JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

P. THIONET

Un problème de sondage parmi les éléments dont la distribution est très dissymétrique

Journal de la société statistique de Paris, tome 96 (1955), p. 192-206

<http://www.numdam.org/item?id=JSFS_1955__96__192_0>

V

VARIÉTÉ

Un problème de sondage parmi les éléments dont la distribution est très dissymétrique

Au cours d'une mission à la Direction Générale des Impôts, nous avons eu à nous occuper de la question suivante :

On considère un *univers*, appelé *groupe d'activité*, constitué d'éléments appelés *redevables*. On classe ceux-ci en 2 *strates* :

— *la strate supérieure* dont la totalité est incluse dans l'enquête,

— *la strate inférieure* où une certaine fraction f des redevables est tirée au sort (avec d'égales probabilités) pour être soumise à l'enquête, les données obtenues étant ensuite pondérées par $1/f$, avant d'être ajoutées à celles de la strate supérieure.

On désire obtenir ainsi des estimations ayant une précision donnée, ou plus exactement ayant *un coefficient de variation donné à l'avance.*

Comment doit-on choisir : 1° la définition de la strate « inférieure » et de la strate « supérieure »,
2° la fraction f,
de façon à soumettre à l'enquête le *nombre minimum de redevables?*

Il s'agit d'un problème manifestement très général. Toutefois il devait être résolu ici, non seulement en théorie, mais encore par *des moyens suffisamment pratiques pour que l'on* puisse obtenir les *résultats numériques* relatifs à plusieurs centaines de groupes d'activité (sans disposer d'un bureau de calcul).

***

Ce problème a conduit à étudier trois questions distinctes :

— La stratification optimum.

— Les courbes de concentration.

— L'agrégation des lois de distribution.

Ces questions vont être évoquées successivement ci-après.

CHAPITRE I

UN PROBLÈME DE STRATIFICATION OPTIMUM

1) Introduction

1.1. — Les théories classiques sur la stratification laissent de côté la recherche du découpage *optimum* en strates. Il existe divers travaux, de nature expérimentale (1), mettant en lumière les réductions de variance qu'on peut

(1) Voir par exemple Lilian H. Madow. *Journal of the Am. Stat. Assoc.*, 1950, pp. 30-47.

obtenir en accroissant *le nombre de strates*, toutes choses égales d'ailleurs et dans les limites du nombre d'unités de sondage (du 1er degré) qu'on désire tirer (car il faut tirer au moins une unité de chaque strate); le gain de variance peut être médiocre si l'on utilise déjà des formules d'estimation qui tiennent compte des mêmes informations supplémentaires servant à découper les strates.

1.2. — Notre cours de sondage (1) était sous presse lorsque nous découvrions l'existence de travaux du statisticien suédois *Tore Dalenius* concernant le problème de stratification optimum. Voici la liste desdits mémoires :

Skandinavisk Aktuarietidskrift (Journal des actuaires scandinaves).

1950, p. 203-213. — *The Problem of Optimum Stratification I.*

1951, p. 133-148. — *The Problem of Optimum Stratification II* (en collaboration avec Miss Margaret Gurney).

1952, p. 61-70. — *The Problem of Optimum Stratification in a special type of design.*

En 1954, un problème voisin de ceux traités par Dalenius nous était posé, et nous avons pu constater le bon accord de nos formules et des siennes.

Peu après, un problème un peu différent était résolu par M. J Desabie à l'I. N. S. E. E. (2).

1.3. — Les recherches du Dr Dalenius portent sur l'influence que peut avoir la variation des frontières arbitraires choisies pour définir les strates, le nombre de strates étant fixé d'avance.

Tout d'abord, il suppose qu'on a réparti l'échantillon entre les strates, proportionnellement à leur effectif (sondage « représentatif »); puis il se préoccupe d'un échantillon à répartition « optimum » (donnant un estimateur de variance minimum), soit que le coût unitaire de l'enquête soit le même dans toutes les strates, soit qu'il y ait lieu d'adopter un coût unitaire différent pour chaque strate.

Dans son troisième mémoire le Dr Dalenius suppose qu'il n'existe plus que *2 strates* dont une est retenue à 100 % dans l'échantillon. C'est exactement notre hypothèse. Mais le Dr Dalenius suppose fixé l'effectif total de l'échantillon et il en déduit quelle stratification conduit à la variance minimum. Nous nous sommes fixé au contraire la précision à atteindre (c'est-à-dire la valeur de la variance) et avons voulu en déduire quelle stratification donnait l'échantillon le plus petit.

1.4. — Le problème de M. Desabie est analogue à ceux du Dr Dalenius, en ce sens qu'il s'agit encore de réduire la variance à son minimum. Il en diffère en ceci que son univers est découpé par une double statification, l'une (fixe) en groupes d'activité, l'autre (mobile) en strates supérieure et inférieure, la limite entre ces 2 strates étant différente d'un groupe d'activité à l'autre.

M. Desabie s'intéresse à l'estimation d'ensemble (tous groupes d'activité réunis), ce en quoi son problème diffère du nôtre (où sont envisagées les estimations relatives à *chaque* groupe).

---

(1) Études théoriques nos 5 et 6, de l'INSEE. Imprimerie Nationale, 1953.

(2) Voir le Bulletin d'Information de l'I.N.S.E.E. (1955), ainsi que les Travaux de l'Association pour le Développement des Techniques d'Étude des Marchés (ADETEM), 1954-1955.

1.5. — Une différence secondaire entre ces divers travaux réside dans la nature des unités et des variables étudiées :

— effectifs de main-d'œuvre des établissements (Desabie).
— revenus des particuliers (Dalenius).
— chiffre d'affaires des établissements (Dalenius, nous-même).

2. Les hypothèses de base

Il a paru nécessaire d'insister sur certaines hypothèses de base, si l'on veut que le présent exposé reste intelligible.

