REVUE DE STATISTIQUE APPLIQUÉE

P. THIONET

La méthode des sondages

Revue de statistique appliquée, tome 9, n° 1 (1961), p. 7-52

<http://www.numdam.org/item?id=RSA_1961__9_1_7_0>

LA MÉTHODE DES SONDAGES

par P. THIONET
Administrateur à l'Institut National de la Statistique
et des Études Économiques

SOMMAIRE

I

II

Le problème général qui se pose est l'étude sur échantillon d'une "population" ou "univers", c'est-à-dire d'un "ensemble". Par exemple, il est très important d'avoir des renseignements sur l'attitude du public à l'égard de certains produits; disons de savoir pourquoi le public n'achète pas une certaine marque de crème à raser; mais il est clair qu'on ne va pas interroger tout le monde, on se limitera à un échantillon qui doit avoir le mérite d'être "représentatif", c'est-à-dire de fournir des renseignements peu déformés sur l'attitude générale.

Quand nous parlons d'étude d'une population, d'un ensemble, nous pensons à quelque chose qui a un sens mathématique précis. Les éléments de cette population présentent des caractères X et nous voulons estimer la valeur moyenne $\bar{x}$ de ces caractères. En particulier, s'il s'agit de caractères qualitatifs à 2 alternatives, X devient une variable (attachée à chaque élément) égale à 1 ou 0, suivant que l'élément présente ou non le caractère en question, et $\bar{x}$ est la proportion d'éléments présentant le caractère.

Ainsi, on se propose d'estimer, par exemple, quelle est la proportion des hommes de ce pays qui n'achètent pas cette crème à raser et qui ne l'ont jamais essayée.

En réalité, on peut se proposer de porter par sondage, c'est-à-dire sur échantillon des jugements plus compliqués qu'une simple estimation; on examinera ci-après les principales méthodes de sondage.

PREMIÈRE PARTIE

MÉTHODES EMPIRIQUES ET MÉTHODES PROBABILISTES

Nous distinguerons parmi les méthodes de sondage deux grands groupes: les Méthodes empiriques et les Méthodes probabilistes.

I - LES METHODES EMPIRIQUES

Il s'agit des méthodes où "le Calcul des Probabilités" n'a pas à intervenir. Pour avoir un échantillon valable, on part de cette idée qu'il faut le composer en choisissant au mieux les éléments qui y seront inclus. On peut, par exemple, choisir des "éléments moyens" (conformes à une certaine notion intuitive de la moyenne) ou encore des éléments faciles à soumettre à l'étude. Voici une forme très perfectionnée de ces méthodes :

La Méthode des quota

C'est la méthode en usage dans la plupart des instituts d'opinion publique ou d'études de marchés : supposons qu'on veuille constituer un échantillon d'hommes de 18 ans et plus, réparti de manière représentative dans l'Ouest de la France, région où l'on se propose de chercher à vendre un certain produit.

On commence par se procurer les résultats du dernier recensement de population concernant les départements de l'Ouest. (Ces résultats sont publiés par les services officiels de Statistique, mais soit faute de crédits, soit délais de publication, il y a parfois certains renseignements qui ne sont pas publiés et qui sont pourtant disponibles, c'est-à-dire non confidentiels; il y a lieu de consulter les services compétents). On connaît ainsi la répartition de ces hommes suivant l'âge, la profession, le groupe de communes de résidence (grandes villes, petites villes, communes rurales, etc.).

Supposons alors qu'on veuille enquêter auprès d'une personne sur 3 000; par exemple, les hommes de 18 ans et plus habitant l'Ouest de la France seront au nombre de 3 millions et on se proposera d'en interroger 1 000.

(Indiquons en passant que 1 000 est petit comme effectif de l'échantillon si l'on veut obtenir des résultats précis; mais que 1 000 est déjà grand si l'on considère les frais d'enquête que cela représente : il faut que l'étude de marché soit payante).

Dans ces conditions, on divise par 3 000 tous les nombres figurant dans les tableaux statistiques dont on dispose pour l'ensemble des départements de l'Ouest; et on obtient (en arrondissant à l'unité) ce qu'on appelle la composition théorique de l'échantillon.

Par ailleurs, on doit disposer d'enquêteurs habitant dans des communes de toutes tailles disséminées dans la région Ouest; ce sont soit les correspondants de la firme d'études de marchés, soit les agents du service commercial de la firme qui organise pour son propre compte l'étude de marché. Supposons qu'on dispose ainsi d'une cinquantaine de points d'enquête.

On va alors répartir entre ces points l'échantillon de 1 000 personnes, de manière à respecter autant que possible la composition de l'échantillon théorique. Puis on adresse à chacun des cinquante enquêteurs (correspondants ou agents commerciaux) une fiche de travail et un jeu de 20 ou 30 questionnaires.

La fiche de travail porte la composition du petit échantillon que l'enquêteur doit "interviewer". Par exemple (si l'enquêteur est M. DURAND) :

M. DURAND, à X .... (commune de 4 000 habitants)

Interrogez 18 hommes (dans votre commune ou aux alentours)

dont	18 à 30 ans	5	cultivateurs	4
	31 à 45 ans	5	ouvriers	7
	46 à 60 ans	4	employés	3
	+ de 60 ans	4	commerçants	2
		18	autres	2
				18

Le questionnaire porte, imprimées, une série de questions que l'enquêteur doit poser à l'enquêté au cours de l'interview; les réponses recueillies sont notées en regard. Telle est la méthode la plus courante. On prescrit, en général, à l'enquêteur de poser les questions telles qu'elles sont imprimées; une légère déformation des questions ou un changement dans l'ordre des questions peut entraîner une variation des réponses, surtout en ce qui concerne les questions d'opinion. Mais avec des enquêteurs très évolués et qui apprennent bien le questionnaire comme un acteur apprend son texte, on peut faire enquêter les mains dans les poches et même sans que les enquêtés s'en aperçoivent; ce n'est pas le cas général.

Ici il convient d'ouvrir une parenthèse : dans le cas qui nous occupe, les enquêteurs ont une autre profession; et le métier d'enquêteur est un métier d'appoint; ils posent tout naturellement leurs questions aux gens qu'ils rencontrent au cours de leur activité professionnelle normale. C'est bien entendu le cas d'un garagiste dépositaire de Citroën, qui pose aux gens venant faire laver ou réparer leur voiture, des questions qu'on peut d'ailleurs lui prescrire de ne poser qu'à des automobilistes; mais ce sera aussi bien le cas pour un enquêteur qui est à la fois clerc de notaire dans un petit bourg rural et un enquêteur d'un institut d'opinion publique. On devine ainsi que le choix des enquêteurs peut avoir éventuellement une influence déformante sur l'échantillon.

Finalement chaque enquêteur retourne (par la poste) au service central, les questionnaires dûment remplis; et on les dépouille pour en tirer les statistiques intéressantes : proportions ou moyennes.

Inconvénients de cette méthode

La méthode n'a pas de justification théorique et peut manifestement conduire à des erreurs graves (si, par exemple, tous les enquêteurs sont recrutés dans le même milieu); en outre, elle nécessite l'emploi de statistiques qui, en France, font souvent défaut dès qu'on sort des renseignements

classiques. (Par exemple, en France, on n'a aucune statistique sur les ressources réelles des ménages).

Avantages de cette méthode

En pratique, ceux qui utilisent la méthode s'en disent satisfaits; car, souvent on aime mieux avoir des renseignements imprécis que pas de données chiffrées du tout. En outre, il y a tellement de causes d'erreur et d'imprécision dans les enquêtes sur les opinions ou le comportement des gens que les erreurs d'échantillonnage n'ont pas besoin d'être réduites au minimum. Un sage principe veut même que l'on s'arrange pour que les erreurs des diverses sources soient du même ordre de grandeur ; dépenser de l'argent pour avoir des erreurs d'échantillonnage 2 ou 3 fois plus petites que les erreurs d'observation n'est pas faire preuve d'une attitude bien réaliste.

Ouvrons ici deux parenthèses :

a) Les erreurs d'observation ou de mesure. Si l'on demande à quelqu'un quelle est la marque de son savon dentifrice et qu'il réponde "Gibbs", il est très possible qu'il fournisse cette réponse parce que c'est le premier nom qu'il lui vient à l'esprit : alors qu'il emploie en réalité une marque peu connue (même de lui).

Si l'on demande à un homme : "Achèteriez-vous un rasoir électrique s'il ne coûtait que 3 000 frs ?" et qu'il réponde OUI (les marques concurrentes coûtant au moins 4 000 frs) il est très possible que finalement il n'en achète pas à 3 000 frs parce qu'il aura appris dans l'intervalle, par un ami, que le rasoir électrique n'était pas aussi commode qu'on le supposait.

Ainsi, les résultats que fournissent les études de marchés ne peuvent pas être précis. Ceci ne veut pas dire que ces études soient inutiles. Bien au contraire, il n'est pas douteux qu'en France elles soient insuffisamment pratiquées. Un industriel achètera facilement de nouvelles machines, de nouveaux brevets, embauchera un nouvel ingénieur et des ouvriers, pour produire finalement un article qui ne répondra pas du tout aux goûts du public. Il dépensera ensuite des sommes énormes en frais de publicité pour essayer d'imposer cet article. Cette façon de faire est irrationnelle.

b) Les budgets d'enquête. Un obstacle aux études de marchés est précisément leur coût. Il semble qu'en France, actuellement, les industriels soient peu disposés à dépenser, par exemple, quelques millions pour une étude de marché préalable à un investissement de quelques centaines de millions. Quant aux instituts d'opinion publique, - s'ils sont autonomes, ce qui est nécessaire pour enquêter objectivement, - ils manquent de ressources financières : en France les résultats de leurs enquêtes ne sont qu'assez exceptionnellement achetés par la Presse (comme cela se fait aux Etats-Unis).

Dans les deux cas, on a donc généralement intérêt (tant qu'une précision quelconque n'est pas demandée pour les résultats) à employer les méthodes empiriques de sondage. Les méthodes probabilistes ne souffrent pas la médiocrité des moyens; et, par exemple, un échantillon tiré au hasard de 10 personnes a beaucoup moins de chances d'être représentatif qu'un échantillon choisi à cette fin.

Les méthodes que nous avons appelées empiriques sont, pour cette raison, appelées de choix raisonné par leurs partisans; les méthodes probabilistes étant appelées alors : méthodes de choix au hasard. ces noms sont tendancieux car ils semblent impliquer la suprématie de la raison sur le hasard; il est plus objectif de dire que les méthodes empiriques comportent la pratique du choix à dessein ou choix intentionné et que les méthodes probabilistes comportent le tirage au sort de l'échantillon.

Nous ne nous occuperons plus dans ce qui suit des méthodes empiriques, estimant qu'il y a lieu, ici surtout, de faire connaître et comprendre les méthodes probabilistes.

II - LES METHODES PROBABILISTES

On se propose donc d'appliquer le calcul des probabilités au problème de l'étude des propriétés d'un ensemble au moyen d'un échantillon.

Pour cela, il faut que l'échantillon provienne d'un tirage au sort ou d'un système de tirages au sort. L'ensemble à étudier doit pouvoir être assimilé à des boules dans une urne (schéma classique de Bernoulli) ou à tout autre système plus compliqué mais reposant sur les mêmes conceptions.

Il faut qu'on soit en mesure de définir les "boules" et ensuite de les tirer au sort. On parlera désormais non de "boules" mais d'"unités de sondage" : on désigne ainsi une entité qu'on doit pouvoir définir et tirer au sort. Considérons, par exemple, un tas de charbon. On veut en tirer un échantillon valable pour étudier son calibre ou sa teneur en cendres. Comment faire ? (le charbon renferme des pierres; les gros morceaux ont une tendance à descendre plus que les petits).

Il est impossible d'adopter le morceau de charbon comme unité de sondage, car qui dit "TIRAGE" au sort dit : possibilité d'attribuer à l'avance à chaque unité de sondage un certain numéro d'ordre et de retrouver dans la masse les unités auxquelles les "numéros gagnants" (à une loterie quelconque) avaient été affectés.

Mais imaginons qu'on transporte depuis la mine le charbon dans les wagonnets d'un petit téléférique; considérons l'unité de sondage constituée par le contenu d'un wagonnet. On peut numéroter ceux-ci et décider de soumettre à l'enquête le contenu des wagonnets n°4, 25 et 47 ; au moment où le wagonnet se vide, on s'arrangera pour que ceux-ci se vident sur la plateforme d'un camion et non sur le tas de charbon; et on aura ainsi obtenu l'échantillon au hasard qu'on voulait.

Vous devez vous demander comment on arrive à tirer des échantillons des populations humaines. On pourrait disposer d'un répertoire complet de tous les habitants du pays : par exemple, ce serait le cas si tout le monde était assujetti à la Sécurité Sociale, avait un numéro d'identification de la S.S. et figurait au fichier de celle-ci. Mais ce n'est pas du tout le cas en France (ou même le fichier de la S.S. est difficilement utilisable). En réalité, on adopte plutôt comme unité de sondage le ménage, c'est-à-dire les personnes habitant le même logement. En effet, il est, en général, possible d'avoir un répertoire (plus ou moins bon d'ailleurs) des logements (par exemple, le fisc en possède un). On pourrait aussi se servir d'un répertoire des immeubles et considérer que tous les habitants de l'immeuble constituent une unité de sondage, etc. Telle est, en gros, la méthode qui convient à une administration d'Etat pour procéder à des enquêtes; elle ne conviendrait pas, bien entendu, à un service privé. On ne s'étendra pas davantage sur cette question.

On va, à présent, passer en revue les principales méthodes probabilistes. On les étudiera ensuite du point de vue des calculs. On distinguera :

1) Le sondage par tirage au sort élémentaire.

2) Le sondage systématique.

3) Le sondage en grappes.

4) Le sondage stratifié.

5) Le sondage à plusieurs degrés.

6) Le sondage à plusieurs phases.

7) L'emploi d'une formule d'estimation utilisant des renseignements supplémentaires.

1/ Le sondage par tirage au sort élémentaire

Considérons un bateau chargé de sacs de sucre. On décharge la cargaison; et le problème se pose, tant pour les douaniers que pour la firme qui achète la cargaison, de connaître le poids exact de cette dernière. En général, on pose successivement chacun des sacs sur la bascule, au fur et à mesure qu'ils sortent du bateau; on note le poids et on fait l'addition. Toutefois, compte tenu de la précision de ce procédé, il est possible de se contenter de peser un échantillon de ces sacs et d'extrapoler, en admettant que c'est le poids moyen de tous les sacs.

Donnons aux sacs un numéro d'ordre i = 1, 2, ..., i, ..., m au fur et à mesure qu'on les débarque. Soit x_i le poids du sac n°i, le poids moyen de l'ensemble des m sacs est :

$$\bar{x} = \sum \frac{x_i}{m}$$

Désignons par S la somme des poids des sacs échantillons ; le poids moyen de ceux-ci (au nombre de n) est $\bar{X} = S \frac{x_i}{n}$. On estime $\bar{x}$ à l'aide de $\bar{X}$ (et $\sum x_i$ à l'aide de $m\bar{X}$).

Pour effectuer le tirage au sort, il faut tirer au sort à l'avance n nombres compris entre 1 et m. Au lieu d'utiliser chaque fois un dispositif tel que les sphères de la Loterie Nationale, on se sert d'une Table de nombres aléatoires établie une fois pour toutes.

Les tables de nombres aléatoires ("cf. random numbers") - Il existe actuellement, à notre connaissance, par ordre d'ancienneté :

1) Les tables de Tippett (Tract for Computers n°15, Cambridge), formées de nombres de 4 chiffres, établies plutôt en vue des applications industrielles.

