replacement sampling, J.A.S.A. 64 (mars. 1969) p. 175-193.

[25] HANURAV (T.V.) - Addendum to [15], J.R.S.S. B 31-1 (1969) p. 192-194 .