APPLICATIONS BIOMÉTRIQUES DE LA THÉORIE DE L'INFORMATION

par M. P. SCHÜTZENBERGER

L'IMPORTANCE accordée par la statistique moderne à la planification préalable des expériences et des observations est peut-être l'aspect le plus frappant de son développement actuel.

Au lieu de considérer comme données antérieurement à l'intervention du statisticien les expériences et leurs résultats et de chercher seulement alors à en extraire le maximum d'informations, le statisticien actuel doit plutôt s'efforcer de prévoir comment doit être organisée la technique d'observation afin qu'en découlent au mieux les renseignements qu'il sollicite. Ainsi sont nées, par exemple, des méthodes telles que l'analyse séquentielle à laquelle nous avons déjà consacré un article dans cette revue (5) qui permet à précision égale de réduire en moyenne de moitié le nombre des observations nécessitées par un contrôle statistique ; c'est aussi le cas des techniques de « design of experiment » de l'analyse de variance grâce auxquelles, à coût égal, la sécurité et la richesse des conclusions possibles sont considérablement augmentées. Les méthodes que nous exposerons ici ont ce même caractère d'économie d'observations et d'accroissement du rendement en information par le choix préalable d'un schéma opératoire adapté au but poursuivi, et elles semblent pouvoir trouver un large emploi aussi bien dans la recherche théorique que dans les applications industrielles et ceci tout particulièrement en biologie où nombre d'épreuves entraînent le sacrifice d'un animal.

Un exemple : Le dépistage de la syphilis.

Pour illustrer d'un cas concret la méthode que nous voulons exposer nous prendrons l'exemple familier du dépistage des réagines syphilitiques par la réaction de Bordet-Wasserman, exemple qui présente en outre l'avantage de correspondre à des essais effectués pendant la guerre aux U.S.A. avec des techniques voisines de celles que nous préconisons.

Prenant avantage de la sensibilité de la réaction, Dorfman (1) avait suggéré d'effectuer celle-ci, non pas sujet par sujet mais sur le mélange des sérums de x sujets : si aucun de ceux-ci ne contient de réagines la réaction sera négative (1) et, d'un seul coup, les x sujets seront diagnostiqués. Si au contraire un ou plusieurs des sujets présentent une réaction positive, il en sera de même pour le mélange des sérums et l'on retestera chacun des individus du groupe : appelons p la probabilité pour un sujet de présenter un B.W. positif, l'on a manifestement les relations suivantes :

Probabilité pour que le mélange des sérums de x sujets donne une réaction négative : $(1 - p)^x$.

Probabilité pour qu'au moins un sujet du groupe de x présente une réaction positive : $1 - (1 - p)^x$.

En moyenne, on effectuera donc pour chaque groupe de x sujets, un nombre de tests égal à $1 + x\,[1 - (1-p)x]$ et l'on voit que pour p petit (ce qui est en général le cas) et x convenablement choisi, cette valeur sera très largement inférieure aux x tests nécessités par la méthode courante.

Par exemple, pour p = 1 0/00, un millier de sujets seront diagnostiqués au moyen de 196 tests en moyenne seulement en utilisant un groupage de sérums 11 par 11, soit une économie de plus de 80 %. Nous verrons plus loin que la technique de Dorfman peut être améliorée et ce nombre réduit à 150 par une application conséquente du même principe général, c'est-à-dire en retestant des mélanges de sérums prélevés dans certains sous-groupes du groupe initial.

L'on voit aussi qu'une méthode analogue aurait pu être employée dans des types de problèmes beaucoup plus variés que ceux considérés par Dorfman. Nous nous limiterons à étudier ici les cas où l'on peut utiliser la méthode pour estimer la proportion des sujets malades dans une certaine population, ou bien encore pour se procurer un certain nombre de sujets dont le sang contiendrait la réagine cherchée et dont on aurait besoin pour telle ou telle recherche de laboratoire ou encore, inversement, pour sélectionner un lot de sujets sains, etc... Ces divers exemples correspondraient respectivement à ce que nous appelerons plus loin les pro-

(1) Il serait probablement plus juste de dire que la réaction est « nulle » plutôt que « négative », mais nous nous conformerons à cet usage dans toute la suite de cet article.

blèmes de *diagnostic, d'estimation* et de *triage,* ce dernier cas se distinguant du premier en ce que le diagnostic d'un sujet malade (ou sain) est considéré comme seul but de l'expérimentation, et que le diagnostic d'un sujet du type non recherché est considéré comme ne présentant aucun intérêt.