2) Les tables de Fisher et Yates (1938) formées de nombres de 2 chiffres, en vue des applications de la statistique à l'agriculture, la biologie, etc.

3) Les tables de Kendall et Babington Smith (Tract for Computers n°24 - 1938) formées de nombres de 5 chiffres.

4) Tables de Burke Horton (U.S.A., 1949) formées de nombres de 5 chiffres.

5) Tables de la "Rand Corporation" (U.S.A., 1955) 200 000 groupes de nombres de 5 chiffres.

Si les tables de Tippett sont formées de nombres relevés dans les recensements anglais, et celles de Fisher et Yates à l'aide des 15ème à 19ème décimales d'une table de logarithmes, les Tables de Kendall sont, en revanche, obtenues par véritables tirages au hasard. Une machine spéciale était construite à cet effet: on fait tourner une roue sur laquelle sont ins-

crits les chiffres de 0 à 9; une lampe au néon illumine le disque à intervalles irréguliers et un observateur lit alors le chiffre placé en face d'un index. Mieux vaudrait photographier ce nombre, car l'expérience a montré que certains observateurs commettaient systématiquement des erreurs de lecture d'un chiffre pour l'autre; mais on a pu éliminer ces observateurs à l'expérience.

On forme un nombre de la Table avec des chiffres tirés ainsi indépendamment les uns des autres. Nombres de Tippett, de Yates ou de Kendall, ces nombres n'ont rien d'"aléatoire", ce qui est aléatoire c'est l'ordre dans lequel ils se trouvent placés sur la Table (l'ordre dans lequel on les a obtenus). Comme les tirages sont nombreux, le hasard fait que certaines colonnes des Tables offrent des rencontres qu'on croirait assez rares. Par exemple, dans la dernière colonne du tableau XXIII de Tippett, on trouve successivement :

6.686
2.674
3.416
4.672
2.200
3.341
1.130
4.442
1.243
7.686

Donc, sur 10 nombres entre 0.000 et 9.999 pris au hasard, 8 sont inférieurs à 5.000. L'utilisation d'une telle Table donnera des mécomptes. Aussi, pour ses propres Tables, Kendall a-t-il indiqué les zones dont il déconseillait l'emploi au lecteur travaillant sur un petit échantillon. Il est clair que pour des centaines ou des milliers de nombres se suivant sur les Tables, le hasard ne peut fournir de pareilles rencontres.

Remarque - Pourquoi n'utilise-t-on pas les annuaires téléphoniques pour composer une table de nombres aléatoires ?

Kendall a pensé effectivement à cette ressource; et il indique qu'elle a dû être écartée (en 1938) parce que les chiffres 5 et 9 entrent bien plus rarement que les autres dans les numéros de téléphone anglais (on peut confondre à l'oreille five et nine, ce serait là une des raisons de ce phénomène).

Emploi de ces Tables - Soit une cargaison de 3.425 sacs de sucre. On va utiliser la Table II de Tippett (ci-jointe) pour tirer au sort un échantillon de 15 de ces sacs. On part de l'angle supérieur gauche, par exemple, on lit les nombres de 4 chiffres et on ne retient que ceux inférieurs à 3.426, jusqu'à ce qu'on ait trouvé 15.

1.254; 3.262; 0.126; 2.372; 0.357; 2.510; 1.658; 1.319; 1.983; 0.330; 1.614; 2.096; 0.511; 0.524; 3.311.

Reste à mettre ces nombres dans l'ordre croissant :

0.126; 0.330; 0.357; 0.511; 0.524; 1.254; 1.319; 1.614; 1.658; 1.983; 2.096; 2.372; 2.510; 3.262; 3.311.

Conclusion - On arrêtera au passage et pèsera les 15 sacs dont les numéros sont ces 15 nombres.

Tableau II

Nombres aléatoires de Tippett

1254	2858	7358	4024	3684	8485	2617	5488
5443	4911	0922	7134	4798	1311	8701	2210
3262	2322	4112	9877	4776	4512	1746	2593
7809	0297	8956	2158	7780	0753	1252	7181
6862	4194	3596	5072	4473	3099	0729	4950
9179	3814	9153	2127	6745	9646	8105	3133
5317	0986	0633	6480	4834	8710	8829	8572
0126	4777	8034	9217	2128	2232	5039	8637
2372	7774	9446	7178	8403	3971	0899	5274
0357	5276	3999	0261	9255	5780	5728	0032
7855	9707	5259	4263	9878	4918	0987	9118
2510	4254	1543	0224	0112	6523	8667	4707
6639	1913	3120	9149	6145	5895	0726	3883
6769	1435	9107	4762	9902	3764	7388	2729
4527	8000	8648	3366	7945	4847	4317	9636
5699	9883	2456	0893	4132	6668	0799	6137
4160	1445	2887	0724	1294	8988	1527	1467
4506	2474	3590	5308	7640	7128	1023	2418
4645	0613	9846	4458	5666	7671	1184	2328
6686	3544	9828	9187	0506	6473	5356	8940
6503	0329	7899	8211	0852	8066	5706	1940
1658	4288	1856	5319	3512	8981	7468	3836
1319	1204	3344	8886	3846	6777	1700	0323
5030	4027	2077	9812	8645	3290	7191	6152
7866	2029	5156	2003	2940	0237	7670	2852
1983	8992	1017	7263	7699	4151	8132	7271
9944	0845	7468	3936	8002	0857	5784	4480
0330	9913	4990	7790	6932	0871	1988	9881
9903	7914	4138	6826	0230	1337	7413	8840
1614	7862	9500	4109	1037	2978	6075	0971
2096	1164	3788	6257	0632	0693	2263	5290
0511	0229	5951	6808	1409	7624	4903	4692
0524	4056	9140	6371	2099	8290	3611	6501
7381	7386	6568	1568	4160	0429	3488	3741
3311	3733	7882	6985	7874	7264	4587	9591
6874	2534	7485	9596	9086	2701	4967	1588
8987	3121	3628	0372	1059	6339	8973	4218
9205	2358	0393	3196	1612	8397	4390	0187
9749	7893	7076	9791	1530	8127	4474	1895
2183	2109	2874	5733	1567	7764	4939	9919
6926	3085	2079	3330	4432	9524	0327	9640
3373	3567	0371	5932	3923	7250	8578	5869
9771	5542	4715	5527	3763	3167	3679	4399
9353	5576	5474	0190	7274	6993	3920	7272
5761	6301	3558	6205	3012	6195	8461	2046

Tableau II (suite et fin)

Nombres aléatoires de Tippett

6850	8122	4455	2940	9945	9688	3588	8311
6616	6760	4938	0066	0391	8898	4753	7402
2633	4255	8755	8434	4334	8992	1260	0547
6837	4898	9527	3526	6536	7703	9941	8454
4525	7772	3888	7243	1330	7115	4897	6618

2/ Le sondage systématique

Avec un peu de bon sens, on aurait plutôt imaginé de peser 15 sacs dont les rangs sont en progression arithmétique, par exemple, de raison 228 (car 3.425 = 228 × 15 + 5) - par exemple : 100; 100 + 228; 100 + 684; etc.

Première question - Est-ce bien un procédé probabiliste ? Nous verrons que oui, pourvu que le 1er nombre de la progression soit désigné au hasard.

Seconde question - Est-ce un bon procédé ? Cela dépend. A priori il est intéressant, car les nombres aléatoires peuvent être, eux, répartis de façon très irrégulière. C'est ainsi que, sur les 15 nombres tirés ci-dessus dans la Table de Tippett, il ne s'en trouve aucun entre 524 et 1.254, - ni entre 2.510 et 3.262; au contraire, on a trouvé 511 et 524; 1.614 et 1.658; 1.893 et 2.096. La progression arithmétique est régulière.

Ceci serait surtout un avantage si les sacs sortaient du bateau dans l'ordre des poids croissants (ou décroissants). Avec un échantillon systématique de 15 sacs, on serait assuré d'en avoir 5 petits, 5 moyens, et 5 gros. La précision du sondage en serait certainement améliorée. S'ils sortent dans un ordre quelconque on ne peut affirmer qu'il en soit ainsi.

En revanche, cette régularité serait certainement défectueuse si les poids des sacs (dans l'ordre où ils se présentent) offraient des fluctuations périodiques, dont la période soit voisine de 228 sacs ou d'une fraction de 228 sacs; suivant le premier sac retenu, on pourrait alors n'avoir que des gros sacs dans l'échantillon, ou, au contraire, n'en avoir que des petits.

En résumé, la qualité d'un sondage systématique dépend de l'ordre dans lequel les unités de sondage sont numérotées. Si cet ordre est au hasard, on démontre que le sondage systématique est tout à fait équivalent à un sondage au hasard.

3/ Le sondage en grappes

On a déjà vu un exemple de "grappe" : le wagonnet du téléférique était une "grappe" de morceaux de charbon. Autre exemple : des boîtes de conserves (jus de fruits) sont livrées en caisses de 12; la caisse est une "grappe" de 12 boîtes.

Problème - A-t-on intérêt à adopter comme unité de sondage la grappe ou l'élément composant la grappe ? Pour le charbon, il n'y avait pas de problème. Mais pour les conserves, on peut hésiter entre la caisse et la boîte.

En général, si l'on veut vérifier la qualité d'un wagon de jus de fruits en prélevant un échantillon de 12 boîtes, il sera beaucoup plus simple de n'ouvrir qu'une caisse et les 12 boîtes de cette caisse; mais il est évident qu'on a un échantillon meilleur en prélevant 12 boîtes au hasard, dut-on ouvrir pour cela 12 caisses différentes.

Il existe cependant quelques cas où c'est exactement le contraire.

4/ Le sondage stratifié

Supposons qu'une usine reçoive du charbon de plusieurs prix et de plusieurs fournisseurs distincts et veuille évaluer le pouvoir calorifique de ce charbon. Il est clair qu'on a intérêt à procéder à des petits sondages séparés et indépendants dans chaque catégorie de charbon livré par chaque fournisseur (en prélevant au besoin des fractions de l'ensemble différentes suivant le prix et le fournisseur); et ceci bien qu'on ne désire pas connaître les différences entre les charbons des diverses origines, simplement parce qu'on a le droit de penser que les livraisons d'un prix et d'un fournisseur donné sont plus homogènes entre elles que si l'on mélange les catégories.

Chaque catégorie (même prix, même fournisseur) constitue ce qu'on appelle une strate.

5/ Le sondage à plusieurs degrés

Revenons à notre wagon contenant des caisses, chaque caisse renfermant 12 articles. On va appeler les caisses : "unités du 1er degré" - les articles : "unités du 2ème degré".

Tirer un échantillon de caisses, puis dans ces caisses un échantillon d'articles constitue un tirage à 2 degrés (1er degré = tirage des caisses).

On peut, bien entendu, faire les tirages à 3 degrés, 4 degrés, ...

6) Le sondage à plusieurs phases

Il s'agit de quelque chose de tout à fait différent (encore qu'il existe des cas limites où la distinction est difficile). Revenons aux boîtes de jus de fruits. Supposons que le contrôle porte, d'une part, sur le poids brut (emballage compris), d'autre part, sur la qualité et le poids net. Le premier contrôle se fait par simple pesée et on a intérêt à y consacrer un grand échantillon; le second contrôle entraîne l'ouverture et la perte de boîtes, on le limite à un petit échantillon. Dans ces conditions le sondage à 2 phases consiste à tirer un grand échantillon de boîtes que l'on pèse, puis (parmi elles) un "sous-échantillon" de boîtes que l'on ouvre.

7/ L'emploi d'une formule d'estimation utilisant des renseignements supplémentaires

Considérons, par exemple, la cargaison de sacs de sucre dont il a déjà été question; mais supposons que le poids de chaque sac soit indiqué, soit par une étiquette, soit sur un état récapitulatif avec un numéro d'ordre reporté sur le sac.

On connaît ainsi que le poids total déclaré de la cargaison.

Mais ce poids déclaré n'est pas forcément exact; et s'il l'était au départ, l'humidité a pu varier, ou bien on a pu vider en partie les sacs en cours de route.

On procède donc à la pesée d'un échantillon de sacs au moment de la réception. On va relever pour cet échantillon à la fois le poids déclaré et le poids actuel réel.

Soit y_i le poids déclaré et x_i le poids actuel réel du sac n°i. On peut estimer le poids actuel réel $\sum x$ de la cargaison entière par :

$$\frac{S x_i}{S y_i} \sum y_i = \frac{\sum y_i}{S y_i} . S x_i$$

On comparera cette formule à celle valable quand il n'y a pas de poids déclaré

$$m\bar{X}_i = m \frac{S x_i}{n} = \frac{m}{n} S x_i$$

Dans un cas, on a extrapolé simplement par le nombre de sacs; dans l'autre on a recours, pour l'extrapolation, au poids déclaré.

Une méthode plus savante consiste à porter sur un graphique les points $(x_i\ y_i)$ des sacs échantillons, à ajuster sur ces points une droite de régression et à estimer $(\sum x_i)$ à l'aide de la formule de régression de cette droite.

CONCLUSION

Les procédés probabilistes dont on dispose sont en somme assez minces. Il est nécessaire de les combiner avec habileté pour aboutir à des plans de sondage pratiques, à la fois suffisamment simples et suffisamment précis. Néanmoins on arrive vite à des situations compliquées.

A titre d'exemple, donnons quelques indications sur la méthode de la pesée géométrique, méthode employée couramment pour évaluer sur pied (et acheter) une récolte de betteraves sucrières. On connaît la superficie A en hectares, ares et centiares du champ dont on achète la récolte (voir le cadastre); on veut donc évaluer le rendement. Pour cela on évalue le poids moyen P d'une betterave et le nombre moyen N de betteraves par hectare. La récolte est A P N .

N et P sont évalués par une opération complexe :

Pour N, c'est un sondage à 2 degrés : 1er degré; échantillon de lignes (les betteraves sont plantées en lignes); 2ème degré : un échantillon de chaînées (avec une chaîne d'arpenteur de 20 mètres) sur ces lignes. On compte les betteraves sur les chaînées-échantillons.

Pour P, c'est un sondage à 3 degrés : 1er et 2ème degrés = idem; 3ème degré = on détermine un petit échantillon de betteraves (par exemple, 1 sur 5) par chaînée-échantillon. Au total les formules mathématiques valables seraient très compliquées.

Cette opération complexe est aussi une sorte de sondage à 2 phases; la 1ère phase consistant à désigner et compter un grand échantillon de betteraves, et la 2ème phase comportant la désignation et l'arrachage d'un sous-échantillon de ces betteraves.

L'opération conduit finalement à arracher et peser une centaine de betteraves à l'hectare, sur un champ qui peut en compter 40 ou 80 000 à l'hectare; c'est un sondage avec une fraction de sondage de l'ordre de 1 à 2 pour mille.

DEUXIÈME PARTIE

ÉTUDE THÉORIQUE DES PRINCIPALES MÉTHODES DE SONDAGE

Quand on traite des applications de la statistique à l'Industrie, on admet, en général, qu'on a affaire à un échantillon de pièces tirées au hasard (d'un univers infiniment grand) par des tirages indépendants. On ne se préoccupe guère du processus par lequel les éléments constituant l'échantillon ont été obtenus; car lorsque ce processus n'est pas l'équivalent à des tirages de boules dans une urne de Bernoulli, les formules mathématiques qu'on emploie couramment ne sont plus valables.

I - LE PROCESSUS DE SONDAGE AU HASARD ELEMENTAIRE

Les tirages d'échantillon ont lieu dans des ensembles d'éléments en nombre fini (et non pas des univers infinis); de ce fait, les éléments (ou unités de sondage) ne sont pas tirés indépendamment les uns des autres.

