PUBLICATIONS DE L'INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS

MÉMOIRES ET CONFÉRENCES SUR LE CALCUL DES PROBABILITÉS, LA STATISTIQUE THÉORIQUE ET APPLIQUÉE, L'ÉCONOMÉTRIE

M. P. SCHÜTZENBERGER

CONTRIBUTION AUX APPLICATIONS STATISTIQUES DE LA THÉORIE DE L'INFORMATION

VOL. III - FASCICULES 1-2 - 1954

PARIS

INSTITUT HENRI POINCARÉ

PUBLICATIONS DE L'INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS

MÉMOIRES ET CONFÉRENCES SUR LE CALCUL DES PROBABILITÉS,
LA STATISTIQUE THÉORIQUE ET APPLIQUÉE, L'ÉCONOMÉTRIE

VOL. III

PARIS
INSTITUT HENRI POINCARÉ
11, Rue Pierre Curie

PUBLICATIONS DE L'INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS
Volume 3 FASCICULES 1 et 2 1954

CONTRIBUTION
AUX
APPLICATIONS STATISTIQUES
DE LA
THÉORIE DE L'INFORMATION

par

M. P. SCHÜTZENBERGER

INTRODUCTION

La variété des outils mathématiques auxquels le calcul des probabilités et la statistique doivent avoir recours augmente chaque jour ; à la théorie des fonctions de variable réelle, qui constituait l'essentiel du bagage analytique nécessaire à leur étude dans le premier quart de ce siècle, se sont ajoutés depuis, le calcul matriciel, la théorie de la transformation de FOURIER, celles des équations aux dérivées partielles, pour ne citer que des exemples classiques.

En retour, sous la pression d'une demande technique sans cesse accrue, qui s'étend de la psychologie à la production industrielle de masse, et suivant les lignes de force de son développement propre, le calcul des probabilités a enrichi de problèmes nouveaux et de solutions originales les disciplines auxquelles ces emprunts avaient été effectués.

L'algèbre n'a pas fait exception à cette tendance comme en témoignent les travaux de Mr. le Pr. FRECHET sur les systèmes d'évènements compatibles et dépendants, qui sont des algèbres de BOOLE, aussi bien que des applications mineures telles que l'emploi des propriétés des corps algébriques pour la solution de certains problèmes de planification des expériences. En particulier, la théorie des treillis semble prédestinée à tenir un rôle important dans de nombreux chapitres du calcul des probabilités.

Le présent travail espère être une illustration nouvelle de ces possibilités et nous avons tenu à conduire l'exposé depuis les fondements abstraits les plus généraux, les relations d'équivalence et d'ordre, jusqu'aux applications à la sérologie ou la génétique mendelienne.

Ce programme déjà trop ambitieux eut sans doute été irréalisable si nous avions dû maintenir à tous les niveaux le degré maximum de généralité. Nous avons donc sacrifié délibérément tout ce qui ne relevait pas des méthodes de la théorie des ensembles finis. Celles-ci d'ailleurs suffisaient pour aborder les problèmes précis que nous avions en vue et on verra, en outre que nombre d'énoncés ont été établis de telle sorte qu'ils s'appliqueraient sans modification à des cas plus généraux. La rançon en serait une formulation plus lourde des démonstrations risquant peut-être parfois d'obscurcir la nature véritable du lien logique qui tresse le raisonnement.

Venons en maintenant à l'objet de cette étude ; la théorie de l'information. Ici aussi, il s'agit d'une discipline nouvelle dont les frontières d'aujourd'hui indiquant la direction des progrès futurs plus que les limites reconnues des concepts et des méthodes. Il était tentant d'en chercher une unification même restreinte et même provisoire et d'appliquer ses techniques à une classe de problèmes qui put mettre en évidence la diversité et l'unité des notions qui se trouvaient ainsi regroupées;

Enfin, nous avons essayé de faire oeuvre pratique. Nous voudrions espérer que les statisticiens, aux prises comme nous avec les réalités parfois si peu accomodantes de la recherche industrielle ou scientifique, trouveront dans ce travail, non seulement des suggestions, mais aussi des résultats directement utilisables.

En fonction de ces objectifs, l'exposé a été scindé en trois parties dont chacune forme un tout ;

I. Dans la première partie, on rappelle certaines notions algébriques qui réapparaîtront constamment par la suite; relations d'ordre, treillis distributifs, enfin treillis de partitions qui généralisent très directement la structure déjà bien connue de treillis de toutes les relations d'équivalence d'un ensemble.

Il s'agit là d'un exposé schématique de résultats presque tous classiques où l'accent est mis sur les concepts spécialement efficaces dans le cas fini tels que les éléments irréductibles ou le groupoïde des intervalles, par exemple.

Les auteurs à notre sens, donnent un sens trop restrictif au terme de valuation des éléments d'un treillis. Nous avons tenté de replacer cette notion dans une perspective plus naturelle, comme fonction numérique associée à la fois aux éléments d'une structure d'ordre et au groupoïde des intervalles que définit celle-ci. La définition proposée permet d'obtenir formellement toutes les valuations d'un treillis de partitions fini.

Trois exercices (calcul de l'expression formelle des cumulants en fonction des moments, - fonction génératrice des probabilités de certaines "statistiques d'ordre" - forme générale des distributions de la statistique quantique) illustrent les possibilités de ces méthodes dans le calcul des probabilités.

II. Dans deux théories indépendantes, avaient déjà été définis des êtres mathématiques nommés "information" ; l'information de FISHER en théorie de l'estimation et l'information de SHANNON en théorie des communications. Dans cette seconde partie on s'efforce de justifier une définition générale des "informations" par une analyse de propriétés que doit a priori posséder un tel être pour généraliser valablement les deux quantités qui viennent d'être évoquées. On aboutit à une classe spéciale de valuations du treillis des partitions des états observables d'une aléatoire. Un opérateur linéaire reste disponible dans cette expression. Pour un choix convenable de celui-ci, on retrouve les deux types spéciaux que l'on voulait unifier mais, aussi, à côté d'êtres apparemment nouveaux tels que les informations de tri, une quantité que nous nous proposons d'appeler

"Information de Wald" en raison du rôle capital que cet auteur lui a fait jouer tacitement dans l'analyse séquentielle.

De la systématisation des résultats découle, comme d'usage, une grande économie de démonstrations et des possibilités d'extensions de concepts tels que l'exhaustivité de DARMOIS, l'additivité, etc ...

Enfin, d'autres grandeurs familières du calcul des probabilités (cumulants, chi-carré) présentent avec les informations des analogies qui sont brièvement discutées.

Comme dans la première partie, l'originalité ne se trouve que dans le mode d'exposition qui permet de rassembler des théories dont le lien n'avait pu être interprété jusqu'ici, que de façon assez superficielle. Cette confrontation fait apparaître des problèmes non encore résolus, d'un intérêt certain, aussi bien du point de vue physique que de celui des mathématiques pures.

III. Les "méthodes de groupage" qui font l'objet de la troisième partie n'ont au contraire presque jamais été étudiées. On construit un modèle mathématique très simple qui représente certains systèmes d'objets tels qu'une seule observation soit éventuellement susceptible de caractériser entièrement plusieurs d'entre eux ; le type en est l'observation du produit d'une série de nombres qui permet d'apprendre, soit qu'ils sont tous différents de zéro, soit que l'un d'eux au moins est nul. Diverses situations concrètes peuvent s'y ramener, semble-t-il, et l'on énonce avec l'aide des concepts algébriques décrits dans la première partie quelques propriétés générales de ce modèle.

L'application des grandes méthodes de la statistique mathématique (test, estimation) conjointement avec les notions développées dans la seconde partie livre une gamme d'exercice dans lesquels se manifestent l'utilité et le sens des diverses informations.

A ce propos, on introduit comme substitut des solutions optimales - trop souvent inaccessibles pour des raisons de complexité combinatoire - la notion de "tactique localement optimale" qui semble nouvelle et qui dépasse d'ailleurs très largement le cadre de ce travail.

Références et bibliographie. Le matériel présenté dans la première partie est le plus souvent classique et presque toutes les indications utiles se trouvent dans l'ouvrage fondamental de G. BIRKHOFF "Theory of Lattices", ou en français dans la "Théorie des Structures" de GLIVENKO. Nous avons jugé superflu d'en reproduire la bibliographie et nous sommes contentés d'introduire dans le cours du texte les références supplémentaires nécessitées par les divers points particuliers étrangers à la théorie des treillis.

Il en est de même pour la troisième partie qui ne fait appel qu'aux éléments les plus familiers de la statistique mathématique quand il ne s'agit pas de notions déjà discutées dans la seconde partie, la seule pour laquelle soit apparu nécessaire de compiler une bibliographie systématique.

Nous exprimons notre reconnaissance respectueuse à Monsieur le Professeur Maurice FRECHET qui a bien voulu nous faire l'honneur de présider la soutenance de cette thèse, à Monsieur le Doyen Albert CHATELET qui fût notre premier Maître de cette Faculté, et à Monsieur le Professeur Georges DARMOIS qui nous a constamment guidé de ses conseils et de ses encouragements.

On verra par ailleurs dans le cours de ce mémoire, la part essentielle qui revient à l'enseignement de Monsieur A. CHATELET et aux travaux de Monsieur M. FRECHET et de Monsieur G. DARMOIS.

Nous ne saurions nons plus ne pas exprimer notre respectueuse gratitude à Monsieur le Professeur Raymond TURPIN qui depuis de nombreuses années nous a associé à ses recherches progénésiques dans le Centre qu'il a fondé à l'Hôpital Saint-Louis et à Monsieur le Professeur Pierre GAVAUDAN qui nous a fait participer à ses travaux de physiologie et dont l'amitié et l'exemple nous ont été le plus précieux des encouragements.

Enfin ce travail n'aurait pu être accompli sans le soutien de l'Institut National d'Hygiène, de son Directeur Monsieur le Professeur Louis BUGNARD, de Monsieur Pierre DENOIX, Chirurgien des Hôpitaux de Paris, et l'aide généreuse du Fonds d'étude de la Société médicale des Hôpitaux de Paris; qu'ils reçoivent, ici le témoignage de notre reconnaissance.

PREMIÈRE PARTIE

PRÉLIMINAIRES ALGÉBRIQUES

I. - RELATIONS D'ÉQUIVALENCE ET RELATIONS D'ORDRE

Etant donné un ensemble E d'éléments a, b, ... une relation d'équivalence (1) sera par définition une relation entre éléments de E, notée aRb, satisfaisant aux trois conditions suivantes :

1° - Réflexivité : pour tout a ; a R a.

2° - Symétrie : pour tout a et b ; si aRb, alors : bRa.

3° - Transitivité : pour tout a, b et c : si aRb et bRc, alors : aRc.

On démontre que R induit sur E une partition en classes d'équivalence X, Y ..... Z. Celles-ci forment un système de sous ensembles de E disjoints, (c'est-à-dire deux à deux sans élément commun) tels que tout élément de E soit contenu dans l'un d'eux et enfin satisfaisant à cette condition que aRb, si et seulement si a et b sont membres du même sous-ensemble.

De deux relations d'équivalence R_1 et R_2 sur un ensemble E R_1 sera dit "plus fine" que R_2 si et seulement si, pour tout a et tout b, aR_1b entraîne a R_2 b.

Dans ce cas, les classes X, Y, Z ..., de R_2 sont elles-mêmes partitionnées en les classes X_1, X_2, ... X_k, Y_1, Y_2,.. Y_k'', Z .... Si A est une classe de R_1 et B une classe de R_2, ou bien A et B sont sans élément commun ou bien A est toute entière contenue dans B.

Enfin, on fera usage de la notion de fermeture transitive $\bar{R}$ d'une relation quelconque R de E; par définition $\bar{R}$ sera la plus fine des relations d'équivalence telle que pour tout a et b si aRb, alors $a\bar{R}b$. Cette définition est justifiée par le fait qu'on peut prouver que $\bar{R}$ est unique pour un E et un R quelconque.

Les relations d'ordre sont des relations :

1° - Réflexives

2° - Transitives

3° - Acycliques c'est-à-dire telles que l'on ait jamais : aRb et bRa, sauf si a = b.

On voit donc qu'elles s'opposent aux relations d'équivalence par le fait que l'on a remplacé la condition de symétrie par son contraire qui est l'acyclicité.

---

(1) Pour un exposé complet de ces notions sur les relations, on consultera avec fruit le travail d'ensemble de : J.RIGUET (1951) - Théorie des relations binaires - (Thèse Paris).

Les exemples les plus familiers de relation d'ordre sont les relations "plus petit ou égal" entre grandeurs réelles et la relation "être contenu dans" entre sous-ensembles et ensembles. Il est commode d'emprunter à ces cas particuliers les notations $\leqslant$ ou $\subset$ pour représenter une relation d'ordre quelconque.

On appellera "ordre dual" d'une relation d'ordre R la relation $\overset{\circ}{R}$ définie par $a\overset{\circ}{R}b$ si, et seulement si, bRa. On vérifie sans peine que $\overset{\circ}{R}$ est bien une relation d'ordre en même temps que R.

Dérivant de la relation d'ordre, on fera usage dans les structures finies de la relation de consécutivité. b sera dit "consécutif" à a'' ou "couvrir a'' si $a \subset b$ et s'il n'existe aucun c différent de a et de b tel que $a \subset c \subset b$.

Si $a \subset b$, une chaîne complète entre a et b sera une suite d'éléments c_i ($0 \leqslant i \leqslant k$) tels que $a = c_0$; $b = c_k$ et qu'enfin pour tout i, c_{i+1} couvre c_i. On appellera k la longueur de la chaîne.

Correspondant en quelque sorte à la notion de classe d'équivalence, nous trouverons ici la notion d'intervalles. Par définition l'intervalle (a, b) sera le sous-ensemble de E formé par les éléments x tels que $a \subset x$ et $x \subset b$. L'expression "intervalle (a, b)" n'aura donc de sens que si $a \subset b$. Si b couvre a, l'intervalle (a,b) sera réduit à ces deux éléments, ce sera donc un intervalle minimum. Les intervalles dégénérés qu'il est commode de considérer formellement seront les intervalles du type (x, x).