Le domaine d'application de la méthode.

Il est évident que le champ d'application de semblables méthodes est beaucoup plus vaste. Sous réserve de certaines hypothèses, on pourra faire appel à elles chaque fois qu'une observation (un test) unique portant sur un groupe de x objets permettra de garantir :

— Soit que ces objets appartiennent tous à un même type, que nous appelerons « *neutre* » (ou, pour abréger N) et qui serait dans l'exemple précédent l'état normal. Nous dirons alors que le résultat du test est « *négatif* » ;

— Soit qu'*au moins un* d'entre eux appartient au type *marqué* (pour abréger M) et qui était représenté dans notre exemple par la présence de réagines dans le sérum. Nous dirons alors que le résultat du test est « positif ».

Nous avons réuni ci-dessous quelques exemples typiques qui rendront tout à fait claire cette notion. On remarquera que dans certains cas la définition des types N ou M dépend essentiellement de la technique d'observation employée. Par exemple dans les déterminations

TABLEAU I

Objets	Technique d'observation	Objets de type « N »	Objets de type « M »
Groupes sanguins O ou A et B ou AB	Réaction d'agglutination avec sérum A.	Groupe A et O	Groupe B et AB
Groupes sanguins O ou B et A ou AB	Réaction d'agglutination avec sérum B.	Groupe B et O	Groupe A et AB
Urines de sujets éventuellement excréteurs d'acide phénylpyruvique..	Test au perchlorure de fer.	Sujets normaux	Sujets excréteurs
Appareils électriques .	Mise en série et vérification du passage du courant.	Appareils en état de marche	Appareils hors d'état de marche
Séquence de calculs numériques ..	Vérification du résultat final (on suppose qu'une compensation des erreurs est absolument improbable).	Opérations exactes	Opérations erronées
Produit de nombres quelconques	Observation de la valeur nulle ou non du résultat final.	Nombres différents de zéro	Nombres égaux à zéro
Produit de chiffres non nuls ..	Observation de la valeur paire ou impaire du résultat final.	Nombres impairs	Nombres pairs
Gènes mendéliens	Détermination du phénotype.	Allèles récessifs	Allèles dominants
Prélèvements bactériologiques	Ensemencement.	Prélèvements aseptiques	Prélèvements septiques

des groupes sanguins les sujets A sont « *neutres* » pour le sérum A mais « *marqués* » pour le sérum B. En revanche les sujets « O » sont toujours « *neutres* » à ce point de vue dans les tests classiques et les sujets A B toujours « *marqués* ». Il n'est pas besoin d'insister par conséquent sur le fait que si, *relativement à l'observation du phénotype,* les gènes récessifs se comportent comme des caractères de type N, il n'existe absolument aucune liaison autre que fortuite entre notre classification et les phénomènes de la génétique (par exemple l'arriération mentale phényl-pyruvique s'hérite selon le mode *récessif* et, relativement à la réaction au perchlorure de fer d'un mélange d'urines, la présence de cette anomalie doit être considérée comme étant de type M).

Enfin il faut encore rappeler que notre discussion sera essentiellement théorique en ce sens que nous n'envisagerons pas dans chaque cas les limitations inhérentes à la technique d'observation : il est évident que quelle que soit la sensibilité du test au perchlorure de fer il serait absurde de tester par une seule réaction en mélangeant leurs urines 2.000 malades mentaux, sous prétexte que ce nombre serait optimal du point de vue du « design of experiment ». Mais il serait dans ce cas précis probablement aussi peu fondé de refuser de procéder à un groupage des sujets disons 10 par 10 qui, représentant une sérieuse économie de temps, permettrait donc d'examiner beaucoup plus de sujets et, par exemple, de mieux estimer dans une population la fréquence du gène responsable de l'anomalie.