Soit i = 1, 2, ..., m, un numéro d'ordre (ou indice) affecté à toutes les unités de sondage de l'ensemble - soit X un caractère étudié prenant les valeurs :

$$x_1,\ x_2,\ \ldots,\ x_i,\ \ldots,\ x_m$$

de moyenne $\bar{x} = \sum \frac{x_i}{m}$. Supposons qu'on se propose d'estimer $\bar{x}$ à l'aide de :

$$\bar{X} = S\ \frac{x_i}{n}$$

moyenne calculée sur un échantillon de n éléments tirés au hasard par la méthode élémentaire, mais sans remettre les boules dans l'urne; (la lettre S désigne la sommation étendue à l'échantillon) : ce qu'on appelle un tirage exhaustif.

On sait que, dans le cas de tirages de Bernoulli (où l'on remet la boule dans l'urne après chaque tirage) on a (σ étant l'écart-type des x_i) :

$$E(\bar{X}) = \bar{x}$$

$$V(\bar{X}) = \frac{\sigma^2}{n}$$

Calculons ces deux quantités dans nos nouvelles hypothèses. Il vient[1]

---

(1) Rappelons, en effet, que l'espérance mathématique E( ) est un opérateur linéaire, c'est-à-dire que E(A + B) = E(A) + E(B); et E(kA) = kE (A).

$$E(\bar{X}) = E\left(\frac{S x_i}{n}\right) = \frac{1}{n} \qquad SE(x_i) = \frac{1}{n} . n . \bar{x} = \bar{x}$$

Pour calculer $V(\bar{X})$, on va donner une méthode de calcul d'aspect compliqué, mais qui présente l'avantage de s'appliquer à un grand nombre d'autres problèmes.

Considérons des variables aléatoires $A_1; A_2, \ldots, A_m$, égales à 1 ou à 0. $A_i = 1$ si l'unité i fait partie de l'échantillon; $A_i = 0$ dans le cas contraire. On a :

$$\text{Probabilité qu } (A_i = 1) = \frac{n}{m}$$

puisque les tirages sont tels que chaque unité ait même probabilité d'être tirée que les autres. Donc :

$$E(A) = 1. \frac{n}{m} + 0\left(1 - \frac{n}{m}\right) = \frac{n}{m}$$

Nous aurons besoin, en outre, de $E(A_i A_j)$ pour $i \neq j$.

La probabilité que $A_i A_j$ soit égal à 1 serait $\left(\frac{n}{m}\right)^2$ si les n tirages étaient indépendants; mais comme on procède par tirages exhaustifs, on a :

$$\text{Prob.}(A_i A_j = 1) = \frac{C_n^2}{C_m^2} = \frac{n\ (n - 1)}{m\ (m - 1)}$$

d'où $$E(A_i A_j) = 1. \frac{n\ (n - 1)}{m\ (m - 1)} + 0 \left[1 - \frac{n\ (n - 1)}{m\ (m - 1)}\right] = \frac{n\ (n - 1)}{m\ (m - 1)}$$

Revenons au calcul de $V(\bar{X})$:

$$V(\bar{X}) = E(\bar{X} - \bar{x})^2 = E(\bar{X}^2 - 2\bar{X}\bar{x} + \bar{x}^2)$$

$$= E(\bar{X}^2) - 2\bar{x}^2 + \bar{x}^2$$

$$= E(\bar{X}^2) - \bar{x}^2$$

$$E(\bar{X}^2) = E(\sum A_i x_i)^2$$

$$= \frac{1}{n^2} E\left[\sum_i A_i^2 x_i^2 + \sum A_i A_j x_i x_j)\right]$$

Remarquons que $A_i^2 = A_i$ (0 ou 1); donc :

$$E(A_i^2) = E(A_i) = \frac{n}{m}$$

$$E(\bar{X}^2) = \frac{1}{n^2}\left[\frac{n}{m} \sum_i x_i^2 + \frac{n\ (n - 1)}{m : (m - 1)} \sum_{ij} x_i x_j\right]$$

Introduisons les 2 premiers moments des x_i

$$\sum x_i = m\bar{x}$$

$$\sum x_i^2 = m(\bar{x}^2 + \sigma^2)$$

en les combinant, il vient :

$$\sum_{ij} x_i x_j = (\sum x_i)^2 - \sum x_i^2 = m\bar{x}^2 - m(\bar{x}^2 + \sigma^2)$$

Portons dans $E(\bar{X}^2)$. On constate que le terme en $\bar{x}^2$ se réduit à $\bar{x}^2$.

$$E(\bar{X}^2) = \bar{x}^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{m-1}\right)$$

d'où

$$V(\bar{X}) = \frac{\sigma^2}{n}\,\frac{m-n}{m-1}$$

Ce résultat est satisfaisant pour l'esprit, car :

1) Si $n = 1$, il vient : $V(\bar{X}) = \sigma^2$ (tirage d'une seule unité)

2) Si $n = m$, il vient : $V(\bar{X}) = 0$

Quand l'échantillon grandit au point de recouvrir tout l'ensemble sondé, il est clair que $\bar{X} = \bar{x}$ et, par conséquent, que $\bar{X}$ n'est plus aléatoire; il est donc normal que sa variance soit nulle.

3) Si m est beaucoup plus grand que n, il vient :

$$V(\bar{X}) \# \frac{\sigma^2}{n},$$

de sorte que le cas des tirages bernoulliens est une première approximation du cas des sondages exhaustifs. Par exemple, si m est de l'ordre du million et n de l'ordre de la centaine ou du millier, on peut parfaitement se placer dans le cas de Bernoulli.

4) Dès que $n > 1$, on a $V(\bar{X}) < \frac{\sigma^2}{n}$ car : $m - n < m - 1$.

Quand on procède comme Bernoulli, on s'expose à tirer 2 fois ou plus la même boule; de sorte qu'il est clair que l'estimation $\bar{X}$ est moins précise que si l'on s'arrange pour avoir, à coup sûr, n valeurs x_i différentes dans l'échantillon. Estimation moins précise correspond à variance plus grande.

En pratique : si $\frac{n}{m} = \frac{1}{5}$, on a (en confondant m et m - 1) :

$$V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{4}{5}$$

on ne peut plus dire que le terme correctif $\left(\frac{m-n}{m-1}\right)$ soit négligeable.

Estimation de $\bar{x}$ à l'aide de sondages au hasard élémentaire

On rappelle que la précision d'une telle estimation se caractérise par l'inverse de la variance :

$$\frac{1}{V(\bar{X})}$$

et qu'on compare 2 estimations différentes en formant le quotient des inverses des variances (quotient appelé efficacité ou efficience).

En se plaçant dans le cas de grands échantillons, où l'on peut admettre que :

$$\frac{\bar{X} - \bar{x}}{\sqrt{V(\bar{x})}}$$

suit une loi normale, proposons-nous d'estimer $\bar{x}$ avec une erreur relative de 2%, par exemple, qui n'ait que 5 chances sur 100 d'être dépassée.

Compte tenu de la loi normale, ceci conduit à écrire :

$$1.96\sqrt{V(\bar{X})} \leqslant \frac{2}{100}\,\bar{x}$$

$$1.96\,\frac{\sigma}{\sqrt{n}} \leqslant \frac{2}{100}\,\bar{x}$$

comme on ne connaît pas ni $\bar{x}$, ni σ, on utilisera leurs estimations, soit :

$$\bar{X} \text{ et } S = \sqrt{\frac{S(x_i - \bar{\bar{X}})^2}{n - 1}}$$

ce qui conduit pratiquement à

$$\left(100\,\frac{s}{\bar{X}}\right)^2 \leqslant n,$$

ceci suppose que l'on ait, a priori, une idée de l'ordre de grandeur de s et de $\bar{X}$, sinon cette formule permettra de corriger l'effectif utilisé afin d'obtenir la précision cherchée.

Dans le cas où l'on veut estimer la fréquence p des individus qui possèdent un caractère donné, par la fréquence f observée dans l'échantillon, l'approximation de la loi binomiale par la loi normale, valable si n est grand et p non trop petit, par exemple si $np(1 - p) > 20$, on obtiendrait avec les mêmes exigences, la condition :

$$1.96\sqrt{\frac{p(1 - p)}{n}} \leqslant \frac{2}{100}\,p$$

soit pratiquement :

$$n \geqslant 10^4\,\frac{1 - p}{p}$$

qui ne pourra être utilisée que si on connaît l'ordre de grandeur de p.

Si dans ce cas, on s'imposait une erreur absolue de 0,02 avec une probabilité 0,05 de n'être pas dépassée dans l'estimation de p, on obtiendrait la condition

$$2\sqrt{\frac{p(1 - p)}{n}} \leqslant 0,02,$$

certainement réalisée si

$$n \geqslant 2500,$$

le produit $p(1 - p)$ étant maximum pour $p = 1 - p = \frac{1}{2}$.

Caractères simultanés mesurés sur la même unité

Soit X et Y des caractères; leurs valeurs sur les unités 1 2 ... i ... m; sont appelées :

$$x_1 \quad x_2 \quad \ldots \quad x_i \quad \ldots \quad x_m$$

$$y_1 \quad y_2 \quad \ldots \quad y_i \quad \ldots \quad y_m$$

Si l'on étudie cette distribution à l'aide d'un échantillon de n unités, on sait que les caractéristiques essentielles en sont :

$$\bar{x} \quad \bar{y} \quad \sigma_x \quad \sigma_y \quad \text{et} \quad \rho,$$

le coefficient de corrélation ρ étant défini par

$$\rho\, \sigma_2 \sigma_y = E[(x_i - \bar{x})(y_i - \bar{y})]$$

où le second nombre est appelé la covariance des x_i et y_i.

On sait les estimer à l'aide respectivement de $\bar{X}$, $\bar{Y}$, s_x, s_y, et r, ce dernier étant tel que :

$$r\, s_x s_y = \frac{S(x_i - \bar{X})(y_i - \bar{Y})}{n - 1}$$

où le 2ème nombre est ce qu'on appelle la covariance de l'échantillon. C'est une extension naturelle de :

$$s_x^2 = \frac{S(x_i - \bar{X})^2}{n - 1}$$

Enfin, le coefficient de corrélation ρ' entre $\bar{X}$ et $\bar{Y}$ n'est pas autre chose que ρ lui-même. Par définition, en effet, on a :

$$\rho' \sigma_{\bar{x}} \sigma_{\bar{y}} = E(\bar{X} - \bar{x})(\bar{Y} - \bar{y})$$

avec

$$\rho' \sigma_{\bar{x}} \sigma_{\bar{y}} = \rho' \frac{\sigma_x}{\sqrt{n}} \frac{\sigma_y}{\sqrt{n}} = \rho' \frac{\sigma_x \sigma_y}{n}$$

et

$$E[(\bar{X} - \bar{x})(\bar{Y} - \bar{y})] = \frac{1}{n} E[(x_i - \bar{x})(y_i - \bar{y})]$$

généralisation de

$$E[(\bar{X} - \bar{x})^2] = \frac{1}{n} E[(x_i - \bar{x})^2]$$

Que deviennent ces résultats dans le cas d'un sondage exhaustif ?

On démontre avec la méthode qui a servi pour

que :

$$V(\bar{X}) = \frac{\sigma^2}{n} \frac{m - n}{m - 1}$$

$$Cov(\bar{X}, \bar{Y}) = \frac{\rho \sigma_x \sigma_y}{n} \frac{m - n}{m - 1}$$

où ρ est toujours le coefficient de corrélation entre les x_i et les y_i.

II - LES PROCESSUS DE SONDAGE SYSTEMATIQUE ET DE SONDAGE EN GRAPPES

Pour l'étude mathématique de ces 2 processus, nous sommes amenés à introduire une notion théorique nouvelle : celle d'"interclass corrélation".

Considérons deux séries de valeurs accouplées, supposées centrées :

α	β	γ			d'écart-type σ
α'	β'	γ	δ'	...	d'écart-type σ'

on sait que le coefficient de corrélation entre les deux lignes est :

$$\frac{E(\alpha\ \alpha')}{\sigma\sigma'}$$

On va généraliser cette notion.

Considérons des séries de valeurs en nombre quelconque, disposées en tableau rectangulaire comme suit :

a	b	c	d	...
a'	b'	c'	d'	...
a"	b"	c"	d"	...

La moyenne générale de tous ces nombres est aussi la moyenne des moyennes de colonne, car :

$$a + a' + a'' + b + b' + ; \ldots = (a + a' + a'') + (b + b' + b'') + \ldots$$

Retranchons de tous les nombres du tableau cette moyenne générale, il vient :

α	β	γ	δ	...
α'	β'	γ'	δ'	...
α''	β''	γ''	δ''	...

On définit l'"interclass correlation" comme suit, en supposant qu'il n'y ait que 3 lignes :

$$\frac{E\left[\frac{\alpha\alpha' + \alpha'\alpha'' + \alpha''\alpha}{3}\right]}{\sigma^2} = \bar{\rho}$$

σ étant l'écart-type du tableau pris dans son ensemble.

On généralise facilement au cas d'un nombre quelconque de lignes.

Variance des moyennes de colonnes

Pour la suite, on aura besoin de connaître la variance des moyennes de colonnes d'un tel tableau, c'est-à-dire :

$$\sigma_1^2 = E\left(\frac{\alpha + \alpha' + \alpha''}{3}\right)^2$$

$$= E\left(\frac{\alpha^2 + \alpha'^2 + \alpha''^2}{9}\right) + 2E\left(\frac{\alpha\alpha' + \alpha'\alpha'' + \alpha''\alpha}{9}\right)$$

$$= \frac{3\sigma^2}{9} + \frac{2\sigma^2}{3}\,\bar{\rho}$$

$$= \frac{\sigma^2}{3}(1 + 2\bar{\rho})$$

Dans le cas d'un tableau à l lignes et c colonnes, on aurait évidemment :

$$\sigma_1^2 = \frac{\sigma^2}{l}[1 + \bar{\rho}(l + 1)]$$

II. 2/ Le processus de sondage systématique

On a vu dans la 1ère leçon qu'on appelait sondage systématique un sondage où l'échantillon est constitué d'éléments dont les numéros d'ordre forment une progression arithmétique.

Supposons que la base de la progression soit 10 et la raison 15. Disposons comme suit la population (tableau à 15 colonnes) :

x_1	x_2	...	x_{10}	...	x_{15}
x_{16}	x_{17}	...	x_{25}	...	x_{30}
x_{31}	x_{32}	...	x_{40}	...	x_{45}
.....					

Pour simplifier, on va supposer que le nombre m d'unités est exactement égal à (lc) (l étant l'effectif n de l'échantillon et c étant ici le nombre 15). L'échantillon systématique est donc formé d'une colonne du tableau; de sorte que la variance de $\bar{X}$ est :

$$\sigma_1^2 = \frac{\sigma^2}{l}[(1 + \bar{\rho}(l - 1)]$$

Si les (n) unités avaient été tirées au hasard, la variance de $\bar{X}$ aurait été :

$$\frac{\sigma^2}{n}\frac{m - n}{m - 1} = \frac{\sigma^2}{l}\frac{lc - l}{lc - 1}$$

L'échantillon systématique est plus précis que l'échantillon stratifié lorsque :

$$1 + \bar{\rho}\,(l - 1) \leqslant \frac{lc - l}{lc - 1}$$

c'est-à-dire :

$$\bar{\rho} < \frac{-1}{lc - 1}$$

En pratique, le nombre $\frac{1}{lc - 1}$ est sensiblement égal à zéro, donc

si $\bar{\rho} < 0$: le sondage systématique est avantageux

si $\bar{\rho} > 0$: le sondage systématique est désavantageux.