Il sera utile de considérer sur les intervalles l'opération de composition associative mais non partout définie suivante :

$$(a,b)\ (c,d) \begin{cases} = (a,d) \text{ si } a \subset b \ ; \ b = c \text{ et } c \subset d \\ \\ \text{non défini dans tous les autres cas.} \end{cases}$$

E étant un ensemble ordonné, l'ensemble de ses intervalles munis de cette loi de composition sera appelé le "groupoïde des intervalles de E " . Les intervalles dégénérés en sont les idempotents puisque (a a) (a a) = (a a) ; Cette définition, généralise et restreint en même temps la notion déjà classique de "groupoïde de BRANDT".

Enfin, étant donnée une relation d'ordre R et une relation d'équivalence R' entre éléments dont les classes sont X, Y ..... toutes deux définies sur le même ensemble, nous dirons que R' est compatible avec R si la relation R entre classes définies par : $X \, R \, Y$ s'il existe a $\in$ X et b $\in$ Y telles que aRb est une relation d'ordre entre ces mêmes classes n'entraînant aucune autre relation de la forme $X\bar{R}Y$ et $Y\bar{R}X$ (ce qui pourrait être le cas si l'on avait à la fois par exemple :

$$a \ , \ a' \in X \text{ et } b, \ b' \in Y \text{ et } aRb \text{ et } b' \ R \ a' \ ).$$

Nous appellerons cette nouvelle structure la structure quotient $E|R'$ de E par R'. On peut démontrer que si R' n'était pas compatible avec R, il existerait une relation d'équivalence R'' unique, moins fine que R', compatible avec R et telle que toute autre relation d'équivalence satisfaisant à ces conditions soit moins fine qu'elle. Nous pouvons donc sans ambiguïté aucune associer à une relation d'ordre R sur E une relation d'ordre $\bar{R}$ sur $E|R''$.

On fera attention au fait qu'à une structure d'ensemble ordonnée correspondent deux familles de relation d'équivalence :

les relations entre éléments dont nous venons de dire un mot et qui induisent des structures quotient de l'ensemble.

les relations entre intervalles qui induisent au contraire des structures quotient du groupoïde des intervalles.

La plus simple d'entre ces dernières est la relation d'isomorphie R définie par :

$(a, b)\ R\ (a', b')$ si et seulement si, il existe une correspondance biunivoque $\longleftrightarrow$ entre les éléments c_i de l'intervalle (a, b) et c'_i de (a', b') telle que si $c_i \longleftrightarrow c'_i$ et $c_2 \longleftrightarrow c'_2$ $c_1 \leqslant C_2$ soit équivalent à $C'_1 \leqslant C'_2$.

Pour terminer, nous rappellerons que si E et E' sont deux ensembles munis chacun d'une relation d'ordre R et R', le produit direct E X E' est muni lui aussi de façon naturelle de la relation d'ordre R'' définie par :

$(a \times a')\ R''\ (b \times b')$ si et seulement si, $a\ R\ b$ et $a'\ R'\ b'$.

Naturellement tous les intervalles $((axa'), (bxa'))$ où a et b sont fixés et où a' parcourt E' sont isomorphes entre eux et isomorphes à E'

II. - TREILLIS

Nous allons maintenant isoler parmi les relations d'ordre une famille satisfaisant à des conditions beaucoup plus strictes qui nous permettront de traiter l'ensemble E comme une structure algébrique.

Définition : Une relation d'ordre $\subset$ sur E est une relation d'ordre latticiel si et seulement si pour toute famille F formée d'éléments a, b ...., c de E il existe un élément x et un élément y tels que :

1° Pour tout $a - F$: $x \subset a \subset y$

2° Pour toute autre paire z, u satisfaisant à la condition précédente: $z \subset x \subset y \subset u$.

Un exemple familier d'ordre latticiel est celui des ensembles convexes du plan réel ordonnés par inclusion; dans ce cas :

y est le plus petit ensemble convexe contenant tous les éléments de F;

x est la partie commune à tous les éléments de F : (celle-ci pouvant d'ailleurs être vide).

On appelle x l'intersection des éléments de F;
y la réunion des éléments de F;

et on note : $x = a \cap b \ldots\ldots\ldots\ldots\ldots\ldots \cap c$

$y = a \cup b \ldots\ldots\ldots\ldots\ldots\ldots \cup c$

ou encore, plus fréquemment, quand aucune confusion n'est à craindre :

$x = a\ b \ldots\ldots\ c$

$y = a + b \ldots\ldots + c$

On peut montrer que ces deux opérations $\cap$ et $\cup$ satisfont aux lois suivantes :

I° - Idempotence : pour tout a : $a \cap a = a = a \cup a$

2° - Associativité : pour tout a, b, c :

$(a \cap b) \cap c = a \cap (b \cap c)$; $(a \cup b) \cup c = a \cup (b \cup c)$

3° - Commutativité : pour tout a et b :

$(a \cap b) = b \cap a$; $a \cup b = b \cup a$

4° Absorption : pour tout a et b :

$a \cap (a \cup b) = a = a \cup (a \cap b)$

Inversement, en prenant ces égalités algébriques comme axiomes on pourrait reconstruire la relation d'ordre par la définition suivante :

$a \subset b$ si et seulement si $a = a \cap b$

ou bien $a \subset b$ si et seulement si $b = a \cup b$

Un ensemble muni d'une relation d'ordre latticiel sera nommé un treillis.

Formellement il sera toujours possible d'adjoindre à un treillis un "plus petit" et "un plus grand" élément, notés, par exemple 0 et 1 tels que l'on ait pour tout a :

$0 \cap a = 0$; $0 \cup a = a = a \cap 1$; $1 \cup a = 1$.

Si le treillis est fini, 0 et 1 peuvent être identifiés respectivement à l'intersection et à la réunion de tous les éléments du treillis.

Dans tout treillis existe une dualité canonique obtenue en permutant les opérations $\cap$ et $\cup$ qui, comme on le voit, entrent de manière parfaitement symétrique dans les définitions. Autrement dit, R et $\tilde{R}$ simultanément sont ou ne sont pas des ordres latticiels.

On observera que si l'ensemble ordonné E est fini et s'il contient un plus grand élément, il suffit pour qu'il soit un treillis qu'il possède une intersection : en effet, pour toute famille F (a, b, ..c), il existe au moins un x tel que $a \in F$ entraîne $a \subseteq x$. S'il en existait plusieurs, constituant une famille (x, y ... z) l'intersection $x_0 = x \cap y \cap \ldots \cap z$ serait la réunion $a \cup b \cup \ldots \cup c$ car :

I° - x satisfaisant à $a \subset F$, ceci entraîne $a \subset x_0$ puisque $a \subset x$ est équivalent à: $a \cap x = a$ et par conséquent si

$a \cap x = a \cap y = a$ on a :

$a = (a \cap x) \cap (a \cap y) = a \cap (x \cap y)$ et donc :

$a \subset x \cap y$

2° - x_0 est minimal parmi les éléments ayant cette propriété de par la définition même de l'intersection.

Le même raisonnement et son extension par dualité permettent de démontrer le résultat suivant :

Tout intervalle d'un treillis est lui-même un treillis.

En effet, on vient de voir que si

$a \subset x \subset b$ et $a \subset x' \subset b$ on a aussi $a \subset x \cap x' \subset x \cup x' \subset b$

Il est trivial de remarquer qu'un sous-treillis c'est-à-dire un sous-ensemble de E fermé pour les deux opérations, réunion et intersection n'est pas nécessairement un intervalle.

TREILLIS LIBRES, ÉLÉMENTS IRRÉDUCTIBLES

La structure de treillis que nous venons de caractériser est encore trop générale et dans les applications on ne se sert le plus souvent que de treillis satisfaisants des axiomes supplémentaires. Ceux-ci peuvent être de plusieurs types dont le plus important est constitué par les "lois universelles" qui postulent une relation algébrique nouvelle reliant les deux opérations $\cap$ et $\cup$.

Par exemple, nous étudierons plus loin les treillis distributifs, c'est-à-dire ceux où entre trois éléments quelconques on a toujours :

$$a \cap (b \cup c) = (a \cap b) \cup (a \cap c) .$$

Appelons "mot latticiel" toute expression formelle bâtie à partir d'une famille de symboles x , y .... z et des deux opérations $\cap$ et $\cup$.

Nous conviendrons que deux mots P_1 (x , y ...z) et P_2 (x, y ...z) sont équivalents si l'on peut les ramener l'un à l'autre par l'application successive des axiomes définissant la relation d'ordre latticiel.

Par exemple (nous employons ici la notation abrégée) les deux mots :

$$P_1 = a\ b \text{ et } P_2 = a(ab + bc)$$

sont équivalents puisque l'on a :

$$a\ b \leqslant b \ ;\ b\ c \leqslant b \ ;\ \text{donc: } a\ b + bc \leqslant b \quad \text{donc :}$$

$$a\ (\ a\ b + a\ c\ ) \leqslant a\ b$$

et d'autre part : $a\ b \leqslant a$; $a\ b \leqslant a\ b + b\ c$;

donc aussi: $a\ b \leqslant a\ (\ a\ b + b\ c\ )$ donc , enfin , l'équivalence cherchée.

L'ensemble de ces mots ou plutôt de ces classes de mots constitue un treillis que l'on appelle le "treillis libre" à n générateurs si n symboles x y ...z sont intervenus dans sa construction.

Soit maintenant $\mathcal{U}$ une loi donnée par l'égalité postulée de deux mots P_1 et P_2 portant sur k symboles et soit T un certain treillis. P_1 et P_2 définissent chacun une fonction dans T des k-uples d'éléments de T. Deux cas sont alors possibles.

- ou bien ces deux fonctions sont toujours égales, auquel cas nous dirons que T satisfait à $\mathcal{U}$ ou encore que T est un $\mathcal{U}$-treillis.

- ou bien au moins pour un k-uple $(x, y, \ldots z)$ on a $P(x_0, y_0, \ldots z_0) \neq t \neq t' = P_2(x_0, y_0 \ldots z_0)$.

Définissons alors une relation d'équivalence $\mathcal{U}$ entre éléments de T par le fait que $t \mathcal{U} t'$ si et seulement si il est possible de les exprimer par deux mots réductibles l'un à l'autre par l'application répétée des axiomes et de l'égalité $P_1 = P_2$.

Il est clair que $T|\mathcal{U}$ est un treillis que l'on appellera le treillis quotient de T par la loi $\mathcal{U}$.

En particulier, si T est le treillis libre à n générateurs $T|\mathcal{U}$ sera le $\mathcal{U}$-treillis libre à n générateurs.

C'est ici qu'intervient une notion nouvelle qui est particulièrement utile dans l'étude de ces structures : celle d'élément irréductible.

Nous dirons que x est un élément $\cup$ irréductible si dans toute représentation $x = y \cup z$; on a $y = x$ ou $z = x$; il revient au même, si le treillis est fini, de dire que x ne couvre qu'un seul élément qu'on notera x_0 .

On définirait de même les éléments $\cap$irréductibles par dualité canonique et l'on peut énoncer le :

> Théorème : Tout treillis est décrit de manière univoque par la seule donnée des relations d'ordre qui existent entre ses éléments $\cup$irréductibles et ses éléments $\cap$ irréductibles.

En effet, tout élément de treillis peut être caractérisé de manière univoque à la fois comme la réunion de tous les éléments $\cup$irréductibles qui sont plus petits que lui ou comme l'intersection de tous les $\cap$irréductibles qui sont plus grands que lui.

Comme d'habitude dans les structures algébriques, on dira que la relation d'équivalence R entre éléments est compatible avec la structure du treillis T si pour tout a et b, a' et b' a R b et a' R' b' entraînent :

$$(a \cap b) R (a' \cap b') \text{ et } (a \cup b) R (a' \cup b').$$

On peut alors énoncer :

> Toute relation d'équivalence R compatible avec T fini est décrite de manière univoque par la donnée des intervalles (x_0, x) (où x est un élément $\cup$ irréductible) qui deviennent des intervalles dégénérés dans le treillis $T|R$.

L'énoncé dual vaut naturellement pour les éléments $\cap$ irréductibles.

TREILLIS DISTRIBUTIFS

Ce sont des treillis où entre trois éléments quelconques, on a toujours l'égalité :

$$(D) : a(b + c) = ab + ac$$

A titre d'exemple, montrons que cet axiome est équivalent à sa forme duale :

$$(\overset{\cup}{D}) : (a + b)(a + c) = a + bc$$

En effet, si D est vrai :

$$(a+b)(a+c) = (a+b)a + (a+b)c = a + ac + bc = a + bc$$

Il en résulte que dans le treillis distributif libre à n générateurs T_n les seuls $\cap$-irréductibles sont les mots $(x + y + \ldots + z)$ et les $\cup$ irréductibles les mots $xy\ldots z$ (où x, y...z constituent une partie quelconque de l'ensemble des générateurs) puisque l'on peut éliminer toutes les parenthèses grâce à D ou à $\overset{\cup}{D}$.

Un élément de T est donc caractérisé de deux manières distinctes comme une famille de parties de l'ensemble des générateurs entre lesquels n'existe aucune relation d'inclusion.

Par exemple :

$ab c + a b d + c d$ est un élément de T_4 et l'élément dual associé serait :

$$(a+b+c)(a+b+d)(c+d) = ac + ad + bc + bd + cd$$

Les exemples de treillis distributifs se rencontrent dans presque toutes les parties des mathématiques. Le plus classique est peut être celui où la relation c représente la relation de divisibilité entre entiers. Les opérations $\cup$ et $\cap$ sont alors les opérations de P.G.C.D. et de P.P.C.M.

Cependant, nous pouvons encore particulariser deux types remarquables à l'intérieur des treillis distributifs.

1° - Les chaînes (ou "ensembles totalement ordonnés") dans lesquelles pour deux éléments quelconques x et y on a toujours soit $x \subset y$ soit $y \subset x$ et dont il est superflu de donner des exemples.