Hypothèses A. — On limite l'étude à une variable unique x, (pratiquement le chiffre d'affaires total C. A.) jugée beaucoup plus importante que les autres (montant des diverses taxes, C. A. à l'exportation, etc...); pour ces dernières les estimations supporteront des erreurs d'échantillonnage qui peuvent être notablement plus grandes que celle du C. A. total.

On va étudier une estimation de $\Sigma\, x$ ou (x), C. A. total de *chaque groupe* d'activité (disons en 1954) et chercher à rendre son coefficient de variation égal par exemple à 1 %, ou 2 % (plus généralement à C_0).

Ceci n'est possible que parce qu'on dispose des données complètes concernant une situation passée (par exemple celle de 1952) et qu'il existe une étroite parenté entre la situation présente (1954) et cette situation passée (1952).

Hypothèses B. — Entre la période de base (1952) et l'époque présente, il est clair que l'*univers* s'est en partie renouvelé; certains redevables ont disparu, quelques nouveaux sont apparus, l'effectif de l'univers s'est plus ou moins modifié; des phénomènes de concentration ont pu apparaître; etc.... On n'en tiendra aucun compte; on supposera que tous les redevables existant en 1954 existaient déjà en 1952 (la contre-partie n'étant pas absolument indispensable). A tout C. A. x_i de 1954 correspond donc un C. A. y_i de 1952.

Hypothèses C. — On simplifiera encore le problème en admettant que, dans la strate inférieure, le coefficient de variation de la variable x (inconnue) a une *valeur numérique pratiquement égale* à celle de la variable y (connue).

Hypothèses D. — On achèvera de simplifier le problème en supposant que les éléments ou unités de sondage sont *rangés sur un axe dans un ordre donné* et que les 2 strates seront déterminées *par un point frontière* qu'il s'agit de placer convenablement sur cet axe.

Pratiquement on classera les « redevables » suivant un critère y, qui sera ici le *chiffre d'affaires* 1952 de ce redevable;

La strate supérieure est par définition formée des plus « gros » redevables (et la strate inférieure des plus « petits » redevables) définis par la condition : y supérieur à la valeur frontière y_0.

Il est alors facile de mettre le problème en équations.

3. La mise en équations

Autres Notations : Strate supérieure : indice 1; strate inférieure : indice 2.
Effectifs de la population : m; de l'échantillon : n.
Fraction sondée (de la strate inférieure) : f. On pose $m_2/m = p$. *Strate infé-*

rieure : Écart type des x_i : σ; coefficient de variation : c. Opérateurs : Variance : V; coefficient de variation : C. V.

Effectif de l'échantillon (2 strates) : $n = m_1 + f\, m_2$
$= m\,[1 - p\,(1 - f)]$

Estimateur employé : $X = (x_1) + m_2 X_2$

Calcul du coefficient de variation de cet estimateur :

On confond m_2 et $(m_2 - 1)$ pour simplifier. Il vient sans difficulté, d'accord avec le Dr Dalenius :

$$C.\ V^2(X) = m\,p\,\frac{\sigma^2}{(x)^2}\,\frac{1-f}{f}$$

4. Résolution théorique de notre problème

Tout revient à *rendre maximum* $p\,(1 - f)$ compte tenu de la condition :

$$\boxed{m\,p\,\frac{\sigma^2}{(x)^2}\,\frac{1-f}{f} = C_0^2} \qquad (1)$$

(par exemple $C_0 = 0{,}01$ ou $0{,}02$),
autrement dit :

$$1 - f = \frac{1}{1 + \dfrac{m\,p\,\sigma^2}{C_0^2 \cdot (x)^2}}; \qquad (1')$$

ce qui revient à *rendre maximum* l'expression

$$z = p \,/\, (1 + h\,p\,\sigma^2)$$
$$\text{avec } h = m/Co^2\,(x)^2. \qquad (2)$$

Ce maximum s'obtient pour

$$dz/dp = o \qquad (3)$$

d'où l'équation de l'optimum :

$$\boxed{\frac{d\,\sigma^2}{d\,p} = \frac{1}{hp^2}} \qquad (3')$$

Conclusion : La résolution pratique du problème suppose donc connue l'expression (au moins empirique) de σ^2 en fonction de p (on ne rencontre pas cette difficulté quand on renverse les conditions de problème comme le font le Dr Dalenius ou M. Desabie)...

Mais on n'a réellement besoin que de la dérivée $d\sigma^2/dp$; la valeur absolue de σ^2 n'intervient plus (sinon pour calculer ensu te la valeur de f). C'est pourquoi il sera raisonnable de substituer à la distribution des C. A. une distribution théorique qui coïncide bien avec l'autre *dans la zone où peut se faire la coupure entre les 2 strates et là seulement* le cas échéant. Cette substitution ne peut conduire à une stratification très différente de l'optimum; elle peut seulement

perturber éventuellement les valeurs de la fraction sondée f correspondante : valeurs qui dans la pratique seront nécessairement perturbées pour d'autres causes : adoption pour $1/f$ d'entiers *simples*. Telle est la justification du paragraphe 5 ci-après.

Remarque : Lorsqu'on suppose la fraction de sondage f suffisamment petite, on peut confondre $(1 - f)$ avec 1 ; ce qui conduit à remplacer la condition (2) par la suivante :
Rendre maximum l'expression :

$$z_0 = p\,(1 - h\,p\,\sigma^2); \qquad (2')$$

d'où l'équation :

$$p^2 \frac{d\,(\sigma^2)}{d\,p} + 2\,p\,\sigma^2 = \frac{1}{h}. \qquad (3'')$$

Comme (3'') est plus compliqué que (3'), on a renoncé à cette « simplification ».