Remarques -

1) On verrait facilement que (en tout état de cause) $\bar{\rho}$ ne peut être supérieur à 1, ni inférieur à $\frac{-1}{l - 1}$.

2) On pourrait voir que, lorsque les éléments de l'univers sont numérotés dans un ordre tel que les x_i décroissent constamment (ou croissent constamment) on a bien $\bar{\rho} < 0$. De même, on pourrait voir que, lorsque les x_i présentent un caractère périodique, de période égale à c ou voisine de c, on a bien

$$\bar{\rho} \geqslant 0$$

II. 3/ Le processus de sondage par grappes

On a vu, précédemment qu'une grappe était un groupement naturel d'éléments, par exemple, constitué par toutes les pièces livrées dans une caisse. En pareil cas on a manifestement un intérêt pratique à ouvrir k caisses et, si celles-ci renferment chacune l pièces, de soumettre à l'analyse toutes ces pièces (au nombre de kl) plutôt que de tirer au sort kl pièces par sondage élémentaire en s'exposant à avoir à ouvrir près de kl caisses distinctes. Mais on peut, d'autre part, penser que, par exemple, si la caisse est tombée ou a été mouillée en cours de route, on a des chances que toutes les pièces qu'elle renferme soient abîmées.

Disposons en un tableau rectangulaire les mesures X de toutes les pièces de l'ensemble, en plaçant dans une même colonne celles des pièces d'une même caisse (= d'une même grappe). On suppose ici (pour simplifier) que toutes les grappes comprennent le même nombre l de pièces.

$$\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1c} \\ x_{21} & x_{22} & \dots & x_{2c} \\ \dots & \dots & \dots & \dots \\ x_{l1} & x_{l2} & \dots & x_{lc} \end{array}$$

c est à la fois le nombre de caisses et de colonnes.

La variance de $\bar{X}$ calculée sur un échantillon de k caisses, c'est-à-dire de k colonnes, est évidemment :

$$V_g(\bar{X}) = \frac{\sigma_1^2}{k} \frac{c - k}{c - 1}$$

où σ_1 est l'écart-type des moyennes de colonne; d'où

$$V_g(\bar{X}) = \frac{\sigma^2}{kl} [(1 + \bar{\rho}(l - 1)] \frac{c - k}{c - 1}$$

Cette expression doit être comparée à la variance d'un échantillon de (kl) pièces prélevées par sondage élémentaire parmi les (cl) pièces de l'ensemble.

$$V(\bar{X}) = \frac{\sigma^2}{kl} \frac{cl - kl}{cl - 1} = \frac{\sigma^2}{k} \frac{c - k}{cl - 1}$$

L'échantillon en grappes sera plus avantageux en ce qui concerne la précision que l'échantillon ordinaire si l'on a :

$$[1 + \bar{\rho}(l - 1)] \frac{1}{l(c - 1)} < \frac{1}{cl - 1}$$

d'où

$$\bar{\rho} < - \frac{1}{cl - 1}$$

C'est la valeur limite qui a été déjà trouvée pour l'échantillon systématique (et qui est évidemment voisine de zéro).

En résumé : $- \frac{1}{l - 1} < \bar{\rho} < - \frac{1}{cl - 1}$: précision améliorée par les grappes

$- \frac{1}{cl - 1} < \bar{\rho} < 1$: précision diminuée par les grappes.

En général, la corrélation $\bar{\rho}$ sera positive et il y aura une certaine corrélation entre les mesures X des éléments d'une même grappe (par exemple, les éléments d'une même caisse seront détériorés en tout ou partie).

Toutefois, il existe des cas où l'on a moins de chances de rencontrer 2 pièces à rebuter dans une grappe, disons de 12 pièces que dans 12 pièces au hasard.

Voici un exemple schématique emprunté à l'industrie :

a) On peut imaginer un tour déréglé : toutes les pièces qui se suivent sortent avec 0,2 mm de trop; on les emballe dans l'ordre dans lequel elles ont été produites, il y a une corrélation positive entre les éléments des grappes (entre les pièces d'une même boîte).

b) Mais on peut imaginer, en revanche, un appareil mal réglé qui, à peu près toutes les 12 pièces commet une erreur; on met les pièces en lots disons de 6 dans l'ordre de leur production. Ainsi un lot sur 2 contient une pièce défectueuse. Il y a corrélation négative entre les éléments des lots; et l'échantillon en grappes (les grappes étant constituées par les lots) est alors plus avantageux que l'échantillon au hasard élémentaire.

Dans le domaine des statistiques humaines; l'échantillon de ménages (grappes d'individus) est de même plus avantageux que l'échantillon de personnes pour étudier le caractère sexe ou le caractère âge.

III - LE PROCESSUS DE SONDAGE STRATIFIE

On appelle strate un sous-ensemble appartenant à l'ensemble à sonder. On est amené à définir des strates lorsqu'on pense que certains sous-ensembles faciles à distinguer sont plus homogènes que l'ensemble total.

On a déjà cité l'exemple des tas de charbon. En voici d'autres :

a) Une cargaison comprend des sacs de diverses origines (l'origine est indiquée sur l'étiquette) et on a des raisons particulières de penser que les sacs provenant d'un certain pays pèsent systématiquement moins que ne l'indique l'étiquette : on va constituer une strate spéciale de ces sacs lorsqu'on voudra estimer par sondage la correction à apporter aux poids déclarés pour obtenir les poids réels. La provenance est le caractère stratificateur.

b) On doit contrôler sur échantillon une livraison de 50 grandes caisses et de 1 000 petites, d'un même produit. On constituera une première strate avec les grandes caisses et on en ouvrira 10 (une sur cinq); une seconde strate comprendra les petites caisses et on en ouvrira 10 également (une sur cent). On utilise la taille des caisses comme caractère stratificateur.

Remarque - On n'a pas besoin de regrouper matériellement les unités de sondage pour former des strates; il suffit qu'on dispose d'un état nominatif de ces unités, indiquant pour chacune sans ambiguïté à quelle strate elle appartient.

°
° °

Soient $U_1, U_2, \ldots, U_h, \ldots, U_l$ les strates composant l'univers. La strate U_h est formée de m_h éléments et la moyenne du caractère X dans cette strate est égale à $\bar{x}_h$. Cette quantité $\bar{x}_h$ sera estimée (à l'aide d'un procédé probabiliste quelconque) par $\bar{X}_h$ satisfaisant à

$$E[\bar{X}_h] = \bar{x}_h$$

On se propose d'estimer la moyenne générale de l'univers

$$\bar{x} = \frac{1}{m} \sum_{h=1}^{l} m_h \bar{x}_h \qquad \left(m = \sum_{h=1}^{l} m_h \right)$$

par une combinaison linéaire $\bar{X}$ adéquate des $\bar{X}_h$.

$$\bar{X} = \sum_{h=1}^{l} \lambda_h \bar{X}_h$$

Imposons aux λ_h la relation

$$E[\bar{X}] = \bar{x}$$

qui traduit que l'estimation $\bar{X}$ de x est "juste" ou "absolument correcte", ou "sans biais" ("unbiased"), ce qui revient à :

$$E\left[\sum_{h=1}^{l} \lambda_h \bar{X}_h\right] = \sum_{h=1} \frac{m_h}{m} \bar{x}_h$$

ou

$$\sum_{h=1}^{l} \lambda_h \bar{x}_h = \sum_{h=1}^{l} \frac{m_h}{m} \bar{x}_h$$

soit $\lambda_h = \frac{m_h}{m}$ quel que soit h

et $$\bar{X} = \sum_{h=1}^{l} \frac{m_h}{m} \bar{X}_h$$

Supposons qu'on ait prélevé un échantillon de n_h unités sur m_h dans la strate h.

On peut vérifier que l'on peut poser :

$$\bar{X}_h = \sum_{i=1}^{n_h} \frac{x_{hi}}{n_h}$$

x_{hi} étant la valeur prise par X sur l'unité n°i de la strate n°h, d'où :

$$\bar{X} = \sum_{h=1}^{l} \frac{m_h}{m} \sum_{i=1}^{n_h} \frac{x_{hi}}{n_h}$$

ou encore $$\bar{X} = \frac{1}{m} \sum_{h=1}^{l} \frac{m_h}{n_h} \sum_{i=1}^{n_h} x_{hi}$$

Fraction sondée - On appelle fraction sondée dans la strate U_h le rapport $f_h = \frac{n_h}{m_h}$;

d'où : $$\bar{X} = \frac{1}{m} \sum_{h=1}^{l} \frac{1}{f_h} \sum_{i=1}^{n_h} x_{hi}$$

Echantillon à fraction sondées égales (ou "échantillons représentatifs") - On appelle ainsi l'échantillon obtenu dans le cas où $f_h = f$ quel que soit h. On a alors :

$$\bar{X} = \frac{1}{mf} \sum_{h=1}^{l} \sum_{i=1}^{n_h} x_{hi}$$

Ce cas est simple et intéressant puisque le sondage peut se dépouiller comme un recensement : on fait la somme totale des valeurs de x_{hi} pour tout h et pour tout i. Cependant, malgré le mot "représentatif", il n'est pas nécessaire, pour avoir un échantillon valable, d'avoir des fractions sondées égales; il suffit d'adapter la formule d'estimation $\bar{X}$ au sondage effectué (et, dans le cas du sondage stratifié, de pondérer les $\bar{X}_h$ par les effectifs m_h des strates).

Si par exemple la strate n°1 comprend 50 caisses, la strate n°2, 1 000 caisses, on obtient une estimation correcte $\bar{X}$ de $\bar{x}$ avec :

$$\bar{X} = \frac{1}{1\,050} (50\ \bar{X}_1 + 1\,000\ \bar{X}_2)$$

Problème - Quels que soient les f_h on obtient une estimation correcte $\bar{X}$ et x. Mais quelles sont les fractions de sondage qui sont les plus intéressantes ? Pour résoudre ce problème il faut passer aux conditions du 2ème ordre : il ne suffit pas que $E[\bar{X}] = \bar{x}$ (et cela n'est même pas nécessaire : il faut seulement que $E[\bar{X}]$ tende vers $\bar{x}$); il faut en outre que $\bar{X}$ soit une estimation précise de $\bar{x}$, c'est-à-dire que $V[\bar{X}]$ soit aussi faible que possible.

Calcul de $V[\bar{X}]$ - Rappelons les règles de calcul suivantes sur l'opérateur V (A et B étant deux aléatoires, k étant une constante) :

$$V[A + B] = V[A] + V[B] + 2\,\mathrm{cov}[A, B]$$

$$V[kA] = k^2 V[A]$$

Il en résulte ici, en supposant que les tirages dans les différentes strates sont indépendants, c'est-à-dire que :

$$\mathrm{cov}[\bar{X}_h, \bar{X}_{k\neq h}] = 0$$

et que les n_h unités tirées dans la strate U_h sont obtenues par sondage élémentaire exhaustif :

$$V[\bar{X}] = V\left[\sum_{h=1}^{l} \frac{m_h}{m} \bar{X}_h\right] = \sum_{h=1}^{l} \left(\frac{m_h}{m}\right)^2 V[\bar{X}_h] = \sum_{h=1}^{l} \left(\frac{m_h}{m}\right)^2 \frac{m_h - n_h}{m_h - 1} \frac{\sigma_h^2}{n_h}$$

en appelant σ_h l'écart-type des $x_{h\alpha}$ ($\alpha = 1, 2, \ldots, m_h$) autour de $\bar{x}_h$ dans la strate U_h

$$\sigma_h = \sqrt{\sum_{\alpha=1}^{m_h} \frac{(x_{h\alpha} - \bar{x}_h)^2}{m_h}}$$

Problème de l'"optimum allocation" (Neyman 1934) - Nous allons maintenant résoudre le problème envisagé ci-dessus : comment choisir les $f_h = \frac{n_h}{m_h}$ pour rendre $V[\bar{X}]$ minimum, compte-tenu de la condition :

$$\sum_{h=1}^{l} n_h = n, \text{ constante donnée,}$$

qui exprime qu'on se fixe a priori le nombre total d'unités prélevées.

C'est un problème d'extremum lié dont la solution est fournie par la méthode de Lagrange.

On peut résoudre directement ce problème en se plaçant dans le cas pratique où l'on peut confondre m_h et $m_h - 1$ et où $\frac{n_h}{m_h} = f_h$ est petit devant 1 .

On a alors : $$m^2 V[\bar{X}] \# \sum_{h=1}^{l} \frac{m_h^2}{n_h} \sigma_h^2$$

Posons : $y_h = \frac{m_h}{n_h} \sigma_h$ et $\bar{y} = \frac{1}{n} \sum_{h=1}^{l} n_h y_h = \frac{1}{n} \sum_{h=1}^{l} m_h \sigma_h$

Si l'on tient compte de l'identité

$$\sum_{h=1}^{l} n_h(y_h - \bar{y})^2 = \sum n_h y_h^2 - n\bar{y}^2$$

il vient :

$$m^2 V[\bar{X}] = \sum_{h=1}^{l} n_h y_h^2 = \sum_{h=1}^{l} n_h \left[\frac{m_h}{n_h} \sigma_h - \sum_{h=1}^{l} \frac{m_h \sigma_h}{n} \right]^2 + \frac{1}{n} \left(\sum m_h \sigma_h \right)^2$$

Dans le dernier membre, le second terme ne contient plus les n_h et le premier terme (positif ou nul) s'annule, pour

$$\frac{m_h}{n_h} \sigma_h = \sum_{h=1}^{l} \frac{m_h \sigma_h}{n}$$

$V[\bar{X}]$ est donc minimum si

$$n_h = n . \frac{m_h \sigma_h}{\sum_{h=1}^{l} m_h \sigma_h}$$

Le calcul complet par la méthode de Lagrange donnerait :

$$n_h = n . \frac{m_h \sigma_h \sqrt{\frac{m_h}{m_h - 1}}}{\sum_{h=1}^{l} m_h \sigma_h \sqrt{\frac{m_h}{m_h - 1}}}$$

relation dans laquelle on pourra toujours pratiquement confondre m_h et $m_h - 1$.

D'où la règle d'utilisation : on doit adopter dans les strates des fonctions de sondages f_h proportionnelles aux écarts-types σ_h de x à l'intérieur des strates.

L'expression de la variance minimum est

$$V_m(\bar{X}) = \frac{1}{n} \left[\sum_{h=1}^{l} \frac{m_h}{m} \sigma_h \right]^2$$

Remarques -

1) En général, l'échantillon tiré n'est pas soumis à une observation unique X mais à un ensemble X, Y, Z ... d'observations. Il ne pourra donc y avoir une répartition qui soit optimum en soi, mais seulement par rapport à une variable X. Si, comme c'est le cas fréquemment, X, Y, Z sont en corrélation étroite les unes avec les autres, en réalisant l'optimum pour l'une des variables, on sera proche de l'optimum pour les autres.

2) Il n'y a aucune raison pour que la solution du problème de Neyman donne pour les n_h des valeurs entières. Comme on doit nécessairement tirer des strates un nombre entier d'unités, on ne pourra jamais réaliser en toute rigueur l'optimum, même vis-à-vis d'un caractère unique X.

3) Souvent il n'est pas nécessaire de se placer exactement dans le cas optimum car $V[\bar{X}]$ est une fonction des n_h qui varie très lentement autour de son minimum.