De même nous ne discuterons pas la possibilité (que nous supposerons systématiquement être exclue a priori) d'une éventuelle interaction entre les objets composant un groupe, soit que cette interaction fasse apparaître un résultat positif alors que tous les objets sont N, soit qu'elle risque au contraire de masquer la présence d'un objet M. Il ne peut s'agir là que de cas d'espèce, propres à telle ou telle technique et c'est le spécialiste seul qui a capacité, sur la base des indications fournies par le mathématicien, pour décider de la validité des méthodes proposées et des limites de leur application.

Il est à noter toutefois que, même dans le cas où il y a un risque d'interaction, divers artifices permettent d'employer ces méthodes : dans le cas du diagnostic de la syphilis, admettons que la probabilité soit inférieure à 1% par exemple, pour que les réagines d'un sujet malade soient masquées par une éventuelle interaction avec le sérum d'un autre sujet au hasard.

Nous groupons les sujets 15 par 15 de telle sorte que chacun d'eux appartienne à 3 groupes et ne se rencontre jamais associé 2 fois avec le même sujet, ce qui est d'ailleurs facile à réaliser (1). Nous avons alors moins de 2,5

(1) Voici la méthode la plus simple : On prend un lot de 225 sujets que l'on numérote conventionnellement par un couple de deux chiffres (a, b) a et b prenant toutes les valeurs de 1 à 15.

Les trois classes de groupages à effectuer sont alors obtenues en réunissant les sujets ayant :

1) Soit le même nombre a.

2) Soit le même nombre b.

3) Soit la même valeur de la différence a-b (si a est plus grand que b, ou $15+a-b$ si a est plus petit que b).

Par exemple (9,1) (12,4) (4,11) et (7,14) appartiennent à un même groupe.

chances sur mille pour qu'un malade passe inaperçu ce qui est une probabilité très faible et nous n'avons cependant effectué que 3 fois plus de test que dans la méthode la plus rapide, soit encore une réduction de plus de 50 % par rapport à la technique habituelle sujet par sujet. Nous ne discuterons pas ici les problèmes d'optimalité qui peuvent se poser en rapport avec une semblable procédure.

La représentation des schémas opératoires.

Dans la méthode de Dorfman le groupage des sérums n'intervenait qu'une seule fois. Il est évident que des schémas plus complexes, dans lesquels divers sous ensembles d'un même ensemble sont testés successivement, doivent être considérés si l'on veut donner une pleine efficacité à la méthode.

La manière la plus claire de décrire un schéma opératoire est d'employer une représentation graphique telle que la suivante :

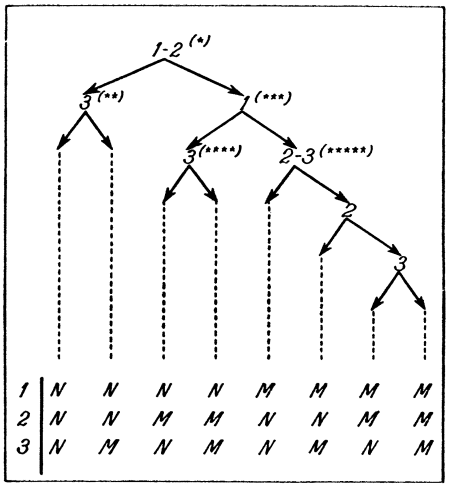

Fig. 1

L'arbre ci-contre (fig. 1) se lit de haut en bas.

Les objets 1 et 2 sont testés ensemble (*).

Si le résultat est négatif (flèche en bas à gauche) l'objet 3 est testé (**).

Si le résultat est positif (flèche en bas à droite) l'objet est alors testé seul (***).

Si ce résultat est négatif, 3 est testé à son tour (****) ; s'il est positif, 2 et 3 sont testés ensemble (*****) etc...

Au bout de chaque « branche » sont indiquées les conclusions que l'on peut tirer à la suite de la série des essais.

Un pareil schéma décrit parfaitement un ensemble de tests puisqu'il indique à chaque étape le nouveau test qu'il convient d'effectuer compte tenu des résultats antérieurs.

Les schémas réguliers.