5. Mode de résolution pratique effectivement utilisé

On a été amené à s'intéresser à la famille de distributions à un paramètre ε, dont la courbe de concentration (1) est représentée par l'équation suivante :

$$pq + \varepsilon\,p - (1 + \varepsilon)\,q = o \qquad (\varepsilon > o)$$

Tout son intérêt est de se prêter à un ajustement rapide sur des données (2) sommaires concernant la distribution. Un calcul (un peu long) donne l'expression suivante de σ^2 en fonction soit de p, soit de q :

$$\sigma^2 = \left[\bar{x}^2 \frac{\varepsilon^2}{3\,(1 + \varepsilon)}\right] \frac{p^2}{(1 + \varepsilon - p)^3} = \left[\frac{\bar{x}^2}{3\,\varepsilon\,(1 + \varepsilon^2)}\right] q^2\,(q + \varepsilon) \qquad (4)$$

L'équation (3') s'explicite comme suit :

soit en p $$\frac{1}{m\,Co^2} \frac{\varepsilon^2}{3\,(1 + \varepsilon)}\, p^3\,(2 + 2\,\varepsilon + p) = (1 + \varepsilon - p)^4 / \qquad (5)$$

soit en q $$3\,q^4 + 2\,\varepsilon\,q^3 = [3\,m\,Co^2\,\varepsilon^2\,(1 + \varepsilon)] \qquad (5')$$

On a retenu cette dernière équation (5') dont la résolution est très simple : On *a mis en table* les valeurs du premier membre de (5') pour $q = 1\ \%$, 2 %, etc... ; 30 % et pour $\varepsilon = 0{,}005$; 0,01 ; 0,015 ; 0,02 ; 0,03 ; ... 0,17 ; 0,18 ; 0,21 ; 0,24 ; 0,27 ; 0,30. (Tableau 1). (On a calculé d'ailleurs certaines valeurs pour q plus grand que 30 %).

D'autre part, on a calculé la valeur du 2e membre, pour $C_a = 1\ \%$ et $C_a = 2\ \%$, pour chacun des groupes d'activité considérés (ce qui suppose qu'on connaisse l'effectif m et le paramètre ε du groupe). Entrant alors dans la table 1 on en a déduit la valeur q correspondante, valeur qui détermine la limite à adopter entre les 2 strates (stratification optimum).

(1) Voir le chapitre 2 ci-après, pour la définition de cette courbe et les calculs correspondants.

(2) Données 1952 (c'est-à-dire yi et non ri) ; on postule que m et ε n'ont pas varié de 1952 à 1954 ; seul x aurait varié. Il suffit de connaître un couple de valeurs (p, q) au voisinage de la coupure pour évaluer ε ; pratiquement on demandait à la mécanographie de fournir (p, q) pour $q = 5\ \%, 10\ \%, 15\ \%$ (et éventuellement 20 % et 30 % pour de rares professions). En général les différentes valeurs de ε ainsi obtenues ne différaient guère entre elles et ne laissaient guère de doute sur la distribution théorique à retenir.

Le paramètre ε caractérise la « *concentration* » du groupe professionnel auquel il se rapporte ; il est d'autant plus petit que le groupe est davantage concentré.

TABLEAU 1. — *Valeurs de* $m C_0^2$ *(où* C_0 *est exprimé en* %)

q % \ ε %	0.5	1.0	1.5	2	3	4	5	6	7	8	9	10
1	—	—	—	—	—	—	—	—	—	—	—	—
2	74	21		—	—	—	—	—	—	—	—	—
3	358	98	47	29	(14)	—	—	—	—	—	—	—
4	1.102	296	139	83	41	26	(18)	—	—	—	—	—
5	2.655	701	326	194	94	58	40	30	23	(19)	—	—
6	5.440	1.426	657	388	186	113	77	57	44	35	29	25
7	10.010	2.605	1.193	703	333	119	135	99	76	61	51	43
8	17.000	4.395	2.000	1.147	552	328	221	161	124	99	81	68
9	29.000	6 975	3.170	1.845	865	510	343	248	190	151	124	104
10	41.000	10.560	4.780	2.780	1.295	761	508	367	280	222	181	152
11	60.000	15 375	6 940	4.025	1.870	1.090	726	523	398	314	256	214
12	84.800	21.670	9.760	5.650	2.610	1.520	1.010	725	550	433	352	293
13	116.400	29.725	13.370	7.710	3.560	2.065	1.370	980	740	582	474	393
14	156.400	39.840	17.900	10 300	4.740	2.750	1.810	1.290	977	767	621	515
15	206.000	52.350	23.500	13.500	6.190	3.580	2.355	1.680	1.265	992	804	665
16	....	67.580	30.250	17 400	7.950	4.590	3.020	2.140	1.615	1.262	1.020	844
17	....	85.900	38.400	22.050	10 050	5 810	3.800	2.710	2.030	1.585	1.280	1.060
18	425.000	107.800	48 200	27.600	12 600	7.240	4.750	3.360	2.520	1.965	1.585	1.305
19	....	133.500	59.600	34 200	15.200	8.930	5.840	4.140	3.100	2.415	1.940	1.600
20	....	163.700	73.000	41 750	19 000	10.900	7 100	5.040	3.770	2.930	2.360	1.925
21	786.000	198.600	88 500	50.600	23 000	13.160	8.570	6.060	4.540	3 520	2.830	2.330
22	....	239.000	106.500	60 900	27 600	15 800	10.260	7.250	5.420	4.210	3.380	2.770
23	....	285.000	127.000	72 500	32 800	18.750	12 200	8.610	6.430	4.980	4.000	3.280
24	....	337.500	150.000	85.700	38 800	22.150	14.775	10.150	7.560	5.860	4.700	3.850
25	....	396.000	....	101.000	45.500	26.000	16.875	11.900	8.850	6.750	5.490	4.500
26	....	464.000	....	117.800	53.100	30.250	19.650	13.800	10.300	7.950	6.380	5.210
27	....	539.000	....	136.600	61.500	35.100	22.750	16.000	12.900	9.200	7.350	6.025
28	....	623.000	....	157.700	71.000	40.500	26.200	18.400	13.690	10 600	8.450	6.910
29	....	716.000	....	181.000	81.500	46.400	30.000	21.100	15.670	12.100	9.660	7.900
30	....	820.000	....	207.000	93.000	53.000	34.250	24.000	17.900	13.800	11.000	9.000