2° - les algèbres de BOOLE dans lesquelles on postule à côté de $\cap$ et $\cup$, l'existence d'une opération unaire biunivoque partout définie la "complémentation" notée $-$ et caractérisée par:

$$\bar{a} \cap a = 0 \quad \text{et} \quad a \cup \bar{a} = 1$$

Il n'est pas nécessaire non plus d'insister sur les algèbres de BOOLE qui sont familières aux probabilistes sous la forme de l'algèbre des propositions et où les symboles $\cap$ et $\cup$ et $^-$ sont interprétés respectivement comme la conjonction (&), la disjonction ($\vee$) et la négation.

On rappellera simplement que dans une algèbre de BOOLE, les seuls irréductibles ("les atomes") couvrent 0 (et dualement les seuls irréductibles sont couverts par 1).

Il est classique aussi que si E est fini l'ensemble de ses parties ordonnées par inclusion constitue une algèbre de BOOLE et l'on connaît l'importance dans la théorie de la mesure, de ces notions étendues convenablement au cas infini.

TREILLIS MODULAIRES

Il existe des cas très importants (idéaux, sous-groupes distingués) où la loi distributive n'est pas vérifiée mais où pourtant existe la loi plus faible suivante : loi modulaire ou loi de DEDEKIND :

Pour tout a, b et c :

$$(M) : a \cap (b \cup (a \cap c)) = (a \cap b) \cup (a \cap c)$$

Par exemple, si l'on considère le treillis formé par les variétés linéaires d'un espace projectif; avec $a \cap b$ comme intersection de a et de b et $a \cup b$ comme "plus petite variété linéaire contenant a et b", on voit que si a_1 est une droite contenue dans le même plan que les points distincts b_1 et c_1, on a : $b_1 \cup c_1$ = une droite et, donc, $a \cap (b \cup c)$ = un point mais (sauf si b_1 ou c_1 sont sur a_1) :

$a_1 \cap b_1 = a_1 \cap c_1 = (a_1 \cap b_1) \cup (a_1 \cap c_1)$ = l'ensemble vide $\emptyset$.

Par contre, on peut démontrer que l'égalité (M) est toujours vraie dans cette structure.

Nous dirons qu'un treillis est "modulaire" s'il satisfait à la loi universelle (M). Comme (M) est identiquement vrai quand (D) l'est elle-même, les treillis distributifs sont à fortiori des treillis modulaires;

A titre d'exemple, nous allons montrer que (M) est équivalent à sa formule duale :

$$(\breve{M}) = a + b (a + c) = (a + b) (a + c)$$

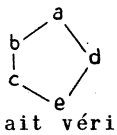
en effet d'après (M) on a :

$$(a + b) (a + c) = (a(a + c) + b) (a + c)$$
$$= a (a + c) + b (a + c) = a + b (a + c)$$

Au contraire de ce que nous avons vu pour les treillis distributifs, il n'est pas connu de forme canonique pour les éléments irréductibles de treillis modulaire libre à k générateurs sauf dans les seuls cas où celui-ci est fini c'est-à-dire quand k = 2 ou 3 .

Les propriétés essentielles des treillis modulaires découlent du résultat suivant :

Dans un treillis modulaire, il n'existe aucun 5-uple d'éléments distincts formant un sous-treillis du type représenté par le schéma suivant où les traits verticaux indiquent des relations de consécutivité :



En effet si la loi M était vérifiée on aurait :

$$b = b\,a = b (c + d) = b (b\,c + d) = b\,c + b\,d = c + e = c$$

La réciproque est vraie et il s'en déduit :

Dans un treillis modulaire fini, toutes les chaînes qui joignent deux éléments ont la même longueur.

Une autre formulation est la suivante qui est connue sous le nom de Lemme de BIRKHOFF.

Si toutes les chaînes d'un treillis sont de longueur finie, la condition nécessaire et suffisante pour que celui-ci soit modulaire est que pour tout x et y, l'assertion : "x et y couvrent $x \cap y$" soit équivalente à : "$x \cup y$ couvre y et x".

Enfin on exprimera plus généralement le théorème suivant ; en désignant par R_0 la fermeture de la relation R entre intervalles définie par :

$$(x , x \cap y) R (x \cup y , y)$$

La condition nécessaire et suffisante pour que le treillis soit modulaire est que R_0 soit une relation d'isomorphisme.

On appelle R_0 la "relation d'équivalence projective" et on voit que dans le contre exemple précédent on avait à la fois $(a , c) R_0 (d , e)$ et $(a,b) R_0 (d, e)$ et que par conséquent $(a , c) R_0 (a, b)$.

Montrons au contraire que R_0 est bien un isomorphisme dans les treillis modulaires. En effet, si v et v' appartiennent à $(x \cap y , x)$, on a la correspondance : $v \to y \cup v$.

$$v' \to y \cup v'$$

$$v \cup v' \to y \cup (v \cup v') = (y \cup v) \cup (y \cup v')$$

et pour tout w, w' dans $(y , y \cup x)$:

$$w \to x \cap w \; ; \; w' \to x \cap w \; ; \; w \cap w \to x \cap (w \cap w')$$

Enfin d'après la loi modulaire, il existe la propriété involutive suivante qui achève d'établir le résultat :

$$v \to y \cup v \to x \cap (y \cup v) = (x \cap y) \cup (x \cap v) = v$$

III. - TREILLIS DE PARTITION

Les autres variétés particulières de treillis que l'on a l'occasion d'utiliser ne sont pas en général definis par des lois universelles, mais par des considérations assez diverses.

Par exemple, un treillis fini est un arbre quand, quelque soient a et b on a : $a + b = 1$ (1 : le plus grand élément) (sauf naturellement si $a \subset b$ ou $b \subset a$).

Ceci revient à dire que chaque élément (sauf 1) ne couvre qu'un seul élément ou encore que tous les intervalles $(0, a)$ $(a \neq 1)$ sont des chaînes.

Les treillis de partition que nous étudierons maintenant sont au contraire définis à partir d'une autre structure. Ils sont un peu plus généraux que les treillis que l'on considère habituellement sous ce nom et qui ne sont rien d'autre que les

structures duales des treillis de toutes les relations d'équivalence d'un ensemble.

Définition : Etant donné un ensemble E, une famille F de parties e_i de E ($i \in i$) sera dite "base de partition" de l'ensemble si elle satisfait aux conditions suivantes :

1° - F contient E lui-même et l'ensemble vide $\emptyset$

2° - F contient en même temps que tout sous-ensemble de ces parties leur intersection au sens de la théorie des ensembles.

3° - Si $e_j \in F$, et si les ensembles deux à deux disjoints e_j ($j \in J \subset I$) sont contenus dans e_i , il existe au moins un système de e_k ($K \in K(I)$ dans F, deux à deux disjoints, et disjoints des e_j , tels que la réunion (toujours au sens de la théorie des ensembles) des e_j et des e_k soit exactement e_i .

On observera que la condition 3° n'est qu'une forme affaiblie de celle par laquelle on définit habituellement la différence de deux ensembles. Qu'elle ne s'y réduit pas est montré par le contre-exemple suivant :

E = a,b,c ; F = (a,b,c),(a,b),(b,c),(a),(b),(c), et $\emptyset$.

F est bien un treillis mais ne contient pas :

$$(a\ c) = (a\ b\ c) - (b)$$

On voit sans peine que :

toute base de partition de E contient aussi une base de partition pour chacun des ensembles, e_i appartenant à F.

Dans la suite, nous n'aurons presque jamais besoin que des deux cas suivants :

1° - E est un ensemble fini de n éléments. F est l'ensemble de toutes les parties de E, ce qui est le cas considéré habituellement comme nous l'avons dit plus haut.

2°-E est l'ensemble des entiers positifs inférieurs à n, F est l'ensemble des entiers compris entre deux valeurs.

Nous considèrerons désormais que E et F sont donnés une fois pour toutes.

Définition : Une partition $W = (e_1)(e_2)\ldots(e_k)$ de E sera une relation d'équivalence sur E dont les classes (les "composants") e_1 , e_2 .... e_k appartiennent à F.

Le résultat suivant est trivial :

Si W est une partition de E, W définit aussi une partition de chacun des sous-ensembles E_i de E formé par la réunion d'un nombre quelconque de ses composants. On appellera ces partitions les restrictions de W à E_i .

On profitera de ceci pour abréger les notations et écrire par exemple $W = (e_1)(e_2)\ldots\ldots(W')(W'')$ .. si e_1 , e_2 ... etc sont des composants de W et $W^{(1)}$ les restrictions de W à e' e''... etc,

les e_1, e_2,... e', e" ... étant, naturellement disjoints et de réunion totale égale à E.

Nous ordonnerons les partitions par la même relation d'ordre ("plus fin") que les relations d'équivalence et nous démontrerons :

Quelque soit F, cet ordre est un ordre latticiel :

En effet, soient $W = (e_1)\ (e_2)\ \ldots.(e_k)$ et $W' = (e'_1)\ (e'2)\ldots.(e'_k)$. Les ensembles $e_i \cap e'_j$ appartiennent à F, sont deux à deux disjoints et ont pour réunion E. Ceux qui ne se réduisent pas à l'ensemble vide sont les composants d'une partition W" plus fine que W et que W' et moins fine que toute autre ayant la même propriété. Donc il existe une intersection des partitions et comme nous opérons dans un domaine fini, il s'en déduit l'existence d'une réunion.

On remarquera à ce propos que ce treillis L_F ne satisfait à aucune loi universelle particulière dans le cas général. Il est cependant distributif dans le cas correspondant au deuxième exemple donné plus haut.

Enfin si W_2 est plus fine que W_1, la notion de "quotient" W' de W_1 par W_2 aura un sens bien clair : W' sera la partition induite par W_1 sur le quotient E' de l'ensemble E par la relation d'équivalence que définit W_2.

Les principales propriétés du treillis L_F découlent des considérations très simples suivantes :

A tout e_i et à toute partition $W\alpha$ de e_i nous associons un symbole que nous appellerons l'opérateur de partition simple $T\alpha$ et nous convenons d'écrire :

$W\ T\alpha = W'$ chaque fois que :

1° - e_i est un composant de W (sinon on convient que l'écriture précédente n'a pas de sens).

2° - les restrictions de W et de W' à $E-e_i$ sont identiques.

3° - La restriction de W' à e_i est précisément $W\alpha$.

Nous composerons ces opérateurs par une loi notée o et définie par :

$T'' = T\,o\,T'$ si et si seulement il existe W , W' et W'' tels que $W\ T = W'$ et $W'' = W'\ T'$ aient un sens.

Manifestement on ne peut avoir ToT' que dans les deux cas suivants :

1° - T' opère sur un composant de la partition $W\alpha$ sur lequel opérait déjà T . Dans ce cas T' o T n'a pas de sens mais ToT' est aussi un opérateur de partition simple.

2° - T' opère sur un composant e_j de W distinct de e_i . Dans ce cas ToT' et T' o T ont tous les deux un sens et peuvent être considérés comme égaux puisqu'ils conduisent à la même partition W''.

L'ensemble des classes de mot en les T formés avec la loi de composition qui ont un sens, et qui ne sont pas réductibles

l'une à l'autre par la loi de commutativité précédente constituent ce que nous appellerons le groupoïde G des opérateurs de partition de E.

Comme on voit, G est isomorphe du quotient du groupoïde des intervalles de L par la relation d'équivalence $\bar{R}$ entre intervalles W qui est la fermeture des relations suivantes :

1° - R_T : $(W, W') R_T (W'', W''')$ si $W' = W T$ et $W''' = W''T$.

2° - C : $(W W') C (W'' W''')$ s'il existe T et T' tels que $W' = W (ToT')$ et $W''' = W'' (T' o T)$.

Nous appellerons $\bar{R}$ l'équivalence normale et nous justifierons son intérêt par les théorèmes suivants :

1° - Si $(W, W') R_T (W'', W''')$ alors les deux intervalles sont isomorphes.

En effet, ils sont tous les deux isomorphes à un intervalle (W_1, W'_1) sur le treillis des partitions de l'ensemble e_1 sur lequel opère T .

2° - L'intervalle (W, W'') où $W'' = W (T_1 o T_2) = W (T_2 o T_1)$ est isomorphe au produit direct des intervalles: $(W T, W'')$ et $(W T_1 , W'')$.

En effet, toute partition de l'intervalle (W , W'') peut s'écrire $(W_1) (W_2) (W_3)$ ou W_1 est unepartition de l'ensemble e_1 sur lequel opère T_1 , W_2 une partition de e_2 et W_3 une partition de $E - e_1 - e_2$.

Enfin on a :

3° - $\bar{R}$ est une relation d'équivalence plus fine que la relation d'équivalence projective $\bar{R}_o$. Deux intervalles $\bar{R}$ équivalents sont donc isomorphes.

Ce résultat est certainement vrai pour la relation C en vertu du théorème précédent. Montrons qu'il en est de même pour R_T.

Si $(W_1 , W_1') R_1 (W_2 , W_2')$, on peut poser par définition:

$$W_1 = (e_o) (W'_1) , W_2 = (e_o) (W'_2)$$

$$W'_1 = (W_o)(W'_1), W'_2 = (W_o) (W'_2)$$

où W_o est la partition effectuée par T sur e_o.

Considérons l'intervalle $\big((W_1 \cup W_2), (W' \cup W'_2)\big)$. Il est isomorphe à chacun des intervalles (W_1 , W'_1) et (W_2 , W'_2) puisque :

$W_1 \cup W_2 = (e_o) (W'_1 \cup W'_2)$ et $W'_1 \cup W'_2 = (W)(W'_1 \cup W'')$.

Qu'illeur est aussi projectivement équivalent résulte du calcul :

$$W_1 \cap (W'_1 \cup W'_2) = W'_2 \quad \text{et} : W_2 \cap (W'_1 \cup W'_2) = W'_2$$

qui est immédiat puisque $W''_i \cap (W''_1 \cup W''_2) = W''_i$

Le même raisonnement aurait pu naturellement être effectué en utilisant l'intervalle $(W_1 \cap W_2 , W'_1 \cup W'_2)$.

L'ensemble de ces résultats peut encore être résumé dans le théorème suivant :

4° - Si W_2 est plus fine que W_1, l'intervalle (W_1, W_2) est isomorphe au produit direct des intervalles (W_{1i}, W_{oi}) correspondant chacun à un composant e_i de W_1 et où W_{1i} et W_{oi} sont respectivement la partition la moins fine et la plus fine de l'ensemble e_i' quotient de e_i par la restriction-appropriée de W_2 considérée comme relation d'équivalence.