4) L'optimum (au sens de Neyman) n'est pas le seul qu'on doive envisager. Si par exemple le coût d'observation d'une unité est variable avec

la strate, il y a lieu de rechercher la répartition optimum des dépenses par strate , la dépense totale étant donnée. On trouve que les fractions sondées doivent être proportionnelles aux écarts-types σ_h par strate et aux inverses des racines carrées des coûts moyens d'observation par strate

$$n_h = n \frac{m_h \frac{\sigma_h}{\sqrt{c_h}}}{\sum_{h=1}^{l} m_h \frac{\sigma_h}{\sqrt{c_h}}}$$

5) Pour déterminer les n_h, il faudrait connaître à l'avance les valeurs de σ_h. En toute rigueur cela est impossible; car pour les connaître, il faudrait connaître la distribution des $x_{h\alpha}$, c'est-à-dire l'univers des x, auquel cas le sondage serait inutile puisque son but est de renseigner sur cet univers.

En réalité, on aura une idée grossière des ordres de grandeur des σ_h ; ou encore, on connaîtra leur valeur lors d'un sondage précédent ou analogue. L'essentiel est de ne pas se tromper grossièrement.

6) Si l'on se trompait grossièrement sur la valeur des σ_h , on se placerait très loin de l'optimum et on s'exposerait à obtenir ainsi une estimation $\bar{X}$ de $\bar{x}$ moins précise que celle obtenue par un sondage non stratifié.

Exemple - Strate 1 : 50 caisses (grandes) - Strate 2 : 1 000 caisses (petites)

$$f_1 = \frac{1}{5} \qquad f_2 = \frac{1}{100}$$

On a adopté des fractions de sondage si différentes parce qu'on a observé une variation considérable entre les poids des grandes caisses, ou entre les nombres de pièces qu'elles doivent contenir, alors que les petites caisses étaient de dimensions très voisines les unes des autres. L'objet du sondage était de s'assurer que la livraison comprenait bien le nombre de pièces (ou le poids) fixé à l'avance.

Dans d'autres circonstances (contrôle de qualité par exemple) on devra porter son attention sur les petites caisses, si on a des raisons de craindre que leur qualité est plus hétérogène que celle des grandes.

L'échantillon à fractions sondées égales

Théorème - L'échantillon stratifié à fractions sondées égales est toujours plus précis que l'échantillon non stratifié (si les strates comportent un nombre suffisamment grand d'unités).

En effet, en assimilant $\frac{m_h - n_h}{m_h - 1}$ à l'unité, ce qui revient à supposer m_h grand et n_h petit devant m_h), on a pour l'échantillon stratifié :

$$V_s [\bar{X}] = \sum_{h=1}^{l} \left(\frac{m_h}{m}\right)^2 \frac{\sigma_h^2}{n_h}$$

et pour l'échantillon ordinaire

$$V_0 [\bar{X}] = \frac{\sigma^2}{n}$$

avec $$\sigma^2 = \frac{1}{m}\sum_{h=1}^{l}\sum_{\alpha=1}^{m_h}(x_{h\alpha} - \bar{x})^2$$

$$= \frac{1}{m}\sum_{h=1}^{l}\left[\sum_{\alpha=1}^{m_h}(x_{h\alpha} - \bar{x}_h)^2 + m_h(\bar{x}_h - \bar{x})^2\right]$$

$$V_0[\bar{X}] = \frac{1}{m^2 f}\sum_{h=1}^{l} m_h\sigma_h^2 + \frac{1}{m^2 f}\sum_{h=1}^{l} m_h(\bar{x}_h - \bar{x})^2$$

soit finalement, en supposant $\frac{n_h}{m_h} = f_h = f$ constant (sondage représentatif)

$$V_s[\bar{X}] = \frac{1}{m^2 f}\sum_{h=1}^{l} m_h\sigma_h^2$$

$$V_0[\bar{X}] = \frac{1}{m^2 f}\sum_{h=1}^{l} m_h\sigma_h^2 + \frac{1}{m^2 f}\sum_{h=1}^{l} m_h(\bar{x}_h - \bar{x})^2$$

En général, le deuxième terme de $V_0[\bar{X}]$ n'est pas nul; il est toujours positif, donc

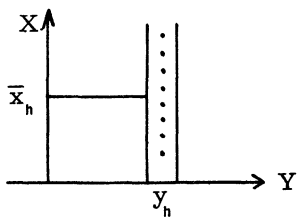
$$V_s[\bar{X}] \leqslant V_0[\bar{X}]$$

L'égalité n'a lieu que si $\bar{x}_1 = \bar{x}_2 = \ldots = \bar{x}_h = \ldots = \bar{x}$. Ainsi, on n'obtient aucun gain de précision à constituer des strates dans une population où les valeurs moyennes de X dans les différentes strates sont égales ou voisines (même si, par ailleurs, les distributions de X dans les strates sont très différentes : très dispersée dans une strate, très peu dans une autre).

Conclusion - En général, il ne faut pas attendre d'une stratification un gain considérable de précision, surtout si X est un caractère qualitatif (variable égale à 0 ou 1). Or, la constitution de certaines strates peut nécessiter un travail préparatoire considérable (si par exemple on doit effectuer certaines mesures sur l'ensemble des unités de sondage pour les classer par strates); alors ce travail ne sera pas "payant" et il y aurait intérêt à accroître l'effectif de l'échantillon.

Qualité du caractère stratificateur Y

Considérons la variable auxiliaire Y qui sert à constituer des strates. Par exemple, quand on crée 2 strates de caisses, les grandes et les petites, Y est leur poids ou leur volume; quand on classe les unités suivant 8 pays de provenance, on a Y = 1, 2, ..., 8, etc.



Portons X et Y sur un graphique rectangulaire.

Les problèmes de stratification sont liés étroitement à la corrélation entre X (caractère étudié) et Y (caractère connu).

Supposons que Y soit une variable continue; partageons son intervalle de variation en segments très petits et très nombreux, c'est-à-dire découpons le plan (X, Y) en bandes très fines, chacune constituant une strate (ceci suppose m et n très grands).

La strate U_h est définie par :

$$y_h - \varepsilon < Y < y_h + \varepsilon$$

Le lieu du point $(\bar{x}_h, y_h)$ n'est autre que la courbe de régression de X en Y. On sait que le rapport des variances est égal à :

$$\frac{V_s [\bar{X}]}{V_0 [\bar{X}]} = 1 - \eta^2$$

η étant le rapport de corrélation de Pearson de X en Y.

On a une réduction substantielle de variance lorsque η est voisin de 1. Il faut donc choisir un caractère Y qui soit en étroite corrélation avec X. C'est ce qu'on ne sait jamais très bien à l'avance et qui explique le peu de succès de certaines stratifications.

Stratification a priori et stratification a posteriori

La théorie qui précède est celle du sondage dans un univers que l'on peut stratifier préalablement.

Il arrive parfois qu'on connaisse les effectifs m_h de chaque strate mais qu'on ne soit pas en mesure d'en tirer un nombre d'unités donné à l'avance. La nature des choses exigera qu'on tire globalement les n unités (comme dans un sondage élémentaire non stratifié); après quoi on saura déterminer à quelle strate appartient une unité tirée donnée.

Soit N_h le nombre d'unités de la strate h ainsi tirées.

On a $\sum_{h=1}^{l} N_h = n$, mais les N_h, à la différence des n_h, sont des aléatoires.

La forme d'estimation ordinaire s'écrit

$$\bar{X} = \frac{1}{n} \sum_{h=1}^{l} \sum_{i=1}^{N_h} x_{hi} = \sum_{h=1}^{l} \frac{N_h}{n} \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}$$

On appelle formule d'estimation stratifiée :

$$\bar{X}' = \sum_{h=1}^{l} \frac{m_h}{m} \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}$$

qui s'applique à la stratification dite "a posteriori".

Il est possible de montrer qu'en première approximation, $\bar{X}'$ a la même précision que $\bar{X}$ (sondage dans un univers stratifié a priori).

Toutefois, il y a lieu de noter qu'on n'est plus maître des fractions de sondage, nécessairement différentes de $f = \frac{n}{m}$, sans quoi on aurait $\frac{m_h}{m} = \frac{N_h}{n}$ et $\bar{X}' = \bar{X}$ mais elles ne peuvent en différer beaucoup. On peut attendre de cette stratification a posteriori un gain modéré. On reprendra plus loin cette question dans l'utilisation de renseignements supplémentaires dans les formules d'estimations.

IV - LE PROCESSUS DU SONDAGE A DEUX DEGRES

Supposons que l'univers à sonder soit constitué de l unités du premier degré repérées par l'indice α, $\alpha = 1, 2, \ldots, l$ (par exemple des caisses), chaque unité primaire renfermant elle-même m_α unités secondaires repérées par l'indice β : $\beta = 1, 2, \ldots, m_\alpha$ (par exemple des pièces, des boîtes ...).

Nous désignerons par $x_{\alpha\beta}$ la valeur du caractère X prise par l'unité secondaire n° β de l'unité primaire n° α et par x_{ij} la valeur du caractère X prise par l'unité secondaire tirée au jème tirage dans l'unité primaire tirée au ième tirage.

x_{ij} est une variable aléatoire - du fait des tirages - alors que $x_{\alpha\beta}$ est une variable certaine. x_{ij} est identique à $x_{\alpha\beta}$ si au ième tirage d'unité primaire on sort celle portant le numéro α et si au jème tirage d'unité secondaire dans l'unité primaire sortie au ième tirage d'unité primaire, on sort celle portant le numéro β.

De même nous désignerons par $x_{\alpha j}$ la valeur du caractère prise par l'unité secondaire sortie au jème tirage dans l'unité primaire non aléatoire qui porte le numéro α et $x_{i\beta}$ sera la valeur du caractère prise par l'unité secondaire numéro β dans l'unité primaire sortie au ième tirage.

Chaque unité primaire forme une sous-population, décrite par sa moyenne $\bar{x}_\alpha$ et sa variance σ_α^2 :

$$\bar{x}_\alpha = \frac{1}{m_\alpha} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta} = \frac{x_{\alpha.}}{m_\alpha}$$

en désignant par $x_{\alpha.}$ la somme

$$x_{\alpha.} = \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$$

$$\sigma_\alpha^2 = \frac{1}{m_\alpha} \sum_{\beta=1}^{m_\alpha} (x_{\alpha\beta} - \bar{x}_\alpha)^2$$

La population totale a pour moyenne par unité primaire

$$\frac{1}{l} \sum_{\alpha=1}^{l} m_\alpha \bar{x}_\alpha = \frac{1}{l} \sum_{\alpha=1}^{l} x_{\alpha.} = \frac{x_{..}}{l}$$

en désignant par $x_{..}$ la somme

$$x_{..} = \sum_{\alpha=1}^{l} x_{\alpha.} = \sum_{\alpha=1}^{l} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$$

(On raisonnera par la suite sur ces grandeurs plutôt que sur les moyennes du fait de l'additivité des sommes).

La variance du caractère X entre les unités secondaires de l'unité primaire α est

$$\sigma_\alpha^2 = \frac{1}{m_\alpha} \sum_{\beta=1}^{m_\alpha} \left(x_{\alpha\beta} - \frac{x_{\alpha.}}{m_\alpha} \right)^2$$

et la variance de $x_{\alpha.}$ est :

$$\sigma^2 = \frac{1}{l} \sum_{\alpha=1}^{l} \left(x_{\alpha.} - \frac{x_{..}}{l} \right)^2$$

Nous allons envisager trois sortes de problèmes dont les deux premiers sont des cas particuliers du troisième :

1/ Tirage de k($< l$) unités primaires sur l et tirage au second degré de toutes les unités secondaires appartenant aux k unités primaires retenues dans l'échantillon (tirage en grappes) (figure 1).

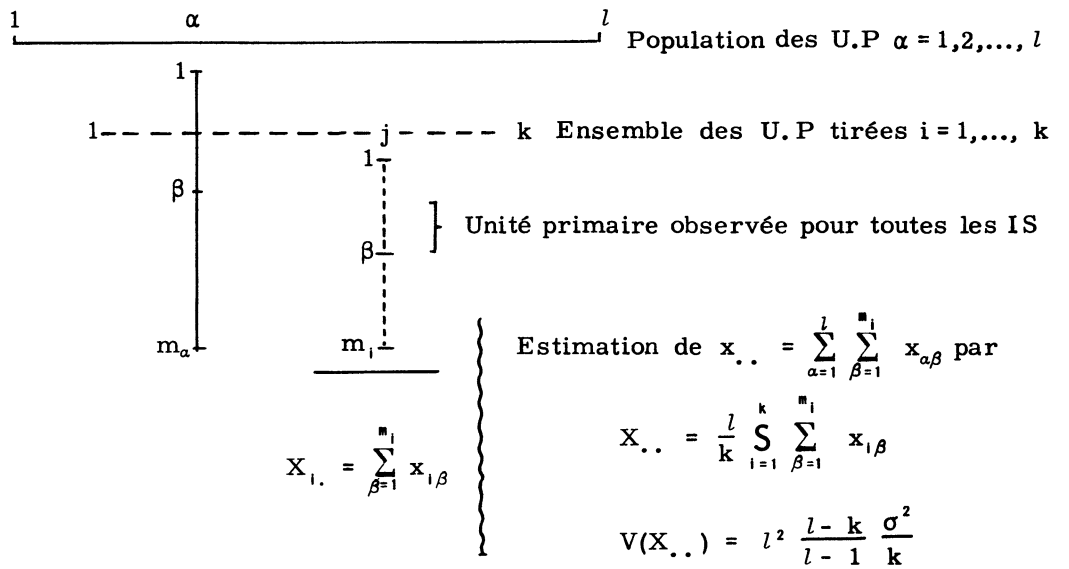


Figure 1

On utilisera alors la notation $x_{i\beta}$

$$i = 1, 2, \ldots, k$$

$$\beta = 1, 2, \ldots, m_i$$

2/ Tirage de la totalité des unités primaires et dans chacune d'elles, tirage de n_α ($< m_\alpha$) unités secondaires sur les m_α qu'elles renferment (tirage stratifié) (figure 2).

On utilisera la notation $x_{\alpha j}$

$$\alpha = 1, 2, \ldots l, \qquad j = 1, 2, \ldots, n_\alpha$$

On conviendra de noter de façon parallèle ce qui concerne population et échantillon. Les sommations sur échantillon seront affectées d'un tildé (~); Ainsi :

Population	Echantillon
$x_{\alpha.} = \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$	$\tilde{x}_i = \sum_{j=1}^{n_i} x_{ij}$
$x_{..} = \sum_{\alpha=1}^{l} x_{\alpha.} = \sum_{\alpha=1}^{l} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$	$\tilde{x}_{..} = \sum_{i=1}^{k} x_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$

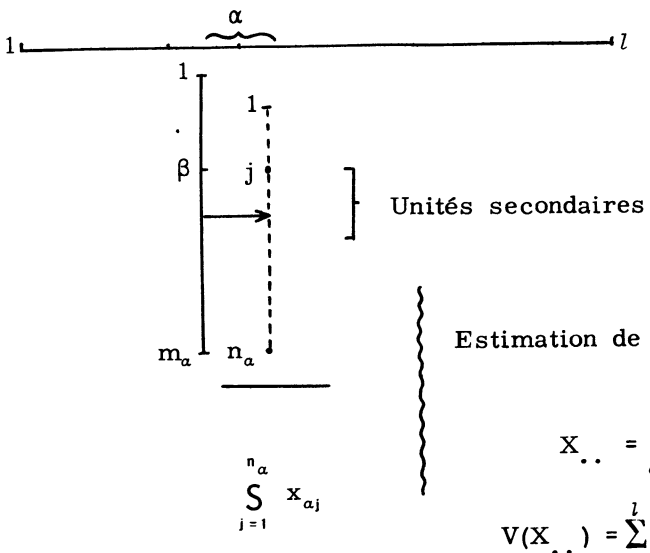


Figure 2

Une notation mixte sera nécessaire :

$$x_{i.} = \sum_{\beta=1}^{m_i} x_{i\beta}$$

On utilisera des grandes lettres comme estimateurs des petites lettres :

$X_{i.}$ est l'estimateur de $x_{i.}$

$X_{..}$ est l'estimateur de $x_{..}$

La notation combinée $\tilde{X}_{..}$ sera alors :

$$\tilde{X}_{..} = \sum_{i=1}^{k} X_{i.}$$

3/ Tirages de k unités primaires sur l, au premier degré, puis dans chaque unité primaire retenue, tirage de n_i unités secondaires sur les m_i qu'elle contient.