Tous les schémas opératoires convenables a priori ne sont d'ailleurs pas utilisables pratiquement en raison de leur complexité trop grande qui risquerait d'entraîner des erreurs de procédure, d'autant plus graves qu'elles pourraient passer entièrement inaperçues. Nous nous limiterons donc à une catégorie particulièrement simple de schémas que nous appelerons « *réguliers* » et qui sont caractérisés par le fait que si un groupe de x objets testés ensemble a donné un résultat positif, les objets de ce groupe ne seront plus jamais testés que conjointement par la suite. Comme nous le verrons plus bas une conséquence de cette limitation est que, à chaque moment de la procédure, l'information que l'on possède relativement à un objet donné α est réduite à l'une des quatre possibilités suivantes :

1) α est M ;

2) α est N ;

3) α appartient à un groupe de x objets ayant donné un résultat positif ;

4) ? on ne sait rien sur α.

Nous excluons ainsi délibérément de ce travail les schémas trop complexes tel que celui indiqué dans la fig. 2

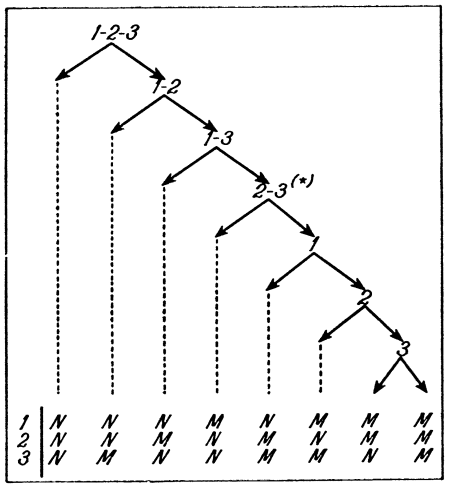

Fig. 2

(qui est pourtant optimal pour le diagnostic de 3 objets quand la probabilité commune que l'un de ceux-ci soit M est inférieure à

$$\frac{\sqrt{65}-7}{8} \neq 0{,}133).$$

puisqu'au moment désigné par (*) sur la figure notre information sur 1 n'appartient à aucune des quatre classes possibles énumérées ci-dessus.

Dans la pratique on peut le plus souvent considérer que la probabilité a priori pour un objet d'être de type M (ou N) a la même valeur « p » pour tous les objets. Nous ferons donc systématiquement cette hypothèse et il est évident que c'est seulement dans ce cas que nous pourrons dire que telle ou telle procédure est « optimale » c'est-à-dire comporte, pour la probabilité donnée p (que l'on connait presque toujours à l'avance au moins de manière approximative) un nombre moyen minimum de tests.

Remarquons d'ailleurs qu'en général la détermination exacte de la procédure « *la plus strictement optimale* »

nécessiterait la connaissance précise (souvent à quelques unités pour cent près) de la valeur de p.

Trop souvent aussi nous sommes incapables de calculer a priori la meilleure procédure sans faire intervenir quelques tâtonnements : il doit donc être entendu que les schémas que nous proposons ne sont « optimaux » que dans un sens bien défini mathématiquement. On verra dans les exemples qu'ils assurent quand même un gain important (souvent dans l'ordre de 50 pour cent) par rapport aux méthodes habituelles.

Quelques remarques importantes s'imposent : si un test sur un ensemble de x objets, a donné un résultat positif et si un second test sur un ensemble de $y < x$ objets choisis parmi ces derniers donne lui aussi un résultat positif, nous n'avons plus aucune information sur les x — y autres objets : ceci serait justifié aisément par le calcul et est d'ailleurs parfaitement intuitif.

Il en résulte que chaque fois que dans une séquence de tests nous aboutirons à un résultat positif, nous pourrons rejeter dans le lot initial certains éléments sur lesquels nous n'avons obtenu absolument aucune information bien qu'ils aient fait partie de groupes testés, ce qui permet de simplifier à la fois la discusion mathématique et la procédure opératoire.

D'autre part, il est évident que si un test sur 2 objets donne un résultat positif et que nous apprenons par la suite que l'un de ces objets est de type *neutre,* il est sûr que l'autre est de type *marqué* sans qu'il soit besoin d'un nouveau test, d'où une nouvelle raison d'économie de ce mode d'expérimentation.

Il n'entre évidemment pas dans nos intentions de faire ici la théorie mathématique de ces questions et nous nous bornerons à donner des règles pratiques et des illustrations concrètes pour chacune des trois variétés de problèmes que nous avons évoqués plus haut.

1) *Diagnostic* : étant donné un certain nombre d'objets, déterminer le type N ou M auquel appartient chacun d'eux.