Enfin rappelons que l'on appelle <u>semi modulaires</u> les treillis tels que :

Si x et y couvrent $x \cap y$ alors $x \cup y$ couvre x et y , ce qui est un affaiblissement du lemme de BIRKHOFF que nous avons donné plus haut pour les treillis modulaires.

On peut établir que :

Les treillis duaux des treillis de partition finis sont semi-modulaires.

Il suffit de voir que W_1 et W_2 ne peuvent couvrir $W_3 = W_1 \cap W_2$ que dans les deux cas suivants :

$$1° - W_1 = (e_i)(W_2')(W_3') \text{ et } W_2 = (W_1')(e_2)(W_3')$$

où W_1' et W_2' sont respectivement des partitions de e_1 et e_2 .

$$2° - W_1 = (W_{11}')(W_3') \text{ et } W_2' = (W_{12}')(W_3')$$

où W_{11}' et W_{12}' sont deux partitions du même e_1 .

Le résultat découle alors de la définition même de la base des partitions.

IV. - FONCTIONS NUMÉRIQUES SUR LES ENSEMBLES ORDONNÉS

Dans la théorie des treillis, il est classique de définir une <u>valuation</u> comme une <u>fonction numérique</u> $|x|$ des éléments x d'un treillis <u>modulaire</u> satisfaisant à l'identité :

$$|x \cup y| + |x \cap y| = |x| + |y|$$

Afin d'étendre aisément cette définition par trop restrictive nous nous placerons d'emblée à un niveau plus général.

E étant un ensemble muni d'une relation d'ordre R et α étant un anneau commutatif, nous appellerons fonctions de (E , R) dans α <u>toute application</u> $f(x,y)$ <u>dans</u> α <u>des paires d'éléments</u> x <u>et</u> y <u>de</u> E <u>qui est nulle sauf quand</u> x R y.

Il s'agit donc plutôt en toute rigueur d'une fonction dans α des <u>intervalles</u> de E mais nous pourrons sans danger de confusion employer cette terminologie plus concise.

Si E est fini, on peut associer (cf. RIGUET, Thèse Paris 1951) à tout f (x,y) une matrice F dont l'élément $a_y^x = f(x,y)$ est nul quand x n'est pas tel que x R y . Il est évident qu'en raison de la transivité et de l'acyclicité de R , les matrices F satisfaisant à cette condition constituent <u>un nouvel anneau</u> $\bar{\alpha}$ dont <u>l'unité</u> I correspond à la <u>fonction Kronekerienne</u> :

$$\delta(x, y) = \begin{cases} 1 & \text{si } x = y \\ 0 & \text{dans tous les autres cas.} \end{cases}$$

De plus, si f (x , x) n'est jamais nul, il existe une matrice unique F telle que $F F^{-1} = I$.

Parmi les fonctions spécialement dignes d'intérêt on notera la fonction de consécutivité dont la matrice C représente la fonction :

$$f(x,y) = \begin{cases} 1 \text{ si } y \text{ couvre } x. \\ 0 \text{ dans tous les autres cas.} \end{cases}$$

et la fonction d'incidence à laquelle nous associerons la matrice S :

$$f(x,y) = \begin{cases} 1 \text{ si } x \subset y. \\ 0 \text{ dans tous les autres cas.} \end{cases}$$

La matrice S^{-1} correspond à la fonction de MÖBIUS de la relation d'ordre car si f (x,y) est définie par $f(x,y) = \sum_{x \subset y} f'(x,y)$ où f' est une autre fonction on a : $F = SF'$ et, par conséquent: $F' = S^{-1} F$.

Enfin, il est classique que l'élément a^x_y de C^n livre le nombre de chaînes distinctes de longueur n joignant x à y. On peut aussi, de manière plus profonde, rattacher certaines propriétés de la relation d'ordre R à celles de la réduite de JORDAN de la matrice S-I .

Nous arrêterons là ces généralités qui sont quelque peu extérieures à l'objet de ce travail mais qui peuvent présenter de l'intérêt pour les probabilistes en raison de leurs applications nombreuses aux problèmes de dénombrement finis. A titre d'exemple nous indiquerons seulement le calcul de l'expression formelle des moments d'un système de variables aléatoires en fonction de leurs cumulants ce qui nous fournira l'occasion de donner la fonction de MÖBIUS du treillis de toutes les partitions d'un ensemble de n objets.

Considérons n variables aléatoires $x_1 \ldots\ldots x_n$ réparties de manière quelconque. A tout composant X_i d'une partition W de l'ensemble des x, associons la valeur moyenne $m(X_i)$ du produit des variables x_i appartenant à X_i et à W elle-même le produit des expressions $m(x_i)$ relatives à chacun des composants. Désignons par K(W) l'expression analogue construite avec les cumulants.

Soit W_o la partition la plus fine; il lui correspond $m(W_o) = K(W_o) = m(x_1)\, m(x_2) \ldots\ldots m(x_n) = K(x_i) \ldots K(x_n)$

Considérons m(W) et K(W) comme deux fonctions de l'intervalle (W, W_o) du treillis de toutes les partitions de l'ensemble des x_i .

Nous supposerons connu le fait que $m(x_1 \ldots x_n)$ est la somme étendue à tous les W des K(W) correspondants, chacun d'eux étant affecté d'un coefficient unité, c'est-à-dire que l'on a : $m(W) = \sum K(W)$ où la sommation est étendue à toutes les W' plus fines que W. Par exemple :

$$M(x\ y\ z) = K(x\ y\ z) + K(x)\, K(y\ z) + K(y)\, K(x\ z) + K(z) K(x\ y) + K(x)\, K(y)\, K(z).$$

On aura donc : $K(x_1, x_2 \ldots x_n) = \sum f(E, W')\, m(W')$ où la somme est étendue aussi à toutes les partitions de E et où

$f(E,W)$ est le coefficient numérique représentant la fonction de MÖBIUS du treillis des partitions de E.

Par définition $f(W_1\ W_2)$ est nulle si W_1 n'est pas plus fine que W_2. Elle est égale à 1 si $W_1 = W_2$ et enfin, pour tout intervalle non dégénéré $(W_1\ W_2)$, elle satisfait à $\Sigma f(W_1, W') = \Sigma f(W', W_2) = 0$ où W' parcourt (W_1, W_2).

Il nous suffit ici de calculer $f(E, W_2)$ mais comme on l'a vu dans le chapitre précédent (E, W_2) est isomorphe à $(E'\ W'_0)$ où E' est le quotient de E par W_2 et où W_0 est la partition la plus fine de E'. Par conséquent :

$f(E, W_2) = f(E', W'_0)$. Il en résulte que $f(E, W_2)$ ne dépend que du nombre h_2 des composantes de W_2. On vérifiera que cette fonction est $(-1)^{h_2}(h_2)!$ pour les premières valeurs de h et on obtient le résultant par récurrence en utilisant l'identité :

$\sum_{i=1}^{m} S_m^i (-1)^i\ i! = 0$. (S_n^i : nombre de STIRLING de 2ème espèce)

puisque S_m^i est le nombre des partitions en i classes de m objets.

De l'expression de $K(x, x_1, \ldots, x_n)$ ainsi obtenue on peut déduire toutes les autres en identifiant certaines variables.

Ainsi par exemple :

$$K(xyz) = m(xyz) - m(x)\ m(yz) - m(y)m(zx) - m(z)m(xy) + 2m(x)m(y)m(z)$$

On en dériverait en faisant : x = z :

$$K(x^2y) = m(x^2y) - 2m(x)m(xy) - m(y)m(x^2) + 2(m(x)^2\ m(y))$$

et en faisant x = y = z

$$K(x^3) = m(x^3) - 3m(x)\ m(x^2) + 2(m(x))^3$$

LES VALUATIONS

Les fonctions qui nous occuperont maintenant sont celles qui dérivent d'une fonction au sens strict des éléments de E et nous introduirons la définition suivante :

$f(x,y)$ sera une fonction simple s'il existe une application $h(x)$ dans α des éléments de E telle que $f(x,y) = h(y) - h(x)$

On voit que pour un système physique évoluant de façon discrète si $f(x,y)$ est considérée comme attachée au passage de l'état x à l'état y, la condition que nous imposons revient à postuler l'existence d'une fonction d'état $h(x)$ indépendant de l'histoire antérieure du système.

Naturellement, nous aurions pu au lieu d'une loi additive. utiliser une loi de composition multiplicative $h(y) = f(x,y)\,h(x)$ et nous ferons souvent usage de cette possibilité.

Plus généralement encore, nous aurions pu supposer que α n'était pas commutatif. Un exemple simple en est fourni par les chaînes de MARKOFF. Dans ce cas $h(t)$ est le vecteur représentant la distribution des probabilités au temps t et $f(t_1, t_2)$ est la matrice décrivant les probabilités de transition.

Nous particulariserons encore les fonctions qui nous intéressent et R étant une relation d'équivalence entre intervalles de E nous définirons enfin les valuations :

La fonction simple $f(x,y)$ sera dite induite par une valuation $h(x)$ des éléments de E si $f(x,y) = f(x',y')$ chaque fois que les intervalles (x,y) et (x',y') sont équivalents selon R.

Dans l'exemple évoqué précédemment des chaînes de MARKOF, nous avions donc bien affaire à une valuation si la chaîne était constante dans le temps et si la relation R était simplement l'égalité de durée des intervalles.

Plus généralement il en serait de même si les états successifs d'un système physique étaient régis par un semi-groupe (ou groupoïde) d'opérateurs T_i. L'on voit alors que notre définition des valuations est une généralisation de la notion de caractère de groupe - telle que la définit WIENER - puisque elle s'exprime en notation multiplicative par :

$h(T_i\, x) = f(T_i')\, h(x)$ où $f(T_i)$ est indépendant de x.

En choissant pour R la relation d'équivalence projective on retrouve bien pour les treillis modulaires la définition classique qui peut s'écrire (en notation abrégée) $[h(x+y) - h(y) = h(x) - h(x\,y)$

Calculons maintenant les expressions $h(x(y+xz)$ et $h(xy + zy)$

dans un treillis quelconque en utilisant l'identité précédente :

$$h(x(y+xz)) = h(x) + h(y+xz) - h(x+y+xz)$$
$$= h(x) + h(y) + h(xz) - h(xyz) - h(x+y)$$
$$= h(xy) + h(xz) - h(xyz)$$
$$h(xy + xz) = h(xy) + h(xz) - h(xyz)$$

On voit que l'égalité de ces deux valeurs entraîne que dans tout treillis, les éléments modulairement équivalents ont la même valuation, quand celle-ci repose sur l'équivalence projective ce qui explique les limitations de la définition classique.

En outre, ceci montre le lien étroit qu'établissent les valuations entre les relations d'équivalence entre éléments et les relations d'équivalence entre intervalles.

On observera enfin que les probabilités associées à un système d'évènements x, y, ... quelconques constituent une valuation particulière, car l'identité fondamentale peut aussi bien s'énoncer sous la forme familière :

$$Pr(x \& y) = Pr(x) + Pr(y) - Pr(x \vee y)$$

Si donc on a affaire à un système d'évènements formant un treillis distributif dont les éléments "& irréductibles" et "V irréductibles" sont désignés respectivement par ai et a et si x est défini par :

$$x = a_1 \& a_2 \& \ldots a_n = \overset{\cup}{a}_1 \vee \overset{\cup}{a}_2 \ldots \vee \overset{\cup}{a}_n$$

on obtient par application répétée de cette identité les deux représentations classiques de $Pr(x)$ comme somme pondérée d'expressions telles que $Pr(a_1 \& a_2 \& \ldots)$ ou que $Pr(\overset{\cup}{a}_1 \vee \overset{\cup}{a}_2 \vee \ldots)$

Revenons maintenant au cas général.

Théorème : Les valuations de E pour R forment un module dont le rang est égal au nombre plus un des classes d'intervalles équivalents définies par R.

La première partie de cette assertion est évidente.

Nous pouvons donc associer à chaque classe C_i d'intervalles (x,y) équivalents une valuation h_i telle que :

$$f(x,y) = \begin{cases} 1 \text{ si } (x,y) \in C \\ 0 \text{ dans tous les autres cas.} \end{cases}$$

Réciproquement, si la fonction $f(x,y)$ déduite de la valuation h_i est telle que $f(x,y) = a_i$ pour $(x,y) \in C_i$, la différence:

$$f(x,y) - \sum a_i (h_i(y) - h(x))$$

est identiquement nulle. Ceci établit le théorème puisque f ne détermine h qu'à une constante près.

Théorème : Toute valuation d'un treillis de partition fini relative à l'équivalence normale R peut se mettre sous la forme :

$$h(W) = g(e_1) + g(e_2) + \ldots\ldots + g(e_k)$$

où $g(e_i)$ est une application dans α du composant e_i de la partition W.

Il est facile de vérifier que si h(W) est de la forme précédente les différences h(W T) -h(W) ne dépendent pas des ensembles e sur lesquels n'opère pas T. Réciproquement, soit W_o la partition la plus fine de E, c'est-à-dire $W_o = (a)(a)(c)\ldots.(x)$. Soit pour une certaine valuation $h(W) = K$. Choisissons arbitrairement dans α les valeurs g(a), g(b) ..telles que leur somme soit K. Pour tout e_i de F appelons W_i la partition la plus fine dont un composant est e_i et définissons

$$g(e_i) \text{ par } h(W_o) + h(W_i) + \sum_{a_j \in e_i} g(a_j)$$

Comme pour toute partition $W = (e_1)(e_2)\ldots(e_k)$, l'intervalle (W_1, W_o) est le produit direct des intervalles $(W', W_o)(W'')W_o)$ $\ldots.(W'_k, W_o)$, on doit avoir :

$$h(W_o) - h(W) = \sum(h(W_o) - h(W_i))$$

ce qui ne peut être identiquement vérifié que si :

$$h(W) = \sum g(e_i)$$

la somme étant étendue à toutes les composantes de W .