q % \ ε %	11	12	13	14	15	16	17	18	21	24	27	30
1	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—	—	—
4	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	—	—	—	—	—	—	—
6	21	(19)	—	—	—	—	—	—	—	—	—	—
7	37	32	28	25	22	20	(19)	—	—	—	—	—
8	58	51	45	40	36	32	29	27	21	—	—	—
9	88	77	67	60	54	48	44	40	31	25	21	(18)
10	129	112	98	87	77	70	63	59	47	39	33	28
11	182	157	137	121	108	97	88	80	62	50	42	35
12	249	214	182	165	147	132	119	108	84	68	56	47
13	332	286	249	219	195	175	158	144	111	89	74	62
14	436	374	320	286	254	228	206	187	144	115	95	80
15	560	481	413	367	325	292	263	239	183	146	120	101
16	711	610	529	464	411	368	331	300	230	183	151	126
17	890	760	660	565	512	458	412	373	285	227	186	155
18	1.100	940	814	714	630	564	506	458	349	277	227	189
19	1.342	1.150	993	870	769	685	616	556	416	336	274	228
20	1.630	1.390	1.200	1.050	928	825	740	670	508	403	328	274
21	1.950	1 665	1.440	1.300	1.110	990	885	800	606	480	390	325
22	2.325	1.980	1.710	1.490	1.315	1.170	1.050	948	716	565	460	382
23	2.745	2.340	2.015	1.760	1.550	1.380	1.240	1.115	845	664	539	447
24	3.225	2.745	2.360	2.020	1.815	1.615	1.445	1.305	980	774	627	520
25	3.760	3.200	2.755	2.400	2.110	1.880	1.680	1.515	1.140	895	726	600
26	4.360	3.700	3.190	2.780	2.440	2.170	1.940	1.750	1.310	1.030	835	690
27	5.030	4.270	3.680	3.200	2.810	2.500	2.235	2.010	1.510	1.185	957	790
28	5.760	4.900	4.210	3.660	3.220	2.860	2.555	2.300	1.725	1.350	1.090	900
29	6.600	5.600	4.810	4.180	3.680	3.260	2.910	2.620	1.960	1.535	1.240	1.020
30	7.450	6.350	5.230	4.750	4.170	3.700	3.300	2.970	2.220	1.740	1.400	1.150

On avait encore besoin de connaître quelle valeur f donner à la fraction de sondage; le calcul donne sans grande difficulté :

$$1/f = 3 + \frac{3\,q}{\varepsilon}$$

Pour plus de commodité, on a établi une Table des valeurs de $1/f$ en fonction

de q et de ε, sur le même modèle que celle des valeurs du 1er membre de (5') (Tableau 2).

A titre d'indication, on a trouvé (pour la ville de Paris avec la nomenclature I.N.S.E.E. des groupes d'activité) :

ε = 1 %, pour les groupes : 297, 31, 37, 623, 70, 732, 79,
ε = 2 %, pour les groupes : 20, 24, 293, 30, 340, 35, 44, 45, 470, ...
ε = 5 %, pour les groupes : 333, 337, 481, 492, 495, 692, 731, 737,...
ε = 15 %, pour les groupes : 338, 526.
ε = 30 %, pour les groupes : 892.

TABLEAU 2. — *Valeurs de* $1/t$

ε % / q %	0.5	1.0	1.5	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	21	24	27	30
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	15	9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3	21	12	9	7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
4	27	15	11	9	7	6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	33	18	13	10	8	7	6	5	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6	39	21	15	12	9	7	6	6	6	5	5	5	5	(4)	—	—	—	—	—	—	—	—	—	—
7	45	24	17	13	10	8	7	6	6	6	5	5	5	5	5	5	4	4	(4)	—	—	—	—	—
8	51	27	19	15	11	9	8	7	6	6	6	5	5	5	5	5	5	4	4	4	4	—	—	—
9	57	30	21	16	12	10	8	7	7	6	6	6	5	5	5	5	5	5	5	4	4	4	4	(4)
10	63	33	23	18	13	10	9	8	7	7	6	6	6	5	5	5	5	5	5	5	4	4	4	4
11	69	36	25	19	14	11	9	8	8	7	7	6	6	6	5	5	5	5	5	5	5	4	4	4
12	75	39	27	21	15	12	10	9	8	7	7	6	6	6	6	6	5	5	5	5	5	4	4	4
13	81	42	29	22	16	13	11	9	8	8	7	7	7	6	6	6	6	5	5	5	5	5	4	4
14	87	45	31	24	17	13	11	10	9	8	8	7	7	6	6	6	6	6	5	5	5	5	5	4
15	93	48	33	25	18	14	12	10	9	9	8	7	7	7	6	6	6	6	6	5	5	5	5	4
16	99	51	35	27	19	15	12	11	10	9	8	8	7	7	7	6	6	6	6	6	5	5	5	5
17	105	54	37	28	20	16	13	11	10	9	9	8	8	7	7	7	6	6	6	6	5	5	5	5
18	111	57	39	30	21	16	14	12	11	10	9	8	8	7	7	7	7	6	6	6	6	5	5	5
19	117	60	41	31	22	17	14	12	11	10	9	9	8	8	7	7	7	7	6	6	6	5	5	5
20	123	63	43	33	23	18	15	13	11	10	10	9	8	8	8	7	7	7	6	6	6	5	5	5
21	129	66	45	34	24	19	15	13	12	11	10	9	9	8	8	7	7	7	7	6	6	6	5	5
22	135	69	47	36	25	19	16	14	12	11	10	9	9	8	8	8	7	7	7	6	6	6	5	5
23	141	72	49	37	26	20	17	14	13	12	11	10	9	9	8	8	8	7	7	7	6	6	6	5
24	147	75	51	39	27	21	17	15	13	12	11	10	10	9	8	8	8	8	7	7	6	6	6	5
25	153	78	53	40	28	22	18	15	14	12	11	10	10	9	9	8	8	8	7	7	7	6	6	5
26	159	81	55	42	29	22	18	16	14	13	12	11	10	9	9	9	8	8	7	7	7	6	6	6
27	165	84	57	43	30	23	19	16	14	13	12	11	10	10	9	9	8	8	8	7	7	6	6	6
28	171	87	59	45	31	24	20	17	15	13	12	11	11	10	9	9	9	8	8	8	7	6	6	6
29	177	90	61	46	32	25	20	17	15	14	13	12	11	10	10	9	9	8	8	8	7	7	6	6
30	183	93	63	48	33	25	21	18	16	14	13	12	11	10	10	9	9	9	8	8	7	7	6	6