2) *Estimation* : estimer par sondage la fréquence du type M (ou N) dans une population supposée très nombreuse.

3) *Triage* : Extraire d'une population (supposée aussi très nombreuse) un nombre donné à l'avance des objets, soit appartenant au type M, soit appartenant au type N.

PROBLÈMES DE DIAGNOSTIC

Règle : *A chaque stade de l'expérience choisir le nombre x d'objets testés simultanément de telle sorte que la probabilité d'un résultat positif ou négatif soit la plus voisine possible de 1/2.*

1er *Exemple :* une technique de préparation d'un produit biologique donne avec une certaine régularité environ 10 % d'unités présentant un défaut. Ce défaut a un caractère « *marqué* » par rapport aux unités qui sont « *neutres* » et relativement à une certaine technique de contrôle (par exemple le défaut est constitué par une certaine impureté provoquant une réaction caractéristique). Comment tester le plus économiquement les résultats de la fabrication d'une série ? : si nous testons chaque unité nous avons à chaque fois 90 chances pour 100 d'obtenir un résultat négatif ; si nous groupons les unités par deux et que nous testons le mélange cette proportion descend à $90/100 \times 90/100 = 81/100$; à un groupage par 3 correspond une probabilité de 73/100 ; à un groupage par 4 : 66 p. 100 ; par 5 : 59 p. 100 ; par 6 : 53 p. 100 ; par 7 : 48 p. 100. C'est donc cette dernière valeur qu'il faut retenir et la méthode optimale consistera (si cela est techniquement possible) à tester les mélanges des prélèvements effectués dans 7 unités qui se trouveront vérifiées d'un seul coup si le résultat du test est négatif. S'il est positif c'est qu'au moins une des 7 unités est marquée.

Si nous testions alors une seule de celle-ci la probabilité d'un résultat négatif serait $\frac{0{,}9 - (0{,}9)^7}{1 - (0{,}9)^7}$, soit à peu près 0,72 ; si nous en testions 2 cette probabilité deviendrait $\frac{(0{,}9)^2 - (0{,}9)^7}{1 - (0{,}9)^7} \neq 0{,}55$ et 0,40 si nous en testions 3 à la fois.

Il convient donc de prendre 2 unités et de les tester ensemble. Si le résultat est positif nous testerons l'une d'elle et nous ne saurons plus rien sur les 5 unités laissées de côté. Il faut donc les retester en bloc, mais il est plus économique d'y adjoindre 2 nouvelles unités afin de se rapprocher du chiffre optimal de 7 conformément aux remarques faites plus haut. On a donc un schéma tel que le suivant (fig. 3).

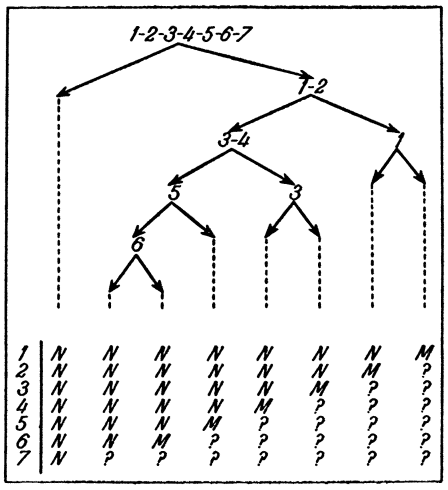

Fig. 3

2e *Exemple :* Soit à déterminer le seuil de gustation d'un sujet à la phényl-thiocarbamide.

L'on peut considérer comme *neutre* (au sens spécial où nous employons ce mot ici) le caractère : « *ne trouve pas le goût pour des concentrations inférieures à C* ». La règle indiquée plus haut nous montre alors que la procédure la plus rapide pour trouver le seuil, consiste à tester le sujet à chaque fois avec des concentrations telles que l'on ait une chance sur 2 de lui voir trouver du

goût à la solution, ce qui est assez facile si l'on possède une estimation empirique de la distribution des seuils dans la population à laquelle appartient le sujet.