Comme les valuations ne sont définies qu'à une constante additive près, il pourra être avantageux d'utiliser la forme:

$$h(W) = g(e_1) - g(E)$$

qui s'adapte commodément au calcul sur un sous-treillis (W', W_o)

UNE ÉQUATION REMARQUABLE

Nous allons supposer maintenant que F contient toutes les parties de l'ensemble fini E et nous allons discuter une équation dont la solution est une valuation de L.

Théorème : Si la fonction $h(e_i, e_j)$ dans $\mathcal{C}$ des paires d'ensembles disjoints satisfait pour tout e_i, e_j, e_k aux deux conditions :

$1° - h(e_i, e_j) = h(e_j, e_i)$

$2° - h(e_i, e_j \cup e_k) + h(e_j, e_k) = h(e_j, e_i \cup e_k) + h(e_i, e_k)$

elle peut être mise sous la forme :

$$h(e_i, e_j) = f(e_i) + f(e_j) - f(e_i \cup e_j)$$

La condition I° implique que l'expression figurant dans 2° est une fonction symétrique de e_i, e_j et e_k que nous pouvons écrire : $h(e_i, e_j, e_k)$

Montrons d'abord par récurrence sur n que

$$h(e_i, e_j \ldots e_{n+1}) = h(e_i, e_j, e_n \; e_{n+1}) + h(e_n, e_{n+1})$$

est aussi une fonction symétrique de ses n + 1 arguments.

Par hypothèse on a :

$$h(e_n, e_{n+1} + h(e_1, e_2 \ldots e_{n-1}, e_n \cup e_{n-1}) =$$

$$h(e_1, e_2 \ldots e_{n-1} \cup e_n \cup e_{n+1}) + h(e_{n-1}, e_n \cup e_{n+1}) + h(e_n, e_{n+1})$$

$$= h(e_1, e_2, .., e_{n-1} \cup e_n \cup e_{n+1}) + h(e_{n-1}, e_n, e_{n+1})$$

ce qui montre que $h(e_1 \ldots . e_{n+1})$ est invariante par permutation de e_{n-1}, e_n et e_{n+1} et comme e_{n-1}, par hypothèse aussi, est un quelconque des n-1 premiers arguments de $h(e_1, e_2, \ldots e_n \; e_{n+1})$, ce résultat intermédiaire est établi.

Montrons maintenant que h est une valuation de treillis des partitions.

Comme F contient toutes les parties de E nous pouvons nous limiter aux opérateurs T qui partitionnent en deux sous-ensembles celui sur lequel ils opèrent.

Soit la partition :

$$W = (e_1 \cup e_2)(e_3 \cup e_4)(e_5)(e_6) \ldots . (e_n)$$

on peut alors vérifier que

$$h(e_1 \cup e_2, e_3 \cup e_4, e_5 \ldots e_n) - h(e_1 \cup e_2, e_3, e_4, e_5, \ldots e_n)$$

$$= h(e_1, e_2, e_3 \cup e_4, e_5 \ldots e_n) - h(e_1, e_2, e_3, e_4, e_5, \ldots e_n)$$

ce qui établit que h est une valuation. D'où le théorème.

Il est important d'observer que quand E n'est pas un espace topologique discret comme nous l'avons supposé, l'équation que nous venons de discuter, peut avoir des solutions qui ne sont pas des valuations. Une équation très analogue se rencontre dans la théorie de ELLENBERG et MAC LANE de l'extension des groupes abeliens - (1943) Ann. of. Math. (2) Vol 41) -

V. - DEUX APPLICATIONS

UNE APPLICATION AUX "STATISTIQUES D'ORDRE"

Nous considèrerons en application des résultats précédents, certaines distributions qui ont été étudiées séparément (1) par les statisticiens et nous montrerons qu'elles peuvent se déduire très simplement de ce seul fait que leur fonction génératrice est une valuation d'un treillis de partition.

Soit une famille de k ensembles A_1, $A_2 \ldots A_k$ comprenant chacun n_i éléments a_{ij}. Il est fréquent en statistique que les a puissent seulement être classés par ordre de grandeur et que l'on ait à tester l'hypothèse nulle selon laquelle les $(\Sigma n_i)!\,(n_1!\,n_2!\ldots n_k!)^{-1}$ permutations possibles seraient équiprobables contre la famille d'hypothèses alternatives selon lesquelles il y aurait toujours une probabilité plus grande que 1/2 qu'un objet a soit plus petit qu'un objet $a_{i'j'}$ si $i < i'$.

Les deux cas limites qui ont été envisagés systématiquement sont :

I° - Celui où k = 2, ce qui fournit un test non paramétrique de la différence de tendance centrale de A_1 et de A_2.

2° - Celui où, au contraire, k étant quelconque, chaque ensemble A_i ne contient qu'un seul objet, ce qui conduit aussi à un test non paramétrique mais cette fois portant sur la corrélation entre le rangement observé et un autre rangement a priori (c'est le coefficient "tau" de KENDALL).

Nous n'étudierons ici que l'aspect algébrique de cette question et nous chercherons les distributions de la variable $R = \sum_{i,i',j,j'} r(i,j,i',j')$

avec $r(i,j,i',j') = \begin{cases} 1 \text{ si } i < i' \text{ et } a_{ij} \text{ observé inférieur à } a_{i'j'} \\ 0 \text{ dans tous les autres cas} \end{cases}$

Il est facile de voir que R fournit la caractérisation la plus immédiate des propriétés d'ordre de l'échantillon observé.

Pour un système de valeurs de n_i donné, nous désignerons par P_x la probabilité pour que $R = x$ dans l'hypothèse H_0 et par $g(n_1, n_2 \ldots n_k) = \sum_x P_x \, t_x$ la fonction génératrice des P_x.

Théorème : g est une valuation multiplicative du treillis des partitions en segment de l'intervalle $(0, n_i)$

D'après les résultats du chapitre précédent, il suffit de montrer que si g' et g'' désignent respectivement les fonctions génératrices relatives à une famille formée de k-1 ensembles de puissance $n_1, n_2 \ldots, n_{i_0} + n_{i_{0+1}} \ldots n$ et à une famille composée seulement de deux ensembles de puissance n_{i_0} et $n_{i_{0+1}}$ on a $g = g'g''$

Cf : KENDALL.M.G.Rank correlation Methods-London 1948
WILCOXON.F. (1945) Biometrics Bulletin 1 p 80-82
HALDANE.J.B.S. an C.A.B.SMITH (1947) Ann.Eug.(14) p 117-124
KEEPING.E.S. (1952) Biometrics (8) p 112-119

Mais cette propriété résulte directement du fait que R est la somme de deux variables aléatoires indépendantes; l'une R' qui ne diffère de R que par le fait que l'on pose $r(i_0, j, i_0+1, j')$ identiquement nulle, c'est-à-dire que l'on ne se préoccupe pas de l'ordre relatif des objets des classes A_{i_0} et A_{i_0+1}; l'autre R'' définie au contraire par :

$r(i, j, i', j') = 0$ sauf si $i = i_0 + 1$ et si l'objet a est observé plus petit que $a_{j'}$.

Par conséquent on peut écrire g sous la forme $\dfrac{\prod \varphi_{n_i}(t)}{\varphi(t)}$

où $\varphi_{n_i}(t)$ est une certaine fonction qui ne dépend que de n_i et que nous allons déterminer.

Pour cela il suffira de considérer le cas très simple d'un ensemble A_1 avec n objets et d'un ensemble A_2 avec un seul objet. Ici, on a évidemment :

$$g = (1 + t + t^2 \ldots . t^h)(n+1)^{-1} = \varphi_n(t)\,\varphi_1(t)\,\varphi_{n+1}^{-1}(t)$$

En raison de l'homogénéité des formules on peut poser :

$$\varphi 1(t) = \frac{1-t}{1-t} = 1$$

et l'on déduit par récurrence du résultat précédent :

$$\varphi n(t) = \prod_{1}^{h} \frac{1-t^i}{1-t}\,(n!)^{-1}$$

d'où l'on peut dériver les distributions déjà connues dans les deux cas que nous avons mentionnés au début, c'est-à-dire :

1° - dans le cas de deux ensembles de puissance respectives n_1 et n_2 =
P_x = coefficient de t_x dans : $\prod_{i=1}^{n_1+n_2}(1-t^i)\left[\prod_{1}^{n_1}(1-t^i)\prod_{1}^{n_2}(1-t^i)\left[\begin{matrix} n_1+n_2 \\ n_1 \end{matrix}\right]\right]^{-1}$

2° - dans le cas de n ensembles formés chacun d'un seul objet

Px = coefficient de t dans : $(n!)^{-1}\prod_{1}^{n}\left(\frac{1-t^i}{1-t}\right)$

APPLICATIONS A L'ANALYSE GÉNÉRALE

Les formules précédentes ont une application algébrique intéressante.

Considérons dans un anneau deux éléments x et y tels que $yx = u\,x\,y$ où u appartient au centre de l'anneau et proposons-nous de calculer :

$$(x+y)^h = z \quad (n : \text{entier positif})$$

Il est immédiat que z est une somme de termes de la forme :

$$x_1^n \quad y^{n-n_1} \quad f_{n,n_1}(u)$$

où $f_{n,n_1}(u)$ est un polynome en u de degré maximum $n_1(n-n_1)$ puisque : $y^b\, x^a = u^{ab}\, x^a\, y^b$.

Mais $f_{n,n_1}(u)\, x^{n_1} y^{n-n_1}$ est la somme de $\frac{z\, n!}{n_1!\,(n-n_1)!} R$

termes élémentaires de la forme $u^R\, x^{n_1}\, y^{n-n_1}$ où R a la même signification que dans le paragraphe précédent. Par conséquent :

$$f_{n\,n_1}(u) = \prod_1^n (1-u^i) \times \prod_1^{n_1} (1-u^i)^{-1} \times \prod_1^{n-n_1} (1-u^i)^{-1}$$

Nous avons utilisé ailleurs (Comptes-Rendus 1953-236-p. 352-353) ce résultat pour donner une solution nouvelle de l'équation fonctionnelle :

$$F(x+y) = F(x)\, F(y) .$$

LA FORME GÉNÉRALE DES DISTRIBUTIONS DE LA STATISTIQUE QUANTIQUE

Dans l'application précédente, c'était une fonction génératrice qui apparaissait comme une valuation; ici au contraire ce seront des distributions de probabilité que nous déduirons directement de la formule générale.

Soit un ensemble E d'objets de puissance finie N. Nous appellerons ces objets des <u>particules</u>.

Soit d'autre part F un second ensemble qui sera l'espace dans lequel se trouvent ces particules.

Les parties de F seront appelées "<u>cases</u>" et on les supposera munies d'une mesure additive.

On définit en général les distributions de la statistique quantique en faisant dès le début des hypothèses sur la nature des particules et on en déduit les expressions classiques de BOLTZMANN, BOSE-EINSTEIN et FERMI (1). Mais dans cette méthode il est parfois difficile de faire la part de ce qui est nécessité logique et de ce qui est raisonnement physique ou encore calcul d'approximation.

Nous nous proposons donc à l'inverse de montrer brièvement comment des impératifs algébriques imposent aux distributions une forme générale qui est <u>indépendante de la nature des particules étudiées</u>.

<u>Théorème</u> : Toutes les distributions de la statistique quantique sont de la forme

$\left(\prod f^{*g_i}(n_i)\right) \left(f^{*\Sigma g_i}(\Sigma n_i)\right)^{-1}$ où f(n) désignant une certaine distribution, $f^{*g}(n)$ représente la distribution dont la transformée de FOURIER est égale à la $g^{ième}$ puissance de celle de f(n)

Cf par exemple : PERRIN.F.1939-Mécanique statistique quantique- XI
FORTET.R.1950-Calcul des probabilités- p 27-31

Désignons par $\Pr\begin{Bmatrix} n_1 , n_2 \ldots\ldots n_k \\ g_1 , g_2 , \ldots\ldots g_k \end{Bmatrix}$ la probabilité conditionnelle de trouver n_i particules dans la case 1 de mesure g_1 , n_2 particules dans la case 2 de mesure g_2 .... etc ... quand on sait que les $N = n_1 + n_2 \ldots\ldots n_k$ particules occupent un domaine de mesure totale $G = g_1 + g_2 \ldots\ldots g_k$

Avec ces notations, le théorème des probabilités composées s'écrit :

$$\Pr\begin{Bmatrix} n_1 & n_2 & n_3 \\ g_1 & g_2 & g_3 \end{Bmatrix} = P_i \begin{Bmatrix} n_1 , & n_2 \\ g & , g_2 \end{Bmatrix} \Pr\begin{Bmatrix} n_1 + n_2 , & n_3 \\ g_1 + g_2 , & g_3 \end{Bmatrix}$$

puisque par hypothèse la distribution de $n_1 + n_2$ particules à l'intérieur du domaine de mesure $g_1 + g_2$ est indépendante de la distribution des autres particules à l'extérieur de celui-ci. Par conséquent, si nous considérons le treillis des partitions en cases de l'espace où sont représentées les particules,

$\Pr\begin{Bmatrix} n_1 , n_2 \ldots\ldots n_k \\ g_1 , g_2 \ldots\ldots g_k \end{Bmatrix}$, en est une valuation multiplicative et peut être mise sous la forme :

$$\prod_i f\begin{pmatrix} n_i \\ g_i \end{pmatrix} \times f\begin{pmatrix} \sum n_i \\ \sum g_i \end{pmatrix}^{-1}$$ où $f\begin{pmatrix} n_i \\ g_i \end{pmatrix}$ désigne une certaine fonction encore indéterminée de n_i et de g_i .

Nous pouvons maintenant appliquer l'axiome des probabilités totales au cas particulier où deux cases seulement sont en cause. Pour toute valeur de g_1 et g_2 :

$$\sum_{n=0}^{n_1 + n_2} f\begin{pmatrix} n \\ g_1 \end{pmatrix} f\begin{pmatrix} n_1 + n_2 - n \\ g_2 \end{pmatrix} \times f\begin{pmatrix} n_1 + n_2 \\ g_1 + g_2 \end{pmatrix}^{-1} = 1$$

C'est-à-dire que pour tout g_1 et g_2 on a l'équation de convolution :

$$f\begin{pmatrix} n_1 + n_2 \\ g_1 \quad g_2 \end{pmatrix} = \sum_{n=0}^{n_1 + n} f\begin{pmatrix} n \\ g_1 \end{pmatrix} f\begin{pmatrix} n_1 + n_2 - n \\ g_1 \quad + g_2 \end{pmatrix}^{-1}$$

et comme il est possible de normaliser f sans que cela apporte le moindre changement aux équations, il s'en déduit le résultat annoncé par une démonstration classique qu'il est inutile que nous reproduisions.