6. AUTRES RÉSOLUTIONS PRATIQUES A L'AIDE D'UNE EXPRESSION EMPIRIQUE DE σ (p)

6.1 — Avant de procéder comme au paragraphe 5, on avait calculé (pour quelques groupes d'activité) les valeurs numériques de σ correspondant à diverses stratifications. On avait constaté que log σ variait linéairement avec p dans l'intervalle de variations de p qui nous intéressait.

En posant donc :

$$\sigma^2 = e^{\lambda p + \mu}$$

on était ramené à résoudre graphiquement l'équation

$$\log \sigma^2 = -\log (h\,\lambda) - 2 \log p \qquad (3' \text{ explicitée})$$

Pour ces quelques groupes d'ailleurs, l'accord est excellent entre les résultats numériques des méthodes des paragraphes 5 et 6.

Mais il faut observer que, pour $p = o$, la strate inférieure disparaît; et la limite de σ^2 est nulle, de sorte que la formule empirique (qui donne e^{μ}) ne convient plus du tout.

En outre la méthode suppose qu'on calcule au moins 2 valeurs numériques de σ^2 pour ajuster λ et μ sur ces données; son application à plusieurs centaines de groupes d'activité supposait donc des calculs préliminaires très longs; elle était irréalisable.

6.2. — On avait également observé pour un groupe d'activité l'existence d'une relation empirique linéaire entre $\log \sigma$ et $\log \log (1/1 - p)$. Le calcul était encore plus lourd et donnait pratiquement les mêmes résultats que le précédent.

7. Autres modes de résolution pratique du problème

7.1. — On a essayé d'employer une courbe de concentration d'équation plus simple que celle du paragraphe 5, soit :

$$q = p^{\alpha};$$

d'où

$$c^2 = \frac{\alpha^2}{2\alpha - 1} - 1;$$

c'est-à-dire que c aurait une valeur *constante* pour toute les stratifications possibles.

La valeur optimum de p correspondante est donnée par :

$$p^{2\alpha - 1} = \frac{m\, C_0^{\alpha}\, (2\alpha - 1)}{2\, (\alpha - 1)^2}.$$

Par exemple, avec $\alpha = 2$, il vient $p = \sqrt[3]{3\, m\, C_0^2/2}$. On n'a pas appliqué cette méthode car on a constaté sur les données 1952 que les $\log q$ n'étaient pas du tout proportionnels aux $\log p$ (bien que, dans un intervalle de variation assez faible, c prenne des valeurs peu différentes les unes des autres, pour une distribution donnée).

7.2. — On a essayé divers autres procédés. Par exemple, il ressort des données que, lorsque q n'est pas trop grand, σ et q sont à peu près proportionnels. On a donc posé

$$\sigma = \frac{\mu}{m}\, q$$

autrement dit

$$c = \mu\, p$$

ce qui conduit à l'équation différentielle suivante de la courbe de concentration

$$\left(\frac{dq}{dp}\right)^2 - 2\, q\, (\mu^2 + p^{-1})\, \frac{dq}{dp} + \frac{q^2}{p^2} = o$$

dont les variables se séparent, et à une autre équation reliant les valeurs optimum de p et q.

Ceci n'a pas paru conduire à des résultats suffisamment simples.

7.3. — On a cherché aussi à employer une courbe de concentration plus compliquée que celle du paragraphe 5 :

On a pensé à une conique à 2 paramètres au lieu d'un.

On a également cherché à remplacer la conique par la courbe d'équation :

$$p\, q^{\alpha} + \varepsilon\, p - (1 + \varepsilon)\, q^{\alpha} = o \qquad (\varepsilon > o)$$

où ε et α auraient été choisis pour ajuster au mieux les données.

On s'est heurté à des calculs numériques trop lourds pour l'application pratique.

CHAPITRE 2

SUR LA DÉFINITION DES LOIS DE DISTRIBUTION PAR LEUR COURBE DE CONCENTRATION

1. La courbe de concentration

Nous avons fait usage de l'*équation de la courbe de concentration* dans les calculs du chapitre 1.

Rappelons la définition de cette courbe, qu'on rencontre chez les auteurs italiens (Gini, Vinci, etc...) plus que chez les auteurs anglo-saxons ou français :

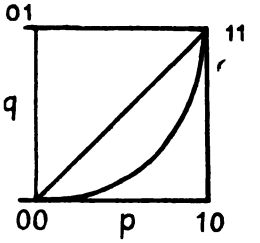
Considérons la distribution d'une variable *positive* x (par exemple le chiffre d'affaires d'un établissement industriel et commercial) ; on introduit 2 variables p et q *comprises entre 0 et 1;* à savoir :

p, proportion des redevables dont le C. A. est inférieur à x;

q, proportion du total des C. A. correspondant aux redevables dont le C. A. est inférieur à x.

Par définition, le lieu du point de coordonnées cartésiennes (p, q) est la *courbe de concentration* (fig. 1) ; soit :

Fig. 1.