3e *Exemple : dépistage d'une erreur dans un tableau d'intercorrélations :* on peut supposer que les erreurs sont très rares et indépendantes et sachant qu'une erreur s'est glissée dans le calcul des intercorrélations entre un grand nombre x de variables, on cherche à « retrouver » celle-ci. La méthode la plus expéditive consiste à utiliser la formule classique pour chaque groupe de y variables

$$\text{var}\,(\Sigma a_i) = \Sigma \text{var}\,(a_i) + 2\,\Sigma \text{cov}\,(a_i, a_j)$$

en calculant directement le membre de gauche et en déduisant la valeur du membre de droite des chiffres à contrôler. D'après les hypothèses faites la chance pour les tests de donner un résultat négatif sera la plus voisine de 1/2 si $\left[\frac{y}{2}\right] = 1/2\left[\frac{x}{2}\right]$ c'est-à-dire si y est l'entier le plus voisin de $\frac{1+\sqrt{2x^2-2x+1}}{2} \neq \frac{x\sqrt{2}}{2}$

Si la vérification donne un résultat positif c'est qu'il y a une erreur parmi les y variables. On opérera alors sur elles comme on a opéré sur les x premières ; si elle donne un résultat négatif $\left[\frac{y}{2}\right]$ covariances auront été vérifiées d'un coup. Pour continuer la vérification, le plus simple est d'employer la même formule mais cette fois-ci en ajoutant z variables aux y précédentes et ceci de telle sorte que :

$$\left[\frac{y+z}{2}\right] - \left[\frac{y}{2}\right] = 1/2\left\{\left[\frac{x}{2}\right] - \left[\frac{y}{2}\right]\right\}$$

Si le résultat est positif c'est qu'il y a une erreur parmi l'une des covariances entre les y premières variables et les z suivantes ou dans les covariances entre ces dernières. Sinon on ajoutera encore z' nouvelles variables conformément à la règle de la probabilité 1/2.

PROBLÈMES D'ESTIMATION

Nous ne démontrerons pas le théorème suivant qui joue un rôle fondamental et qui est d'ailleurs assez intuitif.

Si on se limite aux schémas réguliers tels que nous les avons définis, la précision la plus grande à coût égal est obtenue en choisissant à chaque fois x nouveaux objets, x étant tel que la probabilité d'un résultat négatif soit le plus près possible d'une certaine constante pratiquement égale à 1/5.

Comme on le voit facilement la méthode revient simplement à estimer d'abord q^x puis $\hat{q}$ au lieu de q (q : fréquence des caractères N), a et b désignant respectivement le nombre de test ayant donné un résultat positif et négatif on a : $\hat{q} = \left(\frac{b}{a+b}\right)^{1/x}$ ce qui est l'estimation par le maximum de vraisemblance dont la variance est : $(a+b)^{-1}\, q^{2-x}\,(1-q^x)\, x^{-2}$.

On remarquera que pour $x = 1$ on a la formule habituelle

$$(a+b)^{-1}\, q\,(1-q)$$

et, pour $x = 2$, la formule donnant la variance dans l'estimation de la fréquence d'un gène :

$$0{,}25\; q^{-1}\,(1-q^2)$$

Exemple : l'on sait d'avance que la fréquence d'un certain caractère de type N est de l'ordre de 0,9. Quel doit être le groupage qui donne la plus grande précision dans l'estimation de cette probabilité ?

Les tables indiquent que la quinzième puissance de 0,9 est 0,206, la seizième puissance étant 0,185. On groupera les objets 15 par 15 et l'estimation p de p pour m semblables opérations aura une variance de 0,0123 m au lieu de 0,090 m avec une méthode d'observation objet par objet. *La précision est donc 7 fois plus grande à coût égal.*

Le tableau I donne pour x plus petit ou égal à 20 les valeurs limites de p pour lesquelles chaque groupage est efficace : ainsi dès que p est inférieur à 1/3 il est fructueux de tester par groupes de 2, et par groupes de 3 dès que p est inférieur à 0,525, etc... Réciproquement le tableau II donne les valeurs optimales de x pour les petites valeurs de p.