On peut vérifier que les fonctions f ainsi introduites sont respectivement dans les cas considérés par les physiciens :

- la distribution de POISSON pour la statistique de BOLTZMANN (et ici g est une variable continue).
- la distribution binomiale pour la statistique de BOSE-EINSTEIN.
- la distribution binomiale négative pour la statistique de FERMI-DIRAC.

Dans les deux derniers cas, on observera que g est nécessairement un entier et que, d'autre part, les distributions élémentaires f contiennent un paramètre qui ne semble pas avoir un sens physique et qui d'ailleurs s'élimine de l'expression finale des probabilités.

DEUXIÈME PARTIE

THÉORIE DES INFORMATIONS

I. - DÉFINITION DES INFORMATIONS

Sans doute est-il trop tôt pour que puisse déjà être écrite l'histoire de la théorie de l'information. Aussi bien, est-ce là encore un chapitre tout récent des mathématiques appliquées auquel chaque année, sinon chaque mois ajoute des contributions qui en bouleversent les perspectives et les limites.

Cette théorie, comme on sait, tire ses origines de la pratique des communications et c'est la nécessité de délimiter et de mesurer ce qui était l'objet même de leur travail qui a conduit les fondateurs HARTLEY, TULLER, WIENER .. etc (39, 48. 77, 90) à définir mathématiquement le concept d'information.

Ces recherches ont trouvé leur aboutissement dans la formulation rigoureuse de C. SHANNON dont le mémoire fondamental (72) contient l'essentiel de ce que nous savons sur ce qu'on a appelé l'information sélective ; à une source aléatoire émettant des signaux avec des fréquences données, SHANNON associe un invariant numérique ayant la même expression formelle que l'entropie et il montre que les propriétés les plus significatives des messages émis par la source ne dépendent en définitive que cette quantité.

On peut aussi bien traduire ces résultats en un langage moins spécial (84) et considérer que les signaux sont simplement les résultats d'observations successives effectuées sur un objet dont les états que l'on cherche à identifier sont déterminés par un processus stochastique.

Sur cette base de nombreux travaux ont étendu ou appliqué le concept d'information aux domaines les plus divers : physique, théorique depuis SZILARD (70) avec GABOR (31,32,33) et MANDELBROT (55); optique (BLANC LAPIERRE) (10), physiologie de l'audition (HUGGINS) (42) , sociologie expérimentale (BAVELAS) (5,6,45) linguistique (MANDELBROT) (55), pour ne citer que quelques exemples typiques. Une revue générale sommaire, jusqu'en 1951, a été esquissée par CHERRY (13,14).

Mais, dans un tout autre domaine, il existait un invariant associé à un modèle stochastique que l'on appelait également "information" et qui avait donné lieu à une série importante de travaux (24,25,26,27). Nous voulons parler de l'information de FISHER, bien connue des statisticiens puisque constituant le concept central de la théorie de l'estimation. Les liens entre ces deux quantités apparaissaient d'ailleurs comme fort étroits :

Cependant que l'information de SHANNON livre le nombre minimum d'observations nécessaires pour déterminer exactement l'état d'un système aléatoire, le théorème de FRECHET-DARMOIS, (30,18,19) exorcise tout démon qui prétendrait estimer un para-

mètre inconnu avec une précision supérieure à l'inverse de l'information de FISHER. Enfin, la parenté formelle existant entre les expressions analytiques devait mener à chercher l'unification de ces deux théories qui s'étaient jusque là développées de façon indépendante.

C'est le problème général de l'information qui est d'en fonder la théorie dans un cadre conceptuel dégagé des restrictions imposées par la pratique de la statistique ou celle des communications ; on citera les travaux de BAR HILLEL et CARNAP (12), de MAC KAY (51,52,53,54) de FERON (22,23) de BARNARD (2) auxquels malheureusement pour rester dans les limites fixées, il sera impossible de consacrer ici plus que cette brève mention. Il est vraisemblable qu'à ces préoccupations l'historien futur associera des recherches déjà plus anciennes comme celles suscitées par les conceptions de FISHER sur la signification même des probabilités , dont BARNARD (2) et GOOD (34) ont bien vu la liaison avec le problème de l'information.

Le but du présent exposé est beaucoup plus modeste ; contrairement à la plupart des auteurs que nous venons de citer nous ne chercherons pas à définir par une analyse phénoménologique ou sémantique ce que doit être l' "information en soi" puis à vérifier ensuite que telle ou telle quantité en fournit plus ou moins bien une évaluation numérique. Au contraire, nous laisserons de côté le problème de la nature universelle de l'information pour nous attacher à celui de sa mesure dans des problèmes pratiques précis.

Les deux cas particuliers étudiés par FISHER et par SHANNON nous servant comme de repères, nous essayerons de trouver l'expression analytique la plus générale qui jouisse de leurs propriétés communes et nous définirons celle-ci par des postulats qui ne sont autre chose que les théorèmes relatifs à ces deux invariants.

Même si les nécessités du discours empèchent qu'on l'explicite à chaque fois, il sera donc convenu que par le terme "information" au singulier nous n'entendons rien de plus que "tout type d'expression analytique satisfaisant aux conditions énoncées".

Cet abandon de l'intuition physique ou philosophique pour les cheminements de l'algèbre aura cependant un avantage ; une fois obtenue l'expression générale de l' "information" - à notre sens - nous retrouverons en la particularisant des grandeurs parfois déjà connues mais dans lesquelles on n'avait point encore songé à voir des informations. Il peut y avoir là, à travers la diversité de ses mesures une voie pour des extensions de concept d'information lui-même.

En bref, si l'on nous permet ce latinisme douteux, nous chercherons à construire mathématiquement un "explicandum" puis à l'utiliser, du point de vue du calcul des probabilités bien plus qu'à livrer "l'explicatum" (12) des diverses informations concevables a priori. La tâche de relier à la physique les grandeurs ainsi définies par un raisonnement formel qui a été brillamment entreprise par MANDELBROT (55) reste donc extérieure à l'objet de ce travail.

L'INFORMATION COMME ATTACHÉE A UN PROCESSUS STOCHASTIQUE

C'est la première propriété commune aux informations de SHANNON et de FISHER que d'être relatives à des processus dans lesquels joue un élément de nature aléatoire.

Cette intervention ne s'effectue d'ailleurs pas de la même manière dans les deux cas. Dans celui d'une source qui émet des messages on convient de la considérer comme aléatoire pour exprimer le fait qu'elle est astreinte à produire des signaux avec des fréquences données a priori. Dans celui de l'estimation, selon la formulation définitive de NEYMANN et PEARSON (58,59), on s'interdit au contraire toute hypothèse semblable sur les valeurs inconnues du paramètre, mais toute la théorie repose sur l'existence d'un véritalbe processus stochastique interposé entre le choix arbitraire du paramètre inconnu et l'observateur.

On voit donc qu'il existe une certaine marge de disponibilités dans l'insertion du hasard à l'intérieur du modèle étudié et il nous semble qu'il y a là matière à des développements nouveaux dans un domaine qui apparemment n'a pas encore reçu l'attention qu'il mérite: celui de la possibilité et de la signification d'une sorte d'information dans les processus rigoureusement déterministes que sont les calculs numériques dans un système algébrique quelconque, d'ailleurs.

En effet, dans la perspective où se sont placés la plupart des auteurs, et nous-même, il n'y a aucun changement dans la quantité d'information quand l'on a réussi par exemple à calculer les racines d'une équation dont les coefficients étaient donnés. Il est de même équivalent de connaître les éléments intervenant dans un mot booléen ou de savoir en plus la valeur ("vrai" ou "faux") de celui-ci.

Il semble pourtant qu'il y ait lieu à des théorèmes du type de ceux étudiés par SHANNON pour la transmission des messages qui donneraient a priori la limite inférieure du nombre des opérations nécessaires pour le calcul. Peut être pourrait-on à ce propos évoquer la possibilité encore très vague de faire rentrer les problèmes dans le cadre des phénomènes aléatoires en introduïsant une certaine irreversibilité par l'hypothèse que le calculateur "oublierait" son équation initiale une fois qu'il aurait par exemple trouvé la valeur de la racine sur le feuillet de RIEMANN qui l'intéresse. Les résultats empiriques de LEMOINE (50) en "géométrographie" et les énoncés bien rudimentaires encore de l'algèbre des circuits électriques pourraient peut-être fournir une base concrète à des recherches dans cette voie.

Il est impossible enfin de clore ces remarques heuristiques sans souligner une autre limitation propre cette fois à ce travail et non à l'état d'avancement de la théorie : les mécanismes aléatoires qui interviennent ici seront finis et en quelque sorte intemporels. Il ne sera jamais question que d'une suite limitée de tirages au hasard discrets, c'est-à-dire, en définitive, d'une aléatoire unique dans un espace produit. Les problèmes que pose l'introduction de véritables processus stochastiques tels qu'ils apparaissent dans les travaux de GABOR ou dans les théories de la prédiction de KOLMOGOROF (44) WIENER(91) (Cf aussi(43) et les travaux de VILLE (79-83))dépassent le ca-

dre de cet exposé qui se bornera à la "statique" de l'information comme préliminaire à la "dynamique" qu'annoncent les recherches éminentes qui viennent d'être citées.

L'INFORMATION COMME VALUATION

Nous sommes donc amenés à considérer l'information comme une expression $H(\xi)$ attachée à une aléatoire ξ. A priori la caractérisation la plus générale sera de supposer que $H(\xi)$ est une fonction symétrique $H(X_1, X_2, \ldots X_k)$ des k états possibles X_i que peut prendre ξ lors d'un tirage et nous conviendrons, comme nous l'avons déjà dit, de nous limiter à un nombre fini de ceux-ci. Insistons sur le fait que H pour l'instant est une fonction absolument quelconque qui pourrait, par exemple, dépendre en plus de la probabilité a priori de chaque état, d'un système d'autres grandeurs physiques (distance à l'origine, nombre de fois où ξ a été observé à l'état Xi .... etc... etc...) à condition que chacune de celles-ci soit attachée à l'un des X_i, et qu'on en puisse déduire les valeurs correspondantes à tous les ensembles d'une base de partition F de l'ensemble X des X_i.

Il est commode d'introduire des observateurs O_i dont chacun est muni d'un dispositif lui permettant seulement de repérer ξ entre les composants (Y_{i1}) (Y_{i2})... d'une partition Wi de l'ensemble X des X_i relative à la base F.

Aussi au lieu de parler de l'information attachée à ξ ' où ξ' est une aléatoire dont les états sont Y_{i1}, Y_{i2} ... nous parlerons de l'information H_i sur ξ pour l'observateur O_i qui est caractérisé par cette même partition.

Considérons tous les couples d'observateurs $(O_i, O_{i'})$, où $O_{i'}$ ne diffère de o_i que par l'impossibilité où il est de séparer entre eux certains états $X_1, X_2, \ldots X_k$ qui sont au contraire distincts pour O_i.

La première condition que nous imposerons à H est d'être telle que la différence $H_i - H_{i'}$, soit la même pour tous les couples $(O_i, O_{i'})$. Mettant en correspondance les opérateurs simples de partition T_j et les instruments d'observation capables de "résoudre" comme disent les astronomes les parties X_j en sous-ensembles $X_{j1} \ldots X_{jk}$, il est naturel d'associer à la possession de chacun d'eux un accroissement bien défini de l'information. La condition précédente revient à postuler que cet accroissement est additif quand les ensembles correspondants X sont disjoints

Si l'on admet, ce qui est fort naturel aussi, que le coût de l'installation d'un dispositif d'observation est la somme des coûts de chacun des instruments élémentaires, on voit que cette condition revient à établir une dépendance linéaire entre l'information et ce coût total.

D'après les résultats obtenus dans la première partie, N est donc une valuation de treillis des partitions de X déterminé par l'existence des instruments d'observation et c'est un résultat immédiat que d'énoncer :

La valeur de l'information attachée à l'observation de l'aléatoire ξ est la forme $\Sigma\, g\,(X_i)$ où les X_i sont les ensembles d'états entre lesquels on peut distinguer celui pris par ξ et où g est une application des X_i dans un module α.

CONDITION IMPOSÉE PAR LA RESTRICTION X' DE X

Les expressions $\sum g(X_i)$ sont encore beaucoup trop générales pour être efficaces et nous allons les particulariser par d'autres considérations.

Jusqu'ici nous nous sommes bornés, de fait, à envisager la différence de l'information entre deux observateurs qui étaient censés posséder les mêmes connaissances a priori sur ξ, ce qui éliminait ipso facto toute nécessité d'une référence à l'ensemble des données auxiliaires qu'ils pouvaient avoir à l'avance sur cette aléatoire. Au lieu de cette comparaison en quelque sorte synchronique, considérons maintenant la différence diachronique entre l'information attachée à ξ quand l'observateur sait seulement que $\xi \in X$ et celle qui correspond à une étape intermédiaire du processus d'observation caractérisée par le fait que celui-ci sait en outre que ξ n'est pas dans un certain sous-semble $X'' = X - X'$.

Par exemple, si ξ est une variable numérique entière, disons le numéro d'un billet de loterie, nous devons considérer comme le gain d'information associé à la connaissance d'un premier chiffre la différence entre l'information avant et après cette détermination. Nous distinguerons donc le gain d'information qui n'a de sens que pour chacune des réalisations effectives de l'observation, de l'information proprement dite qui n'a au contraire de sens que pour autant que celle-ci reste encore à faire

Si avant l'observation l'information était donnée par $\sum_{X_i \in X} g\,(X_i)$, on pourrait considérer qu'après, elle devient $\sum_{X_i \in X-X'} g\,(X_i)$ quand le sous-ensemble $X'' = X - X'$ a été exclu par le résultat de l'observation.