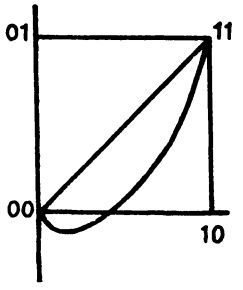
$$g(p, q) = 0$$

l'équation de cette courbe. On a évidemment

$$g(o,o) = g(1,1) = 0$$

ainsi qu'une *condition de croissance*

Fig. 2.



$$\frac{dq}{dp} > 0$$

c'est-à-dire

$$g'_p \cdot g'_q < 0$$

et *une condition de convexité*

$$\frac{d^2q}{dp^2} > 0$$

Remarque : Nous avons lu (vers 1943 dans METRON probablement) que la notion de courbe de concentration s'étendait à une variable pouvant présenter des valeurs négatives, par exemple des *bénéfices* des entreprises (industrielles et commerciales), le total des bénéfices restant positif;
(voir ci-contre la forme de la courbe de concentration).
En pareil cas la condition de croissance ci-dessus disparaît (fig. 2) ; celle de convexité subsiste.

Passage de la loi de probabilité à la courbe de concentration :

Soit :

$f(x)$: la densité de probabilité (dont on suppose l'existence);
m : l'effectif total de la population (nombre de redevables);
$m\,\bar{x} = (x)$ le C. A. total de la dite population.

On voit facilement que :

et que :

$$\left.\begin{aligned} f(x) &= m\frac{dp}{dx} \\ x &= \bar{x}\frac{dq}{dp} \end{aligned}\right\} \quad \text{d'où } f(x) = \frac{m}{\bar{x}} \cdot \frac{1}{\dfrac{d^2q}{dp^2}}$$

2. Exemple : loi de Pareto : $p = 1 - \left(\frac{a}{x}\right)^{\alpha}$

où a désigne le C. A. minimum et où $1{,}6 < \alpha < 1{,}9$ (1). Il vient d'une part :

$$f(x) = \alpha \frac{m}{a} \left(\frac{a}{x}\right)^{\alpha+1}$$

d'autre part :

$$dq = \frac{1}{\bar{x}} \frac{x f(x)}{m} dx = \frac{\alpha}{\bar{x}} \left(\frac{a}{x}\right)^{\alpha} dx$$

d'où

$$q = \frac{\alpha}{\alpha - 1} \frac{a}{\bar{x}} \left[1 - \left(\frac{a}{x}\right)^{\alpha-1}\right] \qquad \text{(avec } \alpha > 1\text{)}$$

Mais

$$\bar{x} = \int_a^{+\infty} x f(x)\, dx = \frac{\alpha}{\alpha - 1} a$$

d'où

$$q = 1 - \left(\frac{a}{x}\right)^{\alpha-1}$$

D'où

$$g(p, q) \equiv (1 - p)^{\alpha-1} - (1 - q)^{\alpha} = 0$$

3. Distributions définies par leur équation $g(p, q) = 0$

3.1. — Si l'on résout en q l'équation de la courbe de concentration, x et $f(x)$ s'en déduisent (par 2 dérivations), en fonction du paramètre arbitraire $\bar{x}$. Ainsi la courbe de concentration définit une certaine famille de distributions statistiques « *semblables entre elles* »; la valeur moyenne de la variable caractérise chacune des courbes de cette famille.

Il est remarquable que les courbes de concentration les plus simples et les plus naturelles ne correspondent pas à des distributions d'usage courant.

3.2. — Nous nous sommes attardé sur le cas où cette courbe serait assimilable à un *arc de conique*, c'est-à-dire :

$$\begin{aligned} g(p, q) &\equiv a\,p\,(1 - p) + b\,q\,(1 - q) + c\,p\,(1 - q) + d\,q\,(1 - p) \\ &\equiv (a + c)\,p + (b + d)\,q - [ap^2 + (c + c)\,pq + bq^2] \end{aligned}$$

Et nous avons trouvé que cet arc présentait la forme voulue, à condition d'avoir simultanément :

$$\begin{aligned} &(a + c)\,(b + d) < o \\ &(b + c)\,(a + b + c + d) < o \\ &(b + d)\,(a\,b - c\,d)\,(a + b + c + d) > o \end{aligned}$$

La courbe ainsi définie s'est révélée trop lourde pour la pratique.

---

(1) Dans une communication de M. Theil au Colloque d'Économétrie, Paris 1955, on indique que suivant les pays, la constante α de Pareto s'échelonne en fait entre 1,1 et 2,6.

3.3. — Nous sommes passé à un seul paramètre, en faisant usage de l'hyperbole équilatère d'équation :

$$pq + \varepsilon p - (1 + \varepsilon) q = 0 \qquad (\varepsilon > 0)$$

qui présentait d'ailleurs le grave défaut d'être symétrique par rapport à la droite d'équation

$$p + q = 1$$

ce qui n'est pas le cas de beaucoup de courbes de concentration réelles. On obtient :

$$x = \frac{\bar{x}\,\varepsilon\,(1 + \varepsilon)}{(1 + \varepsilon - p)^2}$$

$$f(x) = \left[\frac{m}{2}\sqrt{\bar{x}\,\varepsilon\,(1 + \varepsilon)}\right]\frac{1}{x^{3/2}}$$

Cette distribution n'est pas sans quelque analogie avec celle de Pareto. C'est ainsi qu'on a

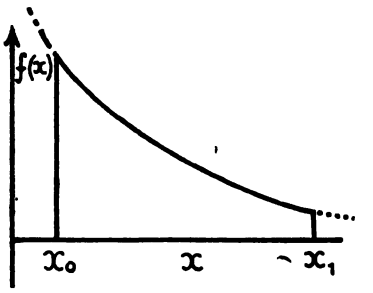


Fig. 3.

Pareto	Hyperbole équilatère
$1 - p = \left(\frac{a}{x}\right)^{\alpha}$	$1 - \frac{p}{1 + \varepsilon} = \left(\frac{a}{x}\right)^{1/2}$
avec $1{,}6 < \alpha < 1{,}9$	avec $a = x(0) = \frac{\bar{x}\,\varepsilon}{1 + \varepsilon}$

Remarque 1. — x ne peut croître à l'infini (ni d'ailleurs être nul), ce qui est sans inconvénient dans la pratique.