TABLEAU I

Valeurs de x	Valeurs limites de p
1	0,67
2	0,51
3	0,31
4	0,30
5	0,25
6	0,22
7	0,19
8	0,17
9	0,15
10	0,14
11	0,13
12	0,12
13	0,11
14	0,10
15	0,10
16	0,09
17	0,09
18	0,08
19	0,08
20	

TABLEAU II

Valeurs de p	Valeurs optimales de x
0,10	15
0,09	17
0,08	19
0,07	22
0,06	26
0,05	31
0,04	39
0,03	52
0,025	63
0,02	78
0,015	105
0,010	157
0,005	320
0,002	798
0,001	1.592

Il est à noter que le gain réalisé par l'emploi du groupage convenable est d'autant plus grand que p est plus petit (il croît sans limite quand p tend vers zéro ce qui est d'ailleurs naturel *a priori*).

Il est intéressant de compléter par quelques remarques élémentaires la discussion dans le cas où, en raison d'une sensibilité insuffisante de test, il se pourrait que le groupage de ces objets risque de donner un résultat négatif alors que le groupe contient un objet marqué. Nous supposerons pour simplifier que l'inverse est impossible (si tous les objets sont N alors le test est toujours négatif) et que l'on connaît la probabilité s d'un résultat faussement négatif.

Avec les mêmes notations que précédemment, q est estimé par : $\hat{q} = \left(\frac{b + a\,(1-s)}{b + as}\right)^{1/x}$ et a pour variance :

$$q^{2-2x}\, s^{-1}\, x^{-2}\, (a+b)^{-1}\, (1-q^x)\, (1-s+sq^x)$$

Une étude complète de l'optimalité de la procédure exigerait une connaissance de s en fonction de n. Nous nous bornerons à faire observer que si s est voisin de 1, ce qui est en général le cas, la variance ne diffère guère de sa valeur dans le cas idéal que par un facteur s^{-1} et que ce changement est pratiquement négligeable par rapport à la réduction considérable que procure le groupage, ce qui montre que la méthode reste applicable sans modification si l'on connaît s avec assez de précision.

PROBLÈME DE TRIAGE

Deux cas sont évidemment à distinguer selon que l'on se préoccupe de trier les objets porteurs du caractère N ou les objets porteurs du caractère M.

Dans le premier cas, le plus simple, il est facile de voir que x doit être le plus voisin possible d'une racine de l'équation : $p^x + x \log p = 0$.

Pour les valeurs courantes de p on obtient les chiffres suivants :

p	=	0,75	: x	= 2
p	=	0,80	: x	= 3
p	=	0,85	: x	= 4
p	=	0,90	: x	= 5
P	=	0,925	: x	= 7
p	=	0,95	: x	= 11

Par exemple, si l'on puise dans un stock très grand d'appareils électriques et si 90 % d'entre eux sont en état de marche, on en tirera en moyenne 2,95 par opération si on les teste 5 par 5, alors qu'un contrôle unité par unité abaisserait ce rendement à 0,9, soit à plus *de trois fois moins.*

Le problème est plus compliqué quand ce sont les objets porteurs de caractère M que l'on cherche à extraire. La méthode que nous proposons n'est qu'asymptotiquement optimale pour des valeurs de p très voisines de 1, c'est-à-dire quand les objets à extraire sont rares.

On peut alors procéder de la manière suivante : on prend un groupe de x objets où x est choisi de telle sorte que p soit le plus voisin possible de 0,5125 et on le teste en bloc. Si le résultat est négatif, on recommence sur un autre lot, puis sur un troisième, etc..., jusqu'à ce que le résultat soit positif. On est alors sûr que ce dernier groupe de x objets en contient au moins un du type marqué. Pour déterminer celui-ci on procède par dichotomies successives en prenant des groupes de $x/2$, $x/4$, $x/8$..., ou plutôt de l'entier le plus voisin de $x/2$, $x/4$, $x/8$... Il est clair que si q est voisin de 1 et si x est grand, on obtiendra en moyenne $1 - q^x$ objets pour un total de $1 + (1 - q^x) \log_2 x$ opérations d'où, par un calcul élémentaire, la condition que nous avons indiquée plus haut relativement à la valeur optimale de x pour p donné à l'avance.

Dans l'exemple précédent si l'on essayait cette fois de trouver des appareils électriques défectueux on les testerait 6 par 6 conformément au schéma opératoire de la figure 4. La méthode permet d'extraire une pièce défectueuse pour 5 essais en moyenne, alors qu'un contrôle pièce par pièce exigerait 2 fois plus d'opérations.