En réalité, il ne saurait en être rigoureusement ainsi : le passage de X à X' a aussi en quelque sorte altéré la nature de chacun des X_i puisqu'ils ne sont plus relatifs à la même catégorie d'épreuves.

Par exemple, comme on avait $\sum_{X_i \in X} P_2\,(X_i) = 1$, les $Pr\,(X_i)$ ne peuvent pas constituer un invariant attaché intrinsèquement à un état X_i, hors de toute référence à la totalité des autres X_i. Pour marquer cette dépendance, nous écrirons explicitement : $H = \sum_{X_i \in X} g_X\,(X_i)$ et $H' = \sum_{X_i \in X'} g_{X'}\,(X_i)$ ce qui fait que g doit être envisagée comme une fonction de deux arguments.

Considérons le cas particulier où $X = X_1 + X_2 + X_3$ et soient deux observateurs O_1 et O_2.

O_1 commence par déterminer si ξ est ou non dans X_1 puis, dans cette deuxième alternative, il regarde alors lequel des deux ensembles X_2 ou X_3 contient ξ. O_2 procède de la même manière mais en commençant par X_3. Par exemple ξ étant, soit nulle, soit finie,

soit infinie, O_1 détermine d'abord si $\xi = 0$ ou $\xi \neq 0$ puis dans ce dernier cas si $\xi^{-1} = 0$ ou $\xi^{-1} \neq 0$.

Au contraire O_2 commence par l'examen de ξ^{-1} puis éventuellement observe ξ. Il est naturel de supposer que l'information attachée à ξ est, avant toute observation, la même pour O_1 et O_2. D'autre part, d'après le principe d'additivité que nous avons déjà utilisé, il est aussi naturel de poser pour O_1 :

$H(X_1, X_2, X_3) = H(X_1, X_2, X_3)$ (valeur de l'information relative à la première observation)

$+ Pr(X_2 + X_3)\, H(X_2, X_3)$ (valeur pondérée de l'information relative à la deuxième observation)

En effet cette relation ne fait que traduire la correspondance linéaire que nous avons déjà postulée entre information et coût d'observation.

On en déduit pour g l'équation fonctionnelle suivante où l'on a posé pour simplifier $P_i = Pr(X_i)$:

$$g_X(X_1) + g_X(X_2) + g_X(X_3)$$

$$= g_X(X_1) + g_X(X_2 + X_3) + (P_2 + P_3)\left(g_{X_2 + X_3}(X_2) + g_{X_2 + X_3}(X_3)\right.$$

$$= g_X(X_3) + g_X(X_1 + X_2) + (P_1 + P_2)\left(g_{X_1 + X_2}(X_1) + g_{X_1 + X_2}(X_2)\right).$$

Y et Z étant deux sous-ensembles ordonnés ($Y \subset Z$) de X posons :

$g_Z(Y) = \frac{Pr(Y)}{Pr(Z)}\, h_Z(Y)$ ce qui est légitime puisque si $Pr(Y)$ était nul, il n'y aurait plus matière à information et il serait donc logique que $g_Z(Y)$ soit nulle.

La première des équations précédentes peut alors s'écrire :

$$P_2\, h_X(X_2) + P_3\, h_X(X_3) - (P_2 + P_3) h_X(X_2 + X_3) = P_2\, h_{X_2 + X_3}(X_2) + P_3\, h_{X_2 + X_3}(X_3)$$

X ne figurant pas au second membre (puisqu'un changement de X multiplie seulement P_2 et P_3 par un facteur et que l'équation est homogène) le premier membre doit être indépendant de X_1 ce qui est le résultat auquel nous voulions aboutir et qui est le plus fort de ceux qui peuvent être obtenus si l'on ne fait pas d'hypothèse sur la nature des invariants qui caractérisent les ensembles X_i. Introduisons alors les restrictions suivantes :

1° - $Pr(Y_i)$ est un élément d'une certaine algèbre commutative de BANACH B (par exemple $Pr(X_i)$ est une fonction analytique deux fois différentielles d'un système de paramètres, ou bien $Pr(X_i)$ est une polynome en certains paramètres $a_1, a_2, \ldots$)

2° - $h_Z(Y)$ est un élément de l'algèbre commutative de BANACH dans laquelle est définie H. Elle est en outre une fonction continue de Pr(Y) et P(Z) et elle ne dépend de Y et de Z que par l'intermédiaire de ces probabilités.

Dans cette hypothèse, $h_Z(Y)$ doit être comme on le voit une fonction de $\frac{Pr(Y)}{Pr(Z)}$ seulement que nous écrivons $h(P(Y)/P(Z))$ et la

dernière équation devient :

$P_2 h(z P_2) + P_3 h(z P_3) - (P_2 + P_3) h(z P_2 + z P_3)$ = une fonction indépendante de l'élément z quelconque de l'algèbre B.

Soit encore, en supposant que $Pr(X_2) = Pr(X_3) = x$:

$2x(h(x z) - h(2 x z))$: une fonction indépendante de z ,

soit encore :

$$h(z x) - h(2 z x) = \text{constante}$$

ce qui est l'équation de SCHRODER (69) dont une solution continue est Log(x z) dans le cas où l'algèbre B se réduit à l'algèbre des nombres complexes.

Posons alors dans le cas général :

$$h(u) = h'(\log u) .$$

On doit avoir pour tout u :

$$h'(\text{Log } u) = h'(\log u + \text{Log } 2) + \text{Constante},$$

ce qui entraîne que h' ( ) soit un opérateur linéaire S .

Nous obtenons ainsi le théorème :

Toute information est la valeur moyenne, étendue à l'ensemble des états, de la résultante de l'application d'un opérateur linéaire S sur le logarithme de la probabilité a priori de chaque état.

Dans tous les cas que nous traiterons, le gain moyen d'information associé à une observation sera non négatif.

On pourra donc ajouter à la définition précédente la restriction :

L'opérateur S doit être tel que l'information correspondante soit toujours positive ou nulle.

REMARQUES SUR L'AXIOMATIQUE PRÉCEDENTE

SHANNON a donné une justification très analogue dans le cas qui l'intéressait et qui se limite a celui où p est un nombre. En outre, il introduit cette hypothèse supplémentaire que s'il existe N états équiprobables a priori, H soit une fonction croissante de N .

Il faut remarquer que l'équation fondamentale qui revient à décomposer l'observation complète de ξ en une série d'observations dichotomiques virtuelles, suffit à elle seule à entraîner que $Pr(x_{1n} \ldots x_k) H(x_1 \ldots x_k)$ soit une valuation. On pourrait donc abréger la démonstration. Il était cependant intéressant d'en analyser au maximum les différentes étapes car les valuations des treillis de partition se rencontrent dans de nombreux chapitres du calcul des probabilités et c'est exclusivement par la pondération $Pr(x_2 + x_3)$ que l'information se distingue d'une quantité telle que le "chi carré" des statisticiens. Nous reviendrons plus loin sur cette question.

II. - THÉORÈMES GÉNÉRAUX SUR LES INFORMATIONS

EXTENSION AU CAS OU N'EST PAS UNE VARIABLE DISCRÈTE

Nous n'avons étudié jusqu'ici qu'une variable susceptible de prendre seulement un nombre fini d'états et il serait important de pouvoir étendre la notion d'information à des cas plus généraux. Nous nous limiterons à celui où ξ est une variable numérique douée d'une densité de probabilité continue $f(x)$ avec $f(x)$ nulle en dehors d'un intervalle fini (a, b).

La démonstration s'étendrait sans peine aux cas plus généraux ((a , b) infini ou ξ vecteur aléatoire).

Considérons une famille infinie de partitions de plus en plus fine de (a b) en intervalles $(a_{ij} ; a_{i,j+1})$ de longueur m_{ij}.

Pour une partition donnée et quel que soit l'opérateur H l'information est :

$$\sum f(x_j) \; (\; m_{ij} \;) \; S \; \mathrm{Log} \; f \; (\; x_j \;) \; m_{ij}$$

$$\text{où} \quad a_{i,j} \leqslant x_j \leqslant a_{i,j+1}.$$

Introduisons maintenant dans f un paramètre t tel que

$$f_t \; (\; x) = \begin{cases} f(x) & \text{si } t = o \\ (a-b)^{-1} & \text{si } t = 1 \end{cases}$$

autrement dit, $f_1 (x)$ est supposée être une <u>distribution rectangulaire</u>. Puisque l'opération consistant à fixer la valeur du paramètre est une opération linéaire :

$$m_{ij} \; f_t \; (x_j) \quad S \quad \left[\mathrm{Log} \; f_t \; (x_j) \; m_{ij} \right]_{t=1}$$

est une information que nous pouvons appeler H_o, donc

$$H_o - H'_o = \bar{H}_o = \quad m_{ij} \; f_o \; (x_j) \; S \; \mathrm{Log} \; f_o \; (\; x_j)$$

est une information et comme $f_o(x_j)$ est continue, H_o tend vers l'intégrale $\int_a^b dx \; f(x) S \; \mathrm{Log} \; f(x)$ quand les partitions de (a,b) sont de plus en plus fines.

<u>Dans les cas où cette expression a un sens</u> nous l'appellerons aussi <u>l'information</u> et nous ne ferons pas de différence entre le cas discret et le cas continu. On remarquera toutefois que H'_o dépend des transformations que l'on pourrait faire subir à la variable x.

Il en est de même pour $\bar{H}$ ce qui est naturel puisque l'opérateur que l'on vient d'employer correspond en quelque sorte à cette partie (infinie) de l'information totale qui est associée à la possibilité d'une identification infiniment précise de la position de x et qui est celle que l'introduction du paramètre puis la soustraction de la quantité associée à t = 1 permet d'éliminer du résultat final.

INFORMATION CONDITIONNELLE

Nous avons déjà fait implicitement usage de cette notion extrêmement commode dûe aussi à SHANNON.

Considérons deux aléatoires η et ζ auxquelles correspondent des treillis de partition T_y et T_z . On a dit antérieurement que cette structure induisait de manière naturelle une base de partition pour l'ensemble X des couples $\xi = (\eta, \zeta)$.

C'est une convention de langage commode que d'appeler pour un opérateur S quelconque "information conditionnelle" la valeur moyenne $H(\eta|\zeta)$ - relativement à tous les états de ζ, de l'information $H(\eta|z_i)$ attachée à l'observation de η quand l'état Z_i de ξ est déjà connu.

Posons:

$$Pr(\eta \in Y_i ; \zeta \in Z_j) = P_{ij} \; ; \; p_i = Pr(\eta \in Y_i) \; ; \; p'_j = Pr(\zeta \in Z_j)$$

$$P_{i|j} = Pr(\eta \in Y_i \mid \zeta \in Z_j) = p_{ij}/p'_j$$

on peut écrire pour tout opérateur S :

$$H(\xi) = \sum_{ij} p_{ij} \, S \, Log \, p_{ij} = \sum_{ij} p_{ij} \, S \, Log \, p_{i|j} + \sum_j \left(\sum_i p_{ij} \right) S \, Log \, p_j$$

$$= \sum_j p'_j \; H(\eta \mid Z_j) + H(\zeta) = H(\eta|\zeta) + H(\zeta)$$

on aurait de même :

$$H(\xi) = H(\zeta|\eta) + H(\eta)$$

Supposons maintenant que η ne dépende que de ξ et ceci par un mécanisme aléatoire tel que l'information qui lui est attachée soit nulle quand ξ est déjà connu ; par exemple, - comme on le verra plus loin - si $S = d^2/dt^2$ (information de FISHER) où t est un certain paramètre, $H(\eta|X_i)$ est nul si t n'intervient pas dans la liaison stochastique entre η et ζ. $H(\eta|\zeta)$ est donc nul aussi.

Comme nous avons supposé que les informations étaient toujours positives $H(\zeta|\eta)$ est nécessairement plus grand que zéro Par conséquent :

$H(\eta) = H(\xi) - H(\zeta|\eta) = H(\zeta) - H(\zeta|\eta)$ est toujours plus petite que $H(\zeta)$.

Ceci correspond bien à l'idée intuitive que l'observation de η ne saurait être aussi efficace que celle de ξ.

VARIABLES INDÉPENDANTES

Supposons maintenant que ξ et η soient indépendantes c'est-à-dire que pour tout i et tout j $P_{i|j} = P_i$. Les calculs précédents montrent que $H(\eta \mid \xi) = H(\eta)$ et par une extension immédiate à un nombre quelconque de variables l'on en déduit le théorème fondamental :

Théorème : L'information est une fonction additive pour la composition des variables aléatoires indépendantes.

Réciproquement, il est intéressant de voir quelle forme doit avoir une valuation H pour être additive relativement à cette opération.

Prenons donc $H(\xi) = \sum g(X_i)$ où g est une fonction quelconque des ensembles X_i qui forment une partition de l'ensemble des états de ξ et supposons que $\xi = \eta \times \zeta$ où η et ζ sont deux aléatoires indépendantes susceptibles de prendre l'un des trois états Y_1, Y_2, Y_3 et Z_1 Z_2 Z_3 respectivement.

Nous écrirons X_{ij} pour $\eta \in Y_i$ et $\xi \in Z_j$ (c'est-à-dire que X_{ij} est l'intersection de X_i et Z_j) et par hypothèse $g(X_{ij})$ devra être une fonction symétrique $g(Y_i, Z_j)$ puisque l'indépendance de η et ξ signifie aussi que les invariants de ξ intervenant dans g peuvent s'exprimer en fonction des invariants de η et de ξ. Ici encore, nous aurons besoin d'un artifice pour éliminer le fait que les trois ensembles Y_1, Y_2 et Y_3 (et naturellement, Z_1 Z_2 et Z_3) ne sont pas indépendants puisqu'ils constituent la totalité des possibles. Nous introduirons pour cela les aléatoires $\bar{\eta}$ et $\bar{\xi}$ obtenues à partir de η et ξ en confondant respectivement Y_2 et Y_3 et Z_2 et Z_3.