On utilisera alors la notation x_{ij}

$$i = 1, 2, \ldots, k \qquad j = 1, 2, \ldots, n_i$$

i et j sont alors des aléatoires pouvant, selon le type de double tirage utilisé, prendre des valeurs appartenant respectivement aux groupes

$$1 \ldots \alpha \ldots l \qquad \text{et} \qquad 1 \ldots \beta \ldots m_\alpha$$

Premier problème : tirage par grappes (figure 1) - On a alors :

$$\tilde{x}_{i.} = X_{i.} = x_{i.} = \sum_{\beta=1}^{m_i} x_{i\beta}$$

et l'estimateur de $x_{..}$ est

$$X_{..} = \frac{l}{k}\,\tilde{X}_{..} = \frac{l}{k}\,\tilde{x}_{..} = \frac{l}{k}\sum_{i=1}^{k}\sum_{\beta=1}^{m_i} x_{i\beta}$$

et

$$V(X_{..}) = l^2\,\frac{l-k}{l-1}\,\frac{\sigma^2}{k}$$

Second problème : tirage stratifié (figure 2) - On a alors :

$$X_{\alpha.} = \frac{m_\alpha}{n_\alpha}\,\tilde{x}_{\alpha.} = \frac{m_\alpha}{n_\alpha}\sum_{j=1}^{n_\alpha} x_{\alpha j}$$

et

$$X_{..} = \tilde{X}_{..} = \sum_{\alpha=1}^{l}\frac{m_\alpha}{n_\alpha}\sum_{j=1}^{n_\alpha} x_{\alpha j}$$

$$V(X_{..}) = \sum_{\alpha=1}^{l} m_\alpha^2\,\frac{m_\alpha - n_\alpha}{m_\alpha - 1}\,\frac{\sigma_\alpha^2}{n_\alpha}$$

Troisième problème : cas général (figure 3) - On a alors :

$$X_{i.} = \frac{m_i}{n_i}\,\tilde{x}_{i.} = \frac{m_i}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

$$X_{..} = \frac{l}{k}\,\tilde{X}_{..} = \frac{l}{k}\sum_{i=1}^{k}\frac{m_i}{n_i}\,\tilde{x}_{i.} = \frac{l}{k}\sum_{i=1}^{k}\frac{m_i}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

$$V(X_{..}) = \sum_{\alpha=1}^{l} m_\alpha^2\,\frac{m_\alpha - n_\alpha}{m_\alpha - 1}\,\frac{\sigma_\alpha^2}{n_\alpha} + l^2\,\frac{l-k}{l-1}\,\frac{\sigma^2}{k}$$

Ainsi la variance apparaît dans le cas général comme la somme des variances écrites plus haut : le premier terme fait intervenir la variance entre unités primaires, le second terme, la variance entres les unités secondaires.

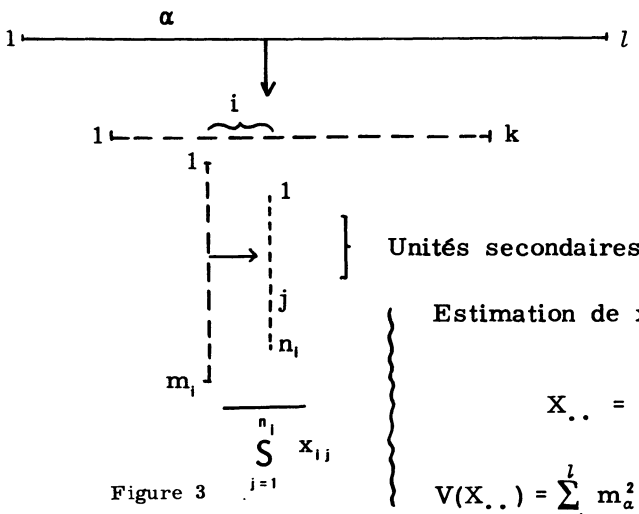


Figure 3

Estimation de $x_{..} = \sum_{\alpha=1}^{l}\sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$ par

$$X_{..} = \frac{l}{k}\sum_{i=1}^{k}\frac{m_i}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

$$V(X_{..}) = \sum_{\alpha=1}^{l} m_\alpha^2\,\frac{m_\alpha - n_\alpha}{m_\alpha - 1}\,\frac{\sigma_\alpha^2}{n_\alpha} + l^2\,\frac{l-k}{l-1}\,\frac{\sigma^2}{k}$$

On a obtenu $X_{..}$ estimation de $x_{..}$. Pour passer de $x_{..}$ à la moyenne (x) par unités primaires ou à celle ($\bar{x}$) par unités secondaires, il suffit de multiplier par une constante :

$$(x) = \frac{x_{..}}{l} \qquad (\bar{x}) = \frac{x_{..}}{m}$$

en posant $m = \sum_{\alpha=1}^{l} m_\alpha$.

Les estimateurs (X) et ($\bar{X}$) de (x) et ($\bar{x}$) se déduisent simplement de $X_{..}$

$$(X) = \frac{X_{..}}{l} = \frac{1}{k} \sum_{i=1}^{k} \frac{m_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$E[(X)] = (x) = \frac{\sum_{\alpha=1}^{l} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}}{l}$$

$$(\bar{X}) = \frac{X_{..}}{m} = \frac{1}{m} \frac{l}{k} \sum_{i=1}^{k} \frac{m_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$$E[(\bar{X})] = (\bar{x}) = \frac{\sum_{\alpha=1}^{l} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}}{m}$$

$$V[(X)] = \frac{1}{l^2} V[X_{..}] = \sum_{\alpha=1}^{l} \left(\frac{m_\alpha}{l}\right)^2 \frac{m_\alpha - n_\alpha}{m_\alpha - 1} \frac{\sigma_\alpha^2}{n_\alpha} + \frac{l-k}{l-1} \frac{\sigma^2}{k}$$

$$V[(\bar{X})] = \frac{1}{m^2} V[X_{..}] = \sum_{\alpha=1}^{l} \left(\frac{m_\alpha}{m}\right)^2 \frac{m_\alpha - n_\alpha}{m_\alpha - 1} \frac{\sigma_\alpha^2}{n_\alpha} + \frac{l^2}{m^2} \frac{l-k}{k-1} \frac{\sigma^2}{k}$$

Cas particulier - $m_\alpha = \bar{m}$ constant $= \frac{m}{k}$. Alors :

$$\bar{X} = \frac{1}{k\bar{m}} \sum_{i=1}^{k} X_{i.} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

et

$$V[\bar{X}] = \sum_{\alpha=1}^{l} \frac{1}{l^2} \frac{\sigma_\alpha^2}{n_\alpha} \frac{\bar{m} - n_\alpha}{\bar{m} - 1} + \frac{1}{\bar{m}^2} \frac{\sigma^2}{k} \frac{l-k}{l-1}$$

$$\# \sum_{\alpha=1}^{l} \frac{1}{l^2} \frac{\sigma_\alpha^2}{n_\alpha} \left(1 - \frac{n_\alpha}{\bar{m}}\right) + \frac{1}{\bar{m}^2} \frac{\sigma^2}{k} \left(1 - \frac{k}{l}\right)$$

Remarque - Très souvent, la seconde partie de la variance sera très grande parce que,

$$\sum_{\beta=1}^{m_\alpha} x_{\alpha\beta} = x_{\alpha.}$$

variant beaucoup d'une unité primaire à l'autre, σ^2 sera grand.

Alors le sondage à 2 degrés sera peu précis et on sera conduit à l'améliorer :

- en stratifiant les unités primaires;
- en tirant les unités primaires avec des probabilités proportionnelles à leur taille m_α et non avec équiprobabilité.

On se bornera sur ces deux points à quelques indications :

1/ Stratification et sondage à 2 degrés

On procède d'abord au classement des unités primaires en différentes strates; à l'intérieur de chacune d'elles, on tire un échantillon a deux degrés (il est en outre possible d'intercaler une seconde stratification, celle des unités secondaires avant le second tirage). Les formules pour chaque strate sont celles données plus haut; on en combine les estimations pour obtenir l'estimation de la moyenne.

Echantillon représentatif - Pour avoir un échantillon qui nécessite le minimum de difficultés de calcul, on a intérêt à tirer une fraction donnée $f' = \frac{k}{l}$ d'unités primaires et une fraction donnée $f'' = \frac{n_i}{m_i}$ d'unités secondaires (parmi les unités primaires tirées) soit au total une fraction $f = f' f''$ des unités secondaires. S'il s'agit d'un univers stratifié, on aura de même intérêt à prendre une même fraction f des unités secondaires par strate. En effet dans ces conditions, les estimations correctes sont obtenues sans pondération et le sondage se dépouille comme un recensement.

En effet :
$$(X) = \sum_{i=1}^{k} \frac{1}{k} \frac{m_i}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

s'écrit dans ce cas, où :
$$k = lf'$$
$$n_i = m_i f''$$
$$(X) = \frac{1}{lf'f''} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

soit
$$(X) = \frac{1}{lf} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

et
$$(\bar{X}) = \frac{1}{mf} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

2/ Probabilités de sortie des unités primaires proportionnelles à leur taille, les unités secondaires étant tirées avec équiprobabilité

Notons qu'un tel tirage n'est réalisable que dans la mesure où on connaît le nombre des unités secondaires contenues dans chaque unité primaire.

Si l'unité primaire α est affectée de la probabilité $\frac{m_\alpha}{m}$ $\left(m = \sum_{\alpha=1}^{l} m_\alpha\right)$ d'être tirée, les formules d'estimation et de variance sont à modifier ainsi :

$$(\bar{X}') = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

est une estimation de

$$(\bar{x}) = \frac{1}{\sum_{\alpha=1}^{l} m_\alpha} \sum_{\alpha=1}^{l} \sum_{\beta=1}^{m_\alpha} x_{\alpha\beta}$$

et
$$V[(\bar{X}')] = \frac{1}{k} \sum_{\alpha=1}^{l} \frac{m_\alpha}{m} \frac{\sigma_\alpha^2}{n_\alpha} \frac{m_\alpha - n_\alpha}{m_\alpha - 1} + \frac{r}{k} \sum_{\alpha=1}^{l} \frac{m_\alpha}{m} (\bar{x}_\alpha - \bar{x})^2$$

où r est un facteur légèrement inférieur à 1.

Application de cette méthode - Si l'univers est formé comme suit :

Unité primaire n°α =	1	2	3	4	5	...	l
Nombre d'unités secondaires de l'unité primaire n°α : m_α =	15	25	47	8	16	...	
$\sum_{u=1}^{\alpha} m_u$: m_α cumulés	015	040	087	095	111	...	625

On forme les totaux cumulés des m_α et on prend dans une table de nombres aléatoires autant de nombres inférieurs à 625 qu'il faut tirer d'unités primaires au total : si l'un d'eux est 089, c'est l'unité n°4 qui est tirée puisque 087 < 089 < 095.

Remarque - On n'examinera pas ici les difficultés provenant de la possibilité de tirer, de cette façon, plusieurs fois la même unité primaire.

Echantillon "représentatif" - Lorsque les unités sont ainsi tirées, l'échantillon "représentatif" n'a pas du tout les caractères indiqués plus haut. Chaque unité secondaire a autant de chances d'être tirée, quelle que soit l'unité primaire à laquelle elle appartient; de sorte que, après avoir tiré k unités primaires sur l, il y a lieu de tirer de chacune le même nombre $n_i = \bar{n}$ d'unités secondaires, si l'on ne veut pas avoir à introduire des coefficients de pondération. On a alors $k\bar{n} = n$

et $$(\bar{X}'_R) = \frac{1}{k\bar{n}} \sum_{i=1}^{k} \sum_{j=1}^{\bar{n}} x_{ij} \qquad \text{(estimation de } \bar{x})$$

Si l'on considère des unités secondaires stratifiées, et si l'on opère un sondage dans chaque strate, il y a lieu de prélever dans chaque strate un nombre d'unités secondaires n_h proportionnel à m_h avec $k_h\bar{n}_h = n_h$

et $$(\bar{X}'_R) = \sum_{h=1}^{l} \frac{1}{k_h\bar{n}_h} \sum_{i=1}^{k} \sum_{j=1}^{\bar{n}_h} x_{hij}$$

En résumé pour avoir un sondage représentatif qui se dépouille comme un recensement, il faut :

soit tirer les unités primaires avec égales probabilités et une fraction donnée f" d'unités secondaires par unité primaire tirée au premier degré;

soit tirer les unités primaires avec des probabilités proportionnelles à leur taille et un nombre $\bar{n}$ donné à l'avance d'unités secondaires par unité primaire tirée au premier degré.

V - LE PROCESSUS DE SONDAGE A PLUSIEURS PHASES

a) Il ne faut pas confondre sondage à plusieurs degrés et sondage à plusieurs phases.

La différence entre sondages à plusieurs degrés et sondages à plusieurs phases a été fixée en 1948, par la Sous-Commission des Sondages des Nations-Unies, qui a normalisé le vocabulaire et les méthodes. On désignait autrefois le sondage à deux phases sous le nom de sous-échantillonnage (sub-sampling) qui peut s'appliquer aussi bien au sondage à deux degrés).

Dans le sondage à plusieurs degrés, on change d'unités de sondage à chaque degré, et les unités de sondage successives s'emboîtent les unes dans les autres. Dans le sondage à plusieurs phases, au contraire, on conserve les mêmes unités de sondage d'un bout à l'autre, mais on en modifie les fractions de sondage d'une phase à l'autre.

b) Voici une méthode de sondage à 2 phases très utile. On a vu que, pour stratifier une population, il était nécessaire d'avoir sur chaque unité de celle-ci un renseignement (en général qualitatif) et (pour que cette stratification soit efficace) un renseignement en corrélation avec la variable X étudiée. Quand on ne possède pas de tels renseignements a priori, mais qu'on peut les recueillir à peu de frais, il n'est pas indispensable de le faire pour toutes les unités de la population.

En effet, on peut (1ère phase) tirer d'abord un grand échantillon de cette population. On recueille sur ces unités les renseignements qui permettent de les classer en strates. Enfin, on tire (2ème phase) un petit échantillon de chaque strate avec des fractions de sondage variables d'une strate à l'autre, en faisant jouer (pour choisir les fractions) des considérations analogues à celles données pour la répartition optimum dans le cas de l'échantillon stratifié ordinaire. C'est alors qu'on procède aux observations ou mesures définitives sur les petits échantillons ainsi obtenus.

Par exemple - On dispose d'une collection de pièces de mécanique dont certaines sont rouillées et dont on veut étudier l'état.

1ère phase - On tire au sort un millier de pièces. On les trie en 3 groupes : pièces intactes, pièces présentant quelques tâches de rouille, pièces très rouillées. On mesure ainsi quelle proportion de ces pièces entre dans chacun des 3 groupes.