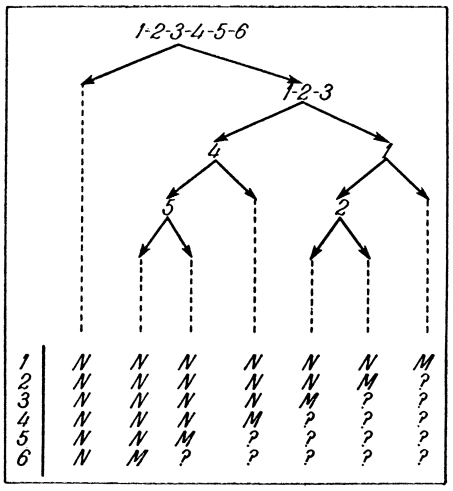

Fig. 4

CONCLUSIONS

S'il nous était permis d'ajouter quelques remarques de nature technique sur les méthodes que nous venons de décrire ce serait pour souligner l'unité qui existe entre les trois problèmes traités : *diagnostic, estimation, triage.* Cette unité qui a des raisons de nature physique profonde tient au fait que dans chaque cas il est possible de construire une expression mathématique appropriée mesurant en quelque sorte la valeur, l'efficacité de chaque observation élémentaire : dans le premier cas cette expression est l'information au sens Wiener-Shannon (7, 11) qui a les dimensions d'une entropie et qui mesure notre ignorance de la nature exacte de chacun des objets. Dans le second cas c'est l'information au sens de R. A. Fisher (2) qui a les dimensions de l'inverse d'une variance et qui mesure notre incertitude sur la vraie valeur de p. Dans le troisième cas, enfin, il s'agit d'un type nouveau d'information que nous avons défini ailleurs (6) et qui mesure d'une certaine manière la richesse de l'échantillon en la catégorie d'objets que l'on cherche à extraire. Dans les trois cas cette information intervient dans une relation d'incertitude étroitement apparentée aux relations d'incertitude de la physique théorique (3) et permet le calcul d'une limite supérieure de l'efficacité de toute méthode concevable : il faut au moins x (p log p + q log q) observations pour diagnostiquer x objets, la variance d'une estimation utilisant x objets ne saurait être inférieure à $xp(1 - p)$ etc... Dans les trois cas selon une démarche extrêmement générale elle aussi, la solution pratique [la « stratégie » dans un sens très voisin de celui de von Neumann (10)] est obtenue en maximisant (compte tenu des particularités de la technique expérimentale) à chaque nouvelle observation la quantité d'information spécifique qu'elle peut apporter.

Enfin dans tous les cas ces « informations » découlent d'une équation fonctionnelle qui les fait apparaître ainsi que nous l'avons montré ailleurs (7) comme les cas les plus simples d'un certain type de fonctions contenant

également l'indice dit « de chi-carré » des statisticiens et même l'expression de certaines distributions de la physique quantique.

BIBLIOGRAPHIE

(1) R. Dorfman. — *Ann. Math. Stat.*, 1943, 14, p. 436.
(2) R. A. Fisher. — *Proc. Roy. Soc.*, 1934, 146, p. 1.
(3) D. Gabor. — *Journ. Ind. Elect. Eng.*, 1946, 93 (111), p. 429.
(4) J. Riguet. — Les relations binaires. *Thèse Paris*, 1951.
(5) M. P. Schützenberger. — *Semaine des Hôpitaux*, 1949, n° 60, p. 2562.
(6) M. P. Schützenberger. — *Comptes rendus Ac. Sc.*, 1951, (232), p. 525.
(7) M. P. Schützenberger. — *Comptes rendus Ac. Sc.*, 1951 (232), p. 1085.
(8) C. Shannon. — *The BELL system Techn. J.*, 1948, 27, p. 379.
(9) J. Ville et M. P. Schützenberger. — *Comptes rendus Ac. Sc.*, 1951 (232), p. 206.
(10) J. von Neuman and O. Morgenstern. — Theory of games and economic behaviour, 1948.
(11) N. Wiener. — Cybernetics, 1948, p. 47.

(*Travail du Centre de Génétique de l'Hôpital St-Louis. Subventionné par l'Institut National d'Hygiène et le Fonds d'Etudes de la Société Médicale des Hôpitaux de Paris*).