Remarque 2. — Pour $p = o : x_0 = \dfrac{\bar{x}\,\varepsilon}{1 + \varepsilon}$

pour $p = 1 : x_1 = \dfrac{\bar{x}\,(1 + \varepsilon)}{\varepsilon}$ } donc $\bar{x} = \sqrt{x_0 \cdot x_1}$

Remarque 3. — Il est possible que, sans la fraude fiscale (qui agit surtout sur les faibles chiffres d'affaires), l'ajustement des C. A. eût exigé un exposant de (a/x) très supérieur à 1/2.

4. — Role de la courbe de concentration dans le problème de la stratification optimum

Montrons comment la relation $g(p, q) = o$ permet d'expliciter l'équation (3') du chapitre 1 :

$$\frac{d(\sigma^2)}{dp} = \frac{1}{a\,p^2} \qquad (3')$$

De

$$x = \bar{x}\frac{dq}{dp} = \frac{(x)}{m}\frac{dq}{dp}$$

on passe à

$$(x_2) = (x)\int_o^p \frac{dq}{dp}\,dp = (x)\,q$$

donc

$$\bar{x}_2 = (x_2)/m_2 = (x)\,q/mp = \bar{x}\,q/p$$

Le deuxième moment de x de la strate inférieure est :

$$\sigma^2 + \bar{x}_i^2 = \frac{1}{mp}\int_o^p \left(\bar{x}\frac{dq}{dp}\right)^2 m dp$$

Calculons :

$$c^2 = \sigma^2/\bar{x}_i^2 \text{; il vient}$$

$$c^2 + 1 = \frac{p}{q^2}\int^p \left(\frac{dq}{dp}\right)^2 dp$$

L'équation (3') de l'optimum s'écrit :

$$\frac{d}{dp}\left(\frac{q^2}{p^2}c^2\right) = \frac{C^2 m}{p^2}, \qquad (3'a)$$

autrement dit :

$$\frac{d}{dp}\left[\frac{1}{p}\int_o^p \left(\frac{dq}{dp}\right)^2 dp - \frac{q^2}{p^2}\right] = \frac{C^2 m}{p^2}, \qquad (3'b)$$

(3'*a*) (3'*b*) donnent la solution générale du problème *d'optimum stratification*, quelle que soit l'équation $g\,(p, q) = 0$ de la courbe de concentration.

Remarque. — Le coefficient de variation de la strate inférieure (dont l'expression vient d'être obtenue) ne dépend que de la courbe de concentration; tandis que l'équation de l'optimum dépend de l'effectif m de la distribution.

Il était à prévoir que, si l'on compare deux distributions *semblables*, à celle qui a le plus fort effectif correspond la valeur optimum de p la *plus élevée* (c'est-à-dire une strate inférieure sondée proportionnellement plus développée).

On observera pourtant que l'équation ne dépend que de $C_o^2 m$. Pour une loi de distribution donnée, la précision $1/C_o$ *varie comme* $\sqrt{m}$, racine carrée de l'effectif de la population sondée.

CHAPITRE 3

SUR L'AGRÉGATION DES LOIS DE DISTRIBUTION

1. *Agrégation de plusieurs distributions.*

1.1. — Considérons par exemple la distribution des C. A. des épiciers, agrégat des deux distributions relatives aux épiciers détaillants et aux épiciers en gros. Soit :

(1) $$g\,(p, q) = 0, \quad g_1\,(p_1\,q_1) = 0,$$

les équations des 2 courbes de concentration correspondantes. On admet que les C. A. des deux distributions sont exprimés dans la même monnaie; la variable x est donc commune aux deux distributions et à leur agrégat; et il vient :

(2) $$\bar{x}\frac{dq}{dp} = \bar{x}_1\frac{dq_1}{dp_1}.$$

ou encore

(2') $$\frac{\frac{1}{m}\frac{\partial g}{\partial p}}{\frac{1}{(x)}\frac{\partial g}{\partial q}} = \frac{\frac{1}{m}\frac{\partial g_1}{\partial p_1}}{\frac{1}{(x_1)}\frac{\partial g_1}{\partial q_1}}$$

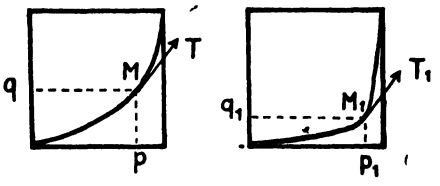


Fig. 4.

Cette relation entre les pentes des 2 courbes de concentration définit *une correspondance* entre points homologues M et M_1 correspondant à une même valeur de la variable x.

1.2. — Soit (P, Q) les coordonnées d'un point de la courbe de concentration de l'agrégat. On a les relations suivantes :

entre les effectifs	$m + m_1 = M$	(3)	$m\,p + m_1\,p_1 = M\,P$	(4)
entre les C. A. totaux	$(x) + (x_1) = (X)$		$(x)\,q + (x_1)\,q_1 = (X)\,Q$	

L'équation de la courbe de concentration de l'agrégat s'obtiendra en éliminant les 4 quantités $p\ q\ p_1\ q_1$ entre les 5 relations (1) (2) et (4).

1.3. — On étend immédiatement ces résultats au cas où l'on envisage plus de 2 distributions.

2. *Sur le groupe des distributions de même nature économique : Problème général*

Divers auteurs ont proposé des expressions (dépendant de certains paramètres) susceptibles de représenter (plus ou moins heureusement) la distribution des revenus ou des salaires, des effectifs, des chiffres d'affaires, etc... d'un pays donné (lois de Pareto, de Gibrat, etc...).