2ème phase - On conserve 1/50 des pièces du premier groupe, 1/5 des pièces du 2ème groupe et toutes celles du 3ème groupe. On examine alors au microscope les pièces ainsi conservées, on les soumet à des observations très minutieuses et on détermine l'importance des dégâts faits par la rouille

L'intérêt du sondage à 2 phases réside dans le fait que l'opération qui suit la 1ère phase est beaucoup plus rapide et moins coûteuse que celle que suit la 2ème phase.

Les calculs à faire

On va supposer qu'on avait μ unités au départ, dont on a tiré m unités à la 1ère phase. On obtient ainsi : $M_1, M_2, \ldots, M_h, \ldots, M_l$ unités dans les strates 1, 2, ..., h, ..., l. On écrit M_h et non m_h car les M_h sont des aléatoires dont la somme M est certaine et égale à m).

On prélève alors $n_1, n_2, \ldots, n_h, \ldots, n_l$ unités sur lesquelles on mesure le caractère x_{hi} ce qui fournit les moyennes $\overline{X}_h$

$$\overline{X}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$$

estimation de

$$\overline{x}_h = \frac{1}{\mu_h} \sum_{\alpha=1}^{\mu_h} x_{h\alpha}$$

μ_h étant le nombre d'unités de la population appartenant à la strate h.

$$\left(\sum_{h=1}^{l} \mu_h = \mu \qquad \text{et} \qquad E[M_h] = \frac{m}{\mu}\,\mu_h\right)$$

On estime enfin $\bar{x} = \frac{1}{\mu}\sum_{h=1}^{l}\sum_{\alpha=1}^{\mu_h} x_{h\alpha} = \frac{1}{\mu}\sum_{h=1}^{l} p_h \bar{x}_h$

par $\bar{X} = \frac{1}{m}\sum_{h=1}^{l}\frac{M_h}{n_h}\sum_{i=1}^{n_h} x_{hi}$

Cette estimation est correcte car $E[\bar{X}] = \bar{x}$,

Le calcul de $V[\bar{X}]$ diffère de celui de la variance d'un échantillonnage stratifié car les M_h sont des aléatoires. Cependant, en première approximation, l'expression suivante est valable :

$$V[\bar{X}] \# \sum_{h=1}^{l}\left(\frac{M_h}{m}\right)^2 \frac{\sigma_h^2}{n_h}\,\frac{M_h - n_h}{M_h - 1}$$

Problème de Neyman (1938) - Neyman a traité en 1938 le problème analogue à celui qu'il avait résolu en 1934; dans le cas actuel où M_h est une aléatoire, comment répartir un échantillon donné entre les strates pour réaliser l'optimum.

On peut se donner l'effectif total $n = \sum_{h=1}^{l} n_h$ de l'échantillon définitif retenu.

On peut aussi se donner les coûts d'enquête au premier degré (c_1) et au deuxième degré (c_2) par unité, ainsi que la dépense totale C.

Il faut réaliser $V[\bar{X}]$ minimum compte tenu de la condition de liaison $C = m c_1 + n c_2$.

Dans le cas plus simple, où l'on cherche à rendre $V[\bar{X}]$ minimum pour $n = \sum_{h=1}^{l} n_h$ donné, on trouve un résultat assez voisin de celui obtenu plus haut. L'optimum s'obtient en prenant les n_h proportionnels à :

$$\sigma_h \sqrt{p_h^2 + \frac{p_h q_h}{m}}$$

où σ_h est l'écart-type de x à l'intérieur de la strate h et où p_h est la proportion des unités de la population qui se trouve dans cette strate :

$$\sigma_h^2 = \frac{1}{\mu_h}\sum_{\alpha=1}^{\mu_h}(x_{h\alpha} - \bar{x}_h)^2$$

$$p_h = \frac{\mu_h}{\mu} \qquad q_h = 1 - p_h$$

Pratiquement q_h est voisin de 1, $\frac{p_h q_h}{m}$ est voisin de 0 et $\sigma_h\sqrt{p_h^2 + \frac{p_h q_h}{m}}$ est voisin de $\sigma_h p_h$ c'est-à-dire $\sigma_h \frac{\mu_h}{\mu}$ voisin lui-même de $\sigma_h \frac{M_h}{m}$. C'est-à-dire que n_h devrait être à peu près proportionnel à $\sigma_h M_h$ comme dans le cas du sondage dans un univers stratifié.

Caractère stratificateur - Nous avons laissé jusqu'ici dans l'ombre le fait que la stratification entre les deux phases est faite en classant les unités de

l'échantillon suivant les valeurs d'un caractère Y (caractère stratificateur). Si ce caractère est en corrélation étroite avec X, le sondage à deux phases améliore l'échantillon par rapport au sondage ordinaire (pour un échantillon d'effectif donné).

L'opération qui consiste à observer ou mesurer Y sur l'échantillon doit être une opération beaucoup plus facile (et moins coûteuse) que celle relative à X; sans quoi on ne ferait qu'ajouter à la dépense ($n c_2$) relative au caractère X une dépense ($m c_1$) du même ordre, sinon plus importante.

Pratiquement, on aura intérêt à procéder par sondage à 2 phases si, par exemple, les relevés relatifs à Y consistent simplement à rechercher dans un document la valeur de Y pour les unités de grand échantillon (exemple : cas d'une cargaison composée de sacs, dont le poids déclaré Y figure sur un état; les sacs portent des numéros de référence; Y est le poids réel, qu'on ne mesurera que sur un petit échantillon).

Quant à la répartition optimum, il est clair qu'en pratique on s'inspirera plus souvent du résultat qualitatif (prélever une fraction d'autant plus élevée que la dispersion est plus grande dans la strate) que du résultat quantitatif.

Echantillon équilibré ("balanced") - Une forme spéciale d'échantillon à 2 phases est l'échantillon équilibré vis-à-vis d'un caractère Y ou d'un ensemble de caractères.

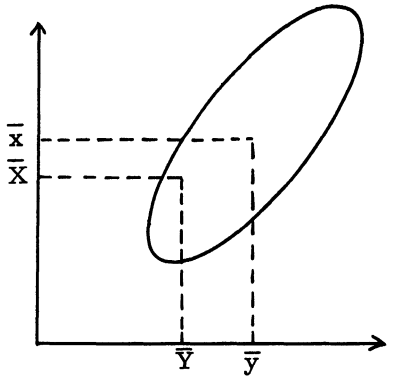
On a vu que l'échantillon stratifié (par rapport à Y) et représentatif avait la même composition en proportions que la population. Tirer un échantillon équilibré par rapport à Y signifie au contraire s'arranger pour que l'échantillon ait la même valeur moyenne $\bar{Y}$ que la population ($\bar{y}$) :

$$\bar{Y} = \bar{y}$$

Lorsque le caractère Y est qualitatif (égal à 0 ou 1) les deux points de vue coïncident mais ce n'est pas le cas général.

Théorie - La théorie mathématique de l'échantillon équilibré est très simple.

Considérons dans le plan XY, le point aléatoire ($\bar{X}$, $\bar{Y}$). Lorsque l'échantillon d'effectif n est assez grand on démontre que le point ($\bar{X}$, $\bar{Y}$) suit une loi (limite) de Laplace-Gauss, de moyenne ($\bar{x}$, $\bar{y}$); s'il y a une forte corrélation entre X et Y, l'ellipse indicatrice a une forme allongée (voir la figure). Imposer $\bar{Y} = \bar{y}$ signifie que le point aléatoire est lié par la condition $\bar{Y} = \bar{y}$.



La variance de $\bar{X}$ non lié est

$$\frac{\sigma_x^2}{n}\,\frac{m-n}{m-1}$$

et celle de $\bar{X}$ lié est :

$$\frac{\sigma_x^2}{n}\,\frac{m-n}{m-1}(1-\rho^2)$$

où ρ est le coefficient de corrélation entre X et Y (voir page).

Ainsi la précision de $\bar{X}$ augmente considérablement avec la corrélation entre X et Y.

Si $\rho = 0,95$, on a $1 - \rho^2 \# \frac{1}{10}$; la précision est donc 10 fois plus grande.

Ceci suppose qu'on connaisse un caractère Y lié au caractère X par une corrélation de 0,95 (ce qui n'est pas fréquent).

Réalisation pratique - En pratique, il n'est pas très commode de réaliser un échantillon équilibré tout en respectant la loi du tirage au sort. Voici la méthode de M. Yates.

On tire n unités au hasard. On recueille ou mesure les Y sur ces unités, soit :

$$y_1 \; y_2 \; \cdots \; y_n$$

On calcule la moyenne $\bar{Y}$; elle diffère de $\bar{y}$.

On rejette l'unité n°1 (donc le nombre y_1) et on tire une nouvelle unité, soit la $(n \pm 1)^e$ unité. On recueille la donnée y_{n+1} et on calcule la nouvelle valeur de $\bar{Y}$; elle diffère encore de $\bar{y}$.

On rejette l'unité n°2 et on tire une nouvelle unité, soit la $(n + 2)^e$, etc.

On finira bien par se rapprocher de $\bar{y}$; et on s'arrêtera (à la m^e unité) lorsque la valeur $(\bar{Y} - \bar{y})$ sera du même ordre que les erreurs de mesure .

Alors, sur les n unités conservées, on procèdera à la partie des opérations relatives au caractère X.

On aura donc recueilli Y sur un échantillon de m unités ($m > n$) et X sur un échantillon de n unités. Il s'agit en somme d'une sorte de sondage à 2 phases.

Généralisation - On peut équilibrer l'échantillon vis-à-vis d'un ensemble de caractères Y, Z ... Mais la réalisation pratique du système de conditions $\bar{Y} = \bar{y}$, $\bar{Z} = z$, ... est difficile.

VI - FORMULES D'ESTIMATION FAISANT INTERVENIR LES RENSEIGNEMENTS SUPPLEMENTAIRES

On a déjà vu à plusieurs reprises ce qu'on entendait par renseignements supplémentaires sur la population sondée.

Pour pouvoir stratifier une population, on a vu qu'il fallait avoir un renseignement supplémentaire sur chaque unité. Pour tirer un échantillon équilibré par rapport à y, il fallait connaître $\bar{y}$.

On se propose de montrer ici de quelle manière il est possible, à l'aide de tels renseignements, d'améliorer les estimations établies à partir d'un échantillon; l'amélioration sera apportée non plus au moment où l'on tire l'échantillon, mais après coup, au moment du dépouillement des données, avant de faire les calculs.

On a déjà indiqué l'une de ces méthodes : celle de la stratification a posteriori.

Supposons que l'échantillon ait été tiré au hasard par le procédé ordinaire (bien que l'idée soit de portée bien plus générale).

On recueille les observations X sur l'échantillon de taille n; l'estimation élémentaire de $\bar{x}$ est

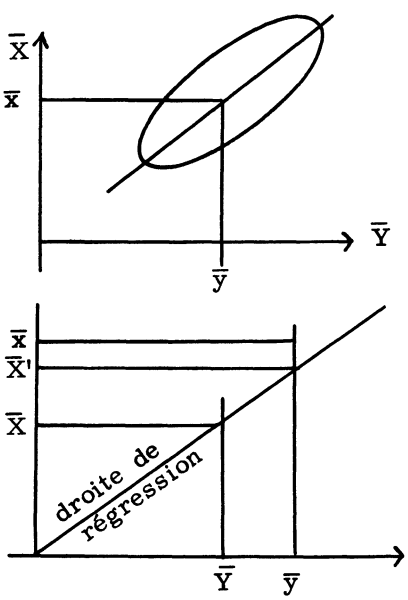
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

avec $$V[\bar{X}] = \frac{\sigma^2}{n} \frac{m - n}{m - 1}$$

Lorsque les unités sont très diverses vis-à-vis du caractère X, σ va être très grand; pour n = 100 par exemple, si σ est grand, $\frac{\sigma}{10}$ sera encore trop grand pour nous.

Supposons qu'on connaisse, pour chaque unité, un caractère Y en corrélation assez étroite avec X. La connaissance de la moyenne $\bar{y}$ de l'univers, de celle $\bar{Y}$ de l'échantillon, permet d'appliquer la formule d'estimation par le quotient pour estimer $\bar{x}$.

On pourrait employer une autre formule de régression (linéaire par exemple) de X et y. Mais nous nous bornerons au cas de l'estimation par le quotient ("ratio-estimation").



L'échantillon n'étant pas équilibré par rapport à Y, on a par exemple $\bar{Y}$ inférieur de 3% à $\bar{y}$. On a

$$\frac{\bar{Y}}{\bar{y}} = \frac{97}{100}$$

On va admettre que $\bar{X}$ est aussi trop petit de 3% et l'estimation $\bar{X}'$ de $\bar{x}$ sera $\frac{100}{97}\bar{X}$, c'est-à-dire $\bar{X}' = \frac{\bar{y}}{\bar{Y}}\bar{X}$.

Pour calculer la variance de $\bar{X}' = \bar{y}\frac{\bar{X}}{\bar{Y}}$ il faut d'abord étudier la variance du quotient de deux variables aléatoires.

Quotient de deux variables aléatoires : formules approchées du 2ème ordre.

Soient A et B deux variables aléatoires d'espérances mathématiques a et b. $\frac{A}{B}$ ne diffère de $\frac{a}{b}$ qu'au premier ordre près.

Posons $A = a + \alpha$, $B = b + \beta$

$$\frac{A}{B} - \frac{a}{b} = \frac{a + \alpha}{b + \beta} - \frac{a}{b} = \frac{ab + \alpha b - ab - a\beta}{b(b + \beta)} = \frac{\alpha b - a\beta}{b^2\left(1 + \frac{\beta}{b}\right)}$$

soit

$$\frac{A}{B} - \frac{a}{b} = \frac{a}{b}\left(\frac{\alpha}{a} - \frac{\beta}{b}\right)\left(1 + \frac{\beta}{b}\right)^{-1}$$

$$\simeq \frac{a}{b}\left(\frac{\alpha}{a} - \frac{\beta}{b}\right)\left(1 - \frac{\beta}{b}\right)$$

$$\simeq \frac{a}{b}\left(\frac{\alpha}{a} - \frac{\beta}{b} + \frac{\beta^2}{b^2} - \frac{\alpha\beta}{ab}\right)$$

Soient γ les coefficients de variation de A, γ' celui de B et ρ le coefficient de corrélation entre A et B.

$$\gamma = \frac{\sqrt{V[A]}}{E[A]} \qquad \gamma' = \frac{\sqrt{V[B]}}{E[B]} \qquad \rho = \frac{Cov(A, B)}{\sqrt{V[A] \cdot V[B]}}$$

On a alors :

$$E\left[\frac{A}{B}\right] = \frac{a}{b}\left(\gamma'^2 - \rho\gamma\gamma'\right)$$

$$\left(\frac{A}{B} - \frac{a}{b}\right)^2 \# \left[\frac{a}{b}\left(\frac{\alpha}{a} - \frac{\beta}{b}\right)\right]^2$$

$$V\left[\frac{A}{B}\right] \# \left(\frac{a}{b}\right)^2\left[\frac{E\alpha^2}{a^2} + \frac{E\beta^2}{b^2} - \frac{2}{ab}\,Cov(\alpha, \beta)\right]$$

soit
$$V\left[\frac{A}{B}\right] = \left(\frac{a}{b}\right)^2\left[\gamma^2 - 2\rho\gamma\gamma' + \gamma'^2\right]$$

Dans le cas de l'estimation par le quotient :

on a
$$E[\bar{X}] = \bar{x}$$

$$E[\bar{Y}] = \bar{y}$$

$$Cov[\bar{X}, \bar{Y}] = \rho\sigma_x\sigma_y$$

$$\gamma_{\bar{x}} = \frac{\sigma_{\bar{x}}}{\bar{x}} = \frac{\sigma_x}{\bar{x}\sqrt{n}}\frac{m - n}{m - 1} = \frac{1}{\sqrt{n}}\gamma_x\frac{m - n}{m - 1}$$

de même
$$\gamma' = \frac{\sigma_{\bar{y}}}{\bar{y}} = \frac{1}{\sqrt{n}}\gamma_y\frac{m - n}{m - 1}$$

$$a)\ E[\bar{X}'] = E\left[\bar{y}\frac{\bar{X}}{\bar{Y}}\right] = \bar{y}\,E\left[\frac{\bar{X}}{\bar{Y}}\right] \# \bar{y} \cdot \frac{\bar{x}}{\bar{y}}\left(\gamma^2 - \rho\gamma\gamma'\right)$$

soit
$$E[\bar{X}'] \# \bar{x} + \frac{\bar{x}}{n}\left(\gamma_y^2 - \rho\gamma_x\gamma_y\right)\frac{m - n}{m - 1},$$

le biais est donc peu différent de

$$B = E[\bar{X}'] - \bar{x} \# \frac{\bar{x}}{n}\left(\gamma_y^2 - \rho\gamma_x\gamma_y\right)\frac{m - n}{m - 1}$$

Ce biais est petit lorsque n est grand.

$$b)\quad V[\bar{X}'] = \bar{y}^2 V\left[\frac{\bar{X}}{\bar{Y}}\right] \# \bar{y}^2\left(\frac{\bar{x}}{\bar{y}}\right)^2\left(\gamma^2 - 2\rho\gamma\gamma' + \gamma'^2\right)$$

soit $$V[\bar{X}'] \# \frac{\bar{x}^2}{n}(\gamma_x^2 - 2\rho\gamma_x\gamma_y + \gamma_y^2)\frac{m-n}{m-1}$$

l'erreur-type de $\bar{X}'$ est donc proportionnelle à $\frac{1}{\sqrt{n}}$, donc de l'ordre de $\sqrt{n}$ par rapport au biais. Si n = 100 on pourra négliger le biais et la variance $V[\bar{X}']$ sera une bonne mesure de la précision de $\bar{X}'$.