On ne semble guère s'être arrêté au fait suivant : la distribution statistique (de revenus, de salaires, etc...) ainsi considérée résulte de la juxtaposition de distributions analogues concernant des sous-populations. Par exemple :

La distribution des établissements industriels et commerciaux suivant leur chiffre d'affaires résulte de la juxtaposition des distributions des chiffres d'affaires des boucheries, des épiceries, des fonderies, des banques, etc...

On se trouve donc devant le problème suivant (1) : existe-t-il des lois de distribution assez générales pour représenter la distribution des *chiffres d'affaires d'un groupe économique?*

— étant entendu qu'en agglomérant plusieurs groupes économiques, comme on le fait en pratique statistique, on devrait retrouver une loi de distribution du *même type.*

(1) A vrai dire, c'est par M. Fonsagrive que nous avons entendu poser d'abord ce problème dans toute sa généralité.

Par exemple : la distribution des C. A. des épiciers devrait s'obtenir en agrégeant celles des épiciers en gros et des épiciers détaillants.

La distribution des C. A. des commerces de l'alimentation devrait s'obtenir en agrégeant celles des épiciers, des bouchers, des boulangers, etc...

Il est bien clair que le type de distribution cherché ne pourrait être qu'extrêmement général.

La question serait de savoir s'il n'existe pas des distributions *formant un groupe* (*au sens mathématique du mot*, cette fois), c'est-à-dire telles qu'en « agrégeant » deux d'eux d'entre elles (de la façon décrite au n° 1) on retrouve une distribution du même groupe.

Le groupe général des distributions (de variable positive) possède bien entendu cette propriété; mais on aimerait être sûr qu'il n'existe pas des groupes plus restreints à l'intérieur de celui-ci.

Il faudrait trouver des fonctions $\mathcal{G}\,(\mathfrak{X}, \mathfrak{Z})$ telles que l'on ait (symboliquement)

$$(5)\qquad \frac{\dfrac{1}{m}\dfrac{\partial \mathcal{G}}{\partial p}}{\dfrac{1}{(x)}\dfrac{\partial \mathcal{G}}{\partial q}} = \frac{\dfrac{1}{m_1}\dfrac{\partial \mathcal{G}}{\partial p_1}}{\dfrac{1}{(x_1)}\dfrac{\partial \mathcal{G}}{\partial q_1}} = \frac{\dfrac{1}{\mathrm{M}}\dfrac{\partial \mathcal{G}}{\partial \mathrm{P}}}{\dfrac{1}{(\mathrm{X})}\dfrac{\partial \mathcal{G}}{\partial \mathrm{Q}}}$$

3. *Problème limité.*

Renonçant à aborder un problème aussi difficile, on a cherché des fonctions $\mathcal{G}$ de forme déterminée, en supposant qu'il existe des relations particulières entre (x), (x_1), m. m_1.

Exemple 1 :

$$\mathcal{G} \equiv \mathfrak{X}^2 + \mathfrak{Z}^2 - 2\,\mathfrak{Z} = 0 \text{ : a-t-on l'identité voulue?}$$

$$\frac{\dfrac{p}{m}}{\dfrac{q-1}{(x)}} = \frac{\dfrac{p_1}{m_1}}{\dfrac{q_1-1}{(x_1)}} = \frac{\dfrac{mp + m_1 p_1}{(m+m_1)^2}}{\dfrac{(x)\,q + (x_1)\,q_1 - (x) - (x_1)}{(x) + (x_1)^2}} \quad ?$$

On a en réalité l'identité ci-dessous :

$$\frac{\dfrac{p}{m}}{\dfrac{q-1}{(x)}} = \frac{\dfrac{p_1}{m_1}}{\dfrac{q_1-1}{(x_1)}} = \frac{\alpha\dfrac{p}{m} + \beta\dfrac{p_1}{m_1}}{\alpha\dfrac{q-1}{(x)} + \beta\dfrac{q_1-1}{m_1}}$$

On ne parvient au résultat voulu que si :

$$\sqrt{\alpha} = \frac{m}{m+m_1} = \frac{(x)}{(x)+(x_1)} \quad \text{(et analogue pour } \sqrt{\beta}\text{)}$$

autrement dit si $\bar{x} = \bar{x}_1$ (agrégation de groupes de même C. A. moyen).

Exemple 2 :

$$\mathcal{G} = (a+c)\,\mathfrak{X} + (b+d)\,\mathfrak{Z} - [a\,\mathfrak{X}^2 + (c+d)\,\mathfrak{X}\mathfrak{Z} + b\,\mathfrak{Z}^2] = 0$$

Il vient pour (2')

$$\bar{x}\,\frac{(a+c)-2ap-(c+d)\,q}{(b+d)-(c+d)\,p-2bq}=\bar{x}_1\,\frac{(a_1+c_1)-2a_1p_1-(c_1+d_1)\,q_1}{(b_1+d_1)-(c_1+d_1)\,p_1-2b_1q_1}$$

En général, même si $\bar{x}=\bar{x}_1$, on ne peut satisfaire à (5).

Exemple 3 :

$$\mathcal{C}_{jj}=\mathfrak{P}\mathfrak{Q}+\varepsilon\,\mathfrak{P}-(1+\varepsilon)\,\mathfrak{Q}=0$$

Il vient pour (2') :

$$\bar{x}\,\frac{q+\varepsilon}{p-(1+\varepsilon)}=\bar{x}_1\,\frac{q_1+\varepsilon_1}{p_1-(1+\varepsilon_1)}$$

et, pour $\bar{x}=\bar{x}_1$, on peut satisfaire à (5); l'agrégation est donc alors possible, le paramètre de la distribution de l'agrégat étant

$$\frac{m\,\varepsilon+m_1\,\varepsilon_1}{m+m_1}=\frac{(x)\,\varepsilon+(x_1)\,\varepsilon_1}{(x)+(x_1)}$$

Ainsi les distributions qui ont été utilisées pour résoudre le problème de stratification optimum ont une propriété intéressante.

P. Thionet.

***