Comme on peut écrire

$$V[\bar{X}] = \frac{\bar{x}^2}{n}\gamma_x^2\frac{m-n}{m-1}$$

on a : $$\frac{V[\bar{X}']}{V[\bar{X}]} = 1 - 2\rho\frac{\gamma_y}{\gamma_x} + \left(\frac{\gamma_y}{\gamma_x}\right)^2$$

On a donc $$V[\bar{X}'] < V[\bar{X}]$$

si $$-2\rho\frac{\gamma_y}{\gamma_x} + \left(\frac{\gamma_y}{\gamma_x}\right)^2 < 0$$

c'est-à-dire si

$$\boxed{\rho > \frac{1}{2}\frac{\gamma_y}{\gamma_x}}$$

En particulier si γ_y et γ_x sont du même ordre de grandeur, il suffira que $\rho > \frac{1}{2}$.

Remarquons que :

$$\frac{V[\bar{X}']}{V[X]} = \left(1 - \frac{\gamma_y}{\gamma_x}\right)^2 + 2\frac{\gamma_y}{\gamma_x}(1-\rho)$$

expression qui tend vers zéro si $\frac{\gamma_y}{\gamma_x}$ tend vers 1 et si $\rho = 1$.

On arrive ainsi, si la corrélation est importante, à une estimation $\bar{X}'$ nettement meilleure que $\bar{X}$.

Application au problème du renouvellement de l'échantillon - Supposons que l'on veuille mesurer par sondages faits à deux époques différentes la variation qui se produit dans l'intervalle. On a intérêt à procéder, si cela est possible, les deux fois sur le même échantillon.

Par exemple, même échantillon de machines-outils, même échantillon d'opérateurs.

Soient alors x_i et y_i les deux observations faites sur la même unité de sondage (i); il y a souvent une corrélation importante entre X et Y et par conséquent le quotient $\frac{\bar{X}}{\bar{Y}}$ peut être estimé avec bien plus grande précision que pour $\bar{X}$ et $\bar{Y}$ estimés sur des échantillons différents(1).

---

(1) Remarquons que $V[\bar{X} - \bar{Y}] = V[\bar{X}] + V[\bar{Y}] - 2\,\mathrm{Cov}[\bar{X}, \bar{Y}]$. La différence $\bar{X} - \bar{Y}$ est elle aussi évaluée avec grande précision.

Toutefois il arrive qu'on ne puisse conserver l'échantillon intégralement, notamment parce qu'une partie de l'enquête disparaît (décès) entre les 2 époques, tandis qu'une autre apparaît (naissances), et aussi (lorsqu'on a affaire à des observations répétées à 3, 4 ... époques successives) parce que l'échantillon observé se déforme.

On est donc souvent amené à renouveler l'échantillon; pour ne pas perdre le bénéfice de la corrélation étroite, on ne le renouvelle que par fractions; par exemple par $\frac{1}{6}$ chaque fois (en permutation circulaire) pour arriver en 6 fois à la disparition totale de l'échantillon primitif.

Supposons qu'on ait affaire, pour simplifier, à un "ratio" tel que le suivant :

$$R = \frac{\bar{X}}{\bar{Y}} = \frac{X_3 + X_2}{Y_2 + Y_1}$$

A l'époque n°2, (3) est la partie de l'échantillon substituée à (1); (2) est la partie commune.

Epoque 1 : (1) + (2); effectifs $n_1 + n_2 = n$

Epoque 2 : (3) + (2); effectifs $n_3 + n_2 = n$

X représente les sommes portant sur x pour tous les éléments d'une partie de l'échantillon

$$V[X_3 + X_2] = n\sigma_x^2$$

$$V[Y_2 + Y_1] = n\sigma_y^2$$

On peut supposer en première approximation $\gamma_x^2 = \gamma_y^2$. D'autre part: $\mathrm{Cov}[X_3 + X_2,\ Y_2 + Y_1] = \mathrm{Cov}(X_2\ Y_2) = n_2\,\rho\,\sigma_x\sigma_y$ en supposant l'indépendance entre (3) et (2), (3) et (1), et (2) et (1). On arrive ainsi à

$$\gamma^2[R] \# \frac{2}{n}\gamma^2\left(1 - \frac{n_2}{n}\rho\right)$$

tandis qu'avec un échantillon entièrement nouveau, on aurait : $\frac{2\gamma^2}{n}$.

Si l'on introduit la fraction renouvelée $\alpha = 1 - \frac{n_2}{n}$, l'efficacité du renouvellement partiel par rapport au renouvellement complet est égale à

$$\frac{1}{1 - (1 - \alpha)\rho}$$

Dans le cas d'un sondage stratifié à plusieurs degrés, les formules sont beaucoup plus compliquées.

Conclusion - On vient de passer rapidement en revue diverses méthodes d'échantillonnage. Il en existe d'autres. Par exemple, par la méthode des réseaux de sondage superposés (qui rappelle les méthodes des "blocks") et l'analyse de variance, il est possible de mettre en évidence, moyennant un accroissement des frais d'opération, l'influence des opérateurs enquêteurs, recenseurs, etc. et la part qu'ils ont dans la variance totale.

ANNEXE

EMPLOI INDUSTRIEL DES MÉTHODES DE SONDAGES AUX ÉTATS-UNIS D'APRÈS W. E. DEMING

Voici quelques exemples pratiques, empruntés à un article du Dr. Deming[1].

A - RECEPTION DU SUCRE RAFFINE IMPORTE AUX U.S.A.

Le sucre arrive par lots importants comprenant souvent jusqu'à 43 000 sacs (modèle standard), pesant chacun environ 100 livres. Les Douanes établissent par sondage le tonnage réel de l'ensemble.

Pour cela, on commence par séparer les sacs endommagés (dont le poids peut être notablement plus petit que 100 livres). La strate des sacs endommagés sera sondée avec une fraction de sondage assez grande et celle des sacs non endommagés avec une fraction de sondage très faible.

Pour les sacs non endommagés, l'unité de sondage adoptée est le "prélèvement" de 8 sacs (contigus); le poids de cette grappe présente un coefficient de variation de l'ordre de 0,1% seulement. Avec un échantillon de 9 prélèvements (prélevés au hasard, indépendamment les uns des autres), l'estimation du poids de la cargaison est connu avec :

3 écarts-types = 0,1%

et il est extrêmement rare que l'écart réel dépasse 3 écarts-types. On a ainsi une bonne précision. En pratique et sans beaucoup de frais, on va jusqu'à 25 prélèvements (de 8 sacs). L'intérêt est qu'on peut alors évaluer avec précision le coefficient de variation, sur l'échantillon (au lieu de l'admettre égal à 0,1%).

B - RECEPTION DE LA LAINE

La laine brute renferme des graisses et autres impuretés dont la proportion peut varier de 30% (entre balles d'un même lot et aussi entre touffes de la même balle). On a besoin d'une méthode permettant de déterminer la proportion de laine lavée que fournira la laine brute. Pour cela le lavage a lieu sur l'échantillon.

La laine est sondée à 2 degrés : 1er degré : échantillon de balles; 2ème degré : échantillon de touffes ("cores") de laine prélevées dans les balles au moyen de sondes spéciales enfoncées dans les balles à des profondeurs variables.

Le problème est de savoir combien l'échantillon doit comprendre de touffes par balle et combien de balles doivent être sondées, - pour atteindre une précision donnée avec une dépense minimum.

---

(1) On the sampling of Physical Materials. Revue de l'Institut International de Statistique - 1950 - N°1/2 - pp. 1 à 20.

Pour arriver à ce résultat, il suffit de savoir écrire la variance d'un échantillon à 2 dégrés de m balles sur M et $\bar{n}$ touffes par balles : $V(\bar{n}, m)$ et la dépense correspondante : $C(n, m) = a\,m + b\,\bar{n}\,m$.

On écrit alors

$$\frac{\partial V}{\partial \bar{n}} + \lambda \frac{\partial C}{\partial \bar{n}} = 0$$

$$\frac{\partial V}{\partial m} + \lambda \frac{\partial C}{\partial m} = 0$$

$$V(\bar{n}, m) = 0,5/100 \text{ (par exemple)}$$

Remarque -

1/ Influence des coûts élémentaires a et b. Pour sonder des laines stockées, le coût (a) du 1er degré du sondage est très élevé vis-à-vis du coût du 2ème degré.

On a alors intérêt à prélever peu de balles et beaucoup de touffes par balle. Pratiquement, l'usage est de prendre $\bar{n}$ égal à 10 touffes par balle.

Au contraire, pour des laines arrivant par camions dans une usine, le coût (a) du 1er degré est minime (car absorbé par le coût normal de manutention des balles); et l'usage est de prendre $\bar{n}$ égal à 1 touffe par balle.

2/ Influence de l'origine des laines. La variance à l'intérieur des balles est 2 fois plus grande pour les laines d'Australie et du Chili que pour les laines d'Argentine. Pour les laines indigènes (des Etats-Unis) c'est la variance entre les balles qui est 2 fois plus grande. La valeur de ces 2 variances intervenant dans l'expression $V(\bar{n}, m)$ il faudrait en tenir compte.

Le tableau suivant (correspondant à la laine d'Argentine et à une erreur-type de 0,5%) donne le nombre de balles à prélever m en fonction de la taille du lot à réceptionner et du nombre de touffes à prélever par balle.

Nombre de balles M du lot	Nombre de touffes à prélever par balle $\bar{n}$			
	1	2	4	6
25	25	19	16	15
50	34	25	21	20
100	40	30	25	24
200	45	34	28	26
500	48	36	30	28
1 000	49	37	31	29

Ainsi, il suffit de sonder 2 fois plus de balles quand le lot comprend 1 000 balles au lieu de 25 (d'ailleurs, pour un lot de 25 balles, presque toutes doivent être sondées).

C - RECEPTION DU TABAC D'IMPORTATION

Contrairement au sucre (exemple A) le tabac arrive par balles de poids excessivement variable; le coefficient de variation pour un même lot est de l'ordre de 3,5% (variant entre 1,5 et 9,4%); et la taille que devrait avoir l'échantillon (si l'on appliquait la même méthode que pour le sucre) serait excessive.

On procède autrement. On utilise l'information supplémentaire constituée par le poids déclaré du tabac; et on étudie le ratio :

$$\frac{\text{poids réel à la réception}}{\text{poids déclaré}}$$

Il existe, en effet, une étroite corrélation entre poids déclaré et poids réel. Le coefficient de variation de ce ratio est seulement de 0,7% (sauf pour le tabac grec où il atteint 2,5%). On obtient donc une précision de 0,25% (satisfaisante) avec un échantillon de 100 balles (sauf pour le tabac grec où il faudrait aller jusqu'à 900 balles).

D - EVALUATION D'UN CAPITAL D'UNE COMPAGNIE DES TELEPHONES

a) Une compagnie des Téléphones possède 973 000 poteaux (avec tout leur équipement); on désire savoir dans quel état se trouve ce matériel, dont on possède l'inventaire. Pour cela on tire (sur l'inventaire) un échantillon (par exemple, de 1 500 poteaux) qu'on fait examiner par un technicien (le poteau existe-t-il encore ? Si oui, dans quelle mesure est-il détérioré ? etc.). Le technicien met une note à chaque poteau échantillon, et la moyenne des notes est applicable aux 973 000 poteaux. Ceci suppose d'ailleurs qu'on a pu définir avec assez de précision la façon de noter et que les divers opérateurs sont, à ce sujet, à peu près d'accord entre eux et avec eux-mêmes.

b) Si le capital de la Compagnie comprend d'autres biens (câbles, pylones, etc.) le problème qui se posera est celui d'obtenir la "répartition optimum" des crédits disponibles entre les divers types de matériel à examiner par sondage.

E - DETERMINER LE POUVOIR CALORIFIQUE OU LA TENEUR EN CENDRES D'UNE CATEGORIE DE CHARBON

Le problème d'échantillonnage qui se pose (avant le problème d'analyse chimique) est délicat. Lorsque le charbon peut être saisi au cours de sa manipulation, on peut tirer au sort rigoureusement un échantillon (disons) de wagonnets. Mais cet échantillon est très gros; il n'est pas question d'analyser plus de quelques centaines de grammes de charbon.

On procède alors à un broyage qui réduit le charbon à l'état de grains moyens, dont on prélève une fraction de la forme $(1/2)^n$ par n partages en 2 moitiés (à la pelle). On recommence alors le broyage pour parvenir à des grains plus fins, puis très fins, etc. On réduit chaque fois l'échantillon.

On opère en somme un sondage à plusieurs degrés; à chaque degré du sondage on fait éclater les unités d'ordre (k) en unités plus petites d'ordre (k + 1). L'effet de chaque broyage est contrôlé par un passage au tamis.

L'unité finale de sondage est un grain de poussière de charbon. C'est sous cet état qu'on livre l'échantillon au chimiste.

Il est essentiel de se rendre compte qu'aucun calcul d'erreur d'échantillonnage ne sera possible tant qu'on n'aura pas analysé les variances correspondant à chaque degré du sondage (on ne l'a jamais fait, croyons-nous).

En outre, le problème se pose souvent d'échantillonner des tas (ou des wagons) de charbon; auquel cas l'échantillon est prélevé avec des sondes qui ne fournissent qu'assez approximativement un échantillon assimilable à celui que donnerait le tirage au sort.

En résumé, on n'a aucune donnée sur la précision de ce genre d'échantillonnage.

oooo

Remarque - Une description détaillée des méthodes réelles d'échantillonnage du charbon dans les divers pays se trouve dans le document (de 166 pages) W/ Coal/ CWP/1 tome III de la commission Economique pour l'Europe (Nations-Unies).