Nous aurons donc, d'une part les quatre équations exprimant

$$H(\eta \times \xi) \quad H(\eta \times \bar{\xi}) \quad H(\bar{\eta} \times \xi) \text{ et } H(\bar{\eta} \times \xi)$$

comme sommes de $H(\eta)$, $H(\bar{\eta})$ $H(\xi)$ et $H(\bar{\xi})$

d'autre part, les huit équations exprimant ces valuations comme sommes de termes $g(X_{ij})$.

Les quatre premières équations montrent que :

$H(\eta \times \xi) - H(\eta \times \bar{\xi}) - H(\bar{\eta} \times \xi) - H(\eta \times \bar{\xi})$ est nul.

On en déduit l'équation :

$$+ g(X_{22}) + g(X_{23}) + g(X_{32}) + g(X_{33})$$

$$-g(X_{22} + X_{23}) - g(X_{32} + X_{33}) - g(X_{22} + X_{32}) - g(X_{23} + X_{33})$$

$$+g(X_{22} + X_{23} + X_{32} + X_{33}) = 0$$

Pour trois ensembles U V V' quelconques posons :

$$g(U_n V) + g(U_n V') - g(U_n (V \cup V')) = G(U; V, V')$$

($U_n V$ désignant l'intersection des ensembles U et V ....

l'équation précédente peut encore s'écrire :

$$G(Y_2 : Z_2, Z_3) + G(Y_3; Z_2, Z_3) = G(Y_2 + Y_3; Z_2, Z_3).$$

On en déduit le théorème :

Théorème : La condition nécessaire et suffisante pour que H soit additive est que g soit telle que pour tout triple Y, Z, Z' la fonction :

$$g(Y \cap Z) + g(Y \cap Z') - g\left((Y \cap Z) \cup (Y \cap Z')\right)$$

soit une fonction additive en Y.

Limitons-nous maintenant au cas des valuations numériques c'est-à-dire à celui où le seul invariant de X intervenant dans $g(X_i)$ est la variable numérique $Pr(X_i)$ On a :

La condition nécessaire et suffisante pour qu'une valuation numérique continue soit additive est qu'elle soit une information de SHANNON.

En effet, posons $g(U) = u\, g'(u)$ $(u = Pr(\xi \in U))$

la condition devient :

$$y\,z\,g'(y\,z) + y\,z'\,g'(y\,z') - (y\,z + y\,z')\,g'(y\,z + y\,z') = y\,K(z\,z')$$

où $K(z\,z')$ est une certaine fonction de z et de z'.

Supposons que $z = z'$, il vient :

$$g(y\,z) - g(2\,y\,z) = \frac{1}{2}\,\frac{K(z,z)}{z}$$

Le second membre étant indépendant de y, doit l'être aussi de z et l'on retombe sur l'équation de SCHRÖDER.

Dans le cas général, on retrouve évidemment les informations H et en outre toutes les expressions de la forme D (H)où D est un opérateur linéaire,mais l'inventaire complet de ces valuations reste semble-t-il encore à faire.

EXHAUSTIVITÉ

Nous pouvons encore appliquer ici cette notion si importante introduite par M. le Professeur DARMOIS (16, 17) (Cf aussi 36,45).

Considérons en effet plusieurs tirages successifs η_1 , η_2... η_n de η et supposons que nous sachions mettre l'espace $\eta_1 \times \eta_2 \times \ldots\ldots \eta_n$ en correspondance biunivoque avec un autre espace $\xi \times \xi'$ satisfaisant à la condition suivante : pour chaque valeur fixée de ξ , ξ' ne dépendra plus de η_i que par un mécanisme stochastique telque pour que l'opérateur S considéré, l'information correspondante est nulle.

Dans ce cas $H(\xi) = \Sigma H(\eta_i)$ et nous dirons que ξ est le résumé exhaustif des η_i

En particulier, si S annule toute expression qui ne dépend pas d'un paramètre θ et si ξ est déterminé à partir de ξ' par un mécanisme constant, ξ est exhaustif pour les η_i.

INFORMATION ET PROBABILITÉ D'ABSORPTION

Le problème fondamental de l'information au sens où nous l'entendons ici est celui du rapport entre les opérateurs S et les êtres mathématiques auxquels correspondent les informations associées. Nous n'avons pu éclaircir pleinement cette question. La difficulté semble dépendre en majeure partie de celle que l'on éprouve à définir qu'elle est précisément la classe de ces êtres. L'analogie qui existe entre la physique quantique où l'on associe aussi opérateurs linéaires et observables pourrait peut-être servir de guide dans cette voie.

Une autre approche qui semble fructueuse est celle qui relie la théorie de l'information à celle des probabilités d'absorption.

Considérons chaque observation comme un mouvement aléatoire dans un certain espace. Si l'on a réussi à attacher un opérateur linéaire à une certaine frontière dans cet espace de telle sorte que pour tout point de celle-ci l'information soit nulle, des théorèmes peuvent être établis qui relient le nombre moyen d'observations à effectuer au gain d'information moyen réalisé à chacune d'elles. Le théorème très général suivant que nous pensons nouveau est essentiel pour ce raisonnement.

Soit u une variable aléatoire qui s'accroit par saut à partir de zéro à chaque temps 1, 2 ...,n de manière absolument quelconque sauf qu'il y a une probabilité égale à 1 pour qu'elle atteigne la valeur K en un temps fini.

Soit $g\ (a_1, a_2 \ldots a_n)$ la longueur moyenne du $(n+1)^{\text{ième}}$ saut quand les positions successives de u forment la séquence $s_i = 0, a_1, a_2 \ldots a_n)$.

Théorème : Si quelle que soit la séquence possible s_i on a :

$$L_0 \leqslant g\ (s_i) \leqslant L_1 ;$$

le temps moyen mis par u pour atteindre K est compris dans l'intervalle $(K/L_1, K/L_0)$.

Démonstration : Nous appellerons $\bar{S}$ l'ensemble de toutes les séquences s_i se terminant par K et S l'ensemble (comprenant $\bar{S}$) de toutes les sous-séquences formées par les premières valeurs des séquences de $\bar{S}$. Nous poserons :

$P\ (s_i)$ = probabilité que la séquence des positions de u soit s_i

On a :

$$(1)\quad g\ (s_i) = \sum_i (a_1 \ldots a_n)\ \frac{Pr\ (0, a_1, a_2, a_n, a_i)}{Pr\ (0\ a_1, a_2 \ldots\ldots, a_n}$$

où la sommation s'étend à toutes les valeurs de a_i.

D'autre part le temps moyen $\bar{T}$ est par définition :

$$(2)\quad \bar{T} = \sum_{i \in j} t\ (\bar{s}_i)\ Pr\ (\bar{s}_i)$$ en notant $t\ (s_i)$

le nombre des termes qui constituent la séquence $\bar{s}_i$.

Cela peut encore s'écrire

$$(3)\quad \bar{T} = \sum_{i \in S} Pr\ (\ s_i\ )$$

où la sommation est étendue cette fois à toutes les séquences de S. En effet à toute séquence : $s_i = (\ 0, a_1, a_2, \ldots a_n)$ entrant dans (2) avec le coefficient $t\ (\ s_i\ ) = n$ correspond dans (3) la somme :

$$Pr\ (\ o\ a_1\ ) + Pr\ (o\ a_1\ a_2\ ) + \ldots\ldots Pr\ (\ o\ a_1\ a_2 \ldots . a_n).$$

Par sommation sur <u>toutes</u> les séquences (1) donne en posant $A\ (s_i\ )$ = dernier terme de la séquence s_i :

$$\sum_{i \in S} Pr(s_i)\ g\ (\ s_i\ ) \sum_{i \in s} \sum_{a} \left(a - A(s_i)\right) P_i\ (s_i) = K$$

puisque tous les a_i (sauf si $A(s_i) = K$) entrent munis du signe plus et du signe moins avec la même somme de probabilité.

Donc si l'on peut poser $g\ (s_i) = L + L'\ (s_i)$ où L est indépendant de s_i, il vient en introduisant cette valeur dans (4):

$$L \sum_{i \in S} Pr\ s_i = K - \sum_{j \in S} L'\ (\ s_i\ )\ Pr(s_i)$$

d'où le résultat annoncé quand L est choisi de telle sorte que l'on ait toujours :

$$L'\ (s_i) \leqslant 0 \quad \text{ou} \quad L'(s_i) \geqslant 0\ .$$

III. - LES INFORMATIONS PARTICULIÈRES

L'INFORMATION DE SHANNON WIENER

Le problème considéré par SHANNON (72) et pour lequel il a été amené à définir cette expression est le problème du codage et nous ne ferons guère ici que retranscrire en un langage plus familier aux statisticiens les résultats obtenus par cet auteur. En effet, comme on l'a très vite remarqué, il est équivalent soit d'étudier une source aléatoire émettant des signaux élémentaires avec des fréquences a priori données, soit de considérer une aléatoire ξ ayant pris l'un des états $X_i \in X$ et de chercher à reconnaître celui-ci (à faire un "diagnostic") par une série d'observations élémentaires.

Chaque observation consistant essentiellement en une partition d'un sous-ensemble de X, une famille d'observations successives permettant un diagnostic complet de ξ (nous dirons une "<u>procédure</u>") correspond à un <u>arbre</u>, au sens que nous avons donné à ce terme dans la première partie et on associera à chaque état X_i le nombre d'observations $L\ (x_i\ )$ qu'il faut effectuer pour apprendre que $\xi \in X_i$. Un état x_i de ξ étant un "message" dans le langage de la théorie des communications, $L(X_i\ )$ n'est autre que la longueur du mot X_i dans le code spécial qu'est une procédure.

Venons-en maintenant à l'information de SHANNON-WIENER, dite aussi "sélective" quoique à vrai dire l'emploi de ce mot soit assez peu explicite, car presque toutes les autres informa-

tions sont aussi sélectives. C'est la seule information numérique, c'est-à-dire, comme on l'a vu, telle que les Pr ($\xi \in X_i$) n'interviennent précisément que par leur valeur numérique. Elle n'est définie qu'à une constante près et nous poserons :

$$H = \sum p_i \ \mathrm{Log}_2 \ 1/p_i \qquad (\mathrm{Log}_2 = \mathrm{Log\ de\ base\ 2}).$$

On voit sans peine que H est toujours positive et qu'elle n'est égale à zéro que pour un système de X_i tel que tous les p_i soient nuls sauf un égal à l'unité. Ce cas correspond exactement à la situation de l'observateur quand il a achevé le diagnostic de ξ et nous pourrons donc appliquer le théorème du chapitre précédent sur les probabilités d'absorption.

Enfin, si ξ ne peut prendre que N états distincts, on vérifie sans peine que H est maximum quand tous les p_i sont égaux à 1/N. Ceci correspond bien à l'idée intuitive que l'on a de l'incertitude la plus grande sur l'état de ξ telle qu'elle s'exprime par exemple dans la théorie du théorème de BAYES. La valeur de H est alors Log_2 N et l'on voit que le choix de la constante était du au souci d'avoir H = 1; pour une observation dichotomique optimale. Enfin, pour clore ses généralités, faisons remarquer que l'inégalité $H \geqslant 0$ n'empêche pas qu'après une observation l'on puisse se trouver avec une incertitude plus grande qu'avant celle-ci. Par exemple si $p_1 = 0{,}98$; $p_2 = 0{,}01$; $p_3 = 0{,}01$, l'information initiale est 0,161; après qu'une première observation ait permis d'écarter X_i, elle devient 1,00. Trop fréquents sont les jours où la quasi certitude se change en doute complet pour qu'il soit nécessaire d'illustrer par des exemples.

Avant de passer à l'étude des propriétés de l'information de SHANNON WIENER, nous donnerons quelques indications sur son expression analytique.

Posons $S_r = \sum p^r$ (donc $S_1 = 1$) et

$$b\ (k,r) = \begin{bmatrix} k \\ r-1 \end{bmatrix} (-1)^{r-1} \text{ pour } r \geqslant 2 \text{ et } 1/h \text{ pour } r = 1.$$

Pour toutes les valeurs de k on a l'inégalité :

$$\mathrm{Log}\ H > \sum_{r=1}^{K+1} b\ (k,r)\ S_r$$

le second membre livrant un développement asymptotique de H en fonction des S .

En effet, il suffit de développer chacun des termes Log 1/pi sous la forme pi $p_i \sum_{r=1}^{\infty} (1-p_i)^r$ puis de sommer sur les indices i. On obtient ainsi pour les premières valeurs de k_i:
$[H\ \mathrm{Log}_2 > 11/6 - 3\ S_2 + 3/2\ S_3 - S_4 > 3/2 - 2\ S_2 + 1/2\ S_3 > 1 - S_2$

Qu'on ne trouve par cette méthode qu'un développement asymptotique, valable seulement pour k fini est évidemment lié au fait que déjà dans le cas d'une variable dichotomique H n'est pas développable en série de puissances de la variance pq (Cf. plus bas "Information de WALD".) Nous n'insisterons pas plus sur ce point qui ne semble pas avoir été remarqué, mais qui présente peut être un certain intérêt en liaison avec le problème d'exprimer H en fonction des autres invariants classiques de la distribution de ξ .

On peut signaler enfin que les résultats précédents donnent la formule commode :

$H \operatorname{Log}_2 \# \operatorname{Log} N + 1/2$ var t_i , quand les p_i sont de la forme $(1+t_i)N^{-1}$ avec t_i tendant uniformement vers zéro.

Venons-en maintenant aux propriétés de l'information H. On a d'abord le théorème fondamental de SHANNON :

> Théorème : Si l'information attachée à chaque observation est toujours comprise entre deux limites H_1 et H_2 , le nombre moyen L d'observations nécessaires pour établir le diagnostic de ξ est compris entre H/H_1 et H/H_2 ou H est l'information a priori sur ξ .

Ceci est une application immédiate du théorème sur les probabilités d'absorption. Un résultat plus profond du aussi à SHANNON est le suivant que nous formulerons ainsi :

> Si toutes les procédures ne comportant que des observations en K composantes (ou moins) sont possibles on peut en trouver une telle que $L \leqslant 1 + H \dfrac{\operatorname{Log}_2}{\operatorname{Log} K}$.

Rangeons les X_i par ordre de probabilités p_i décroissantes et associons à chaque X_i la somme $P_i = \sum_{1}^{i} P_j$ avec $P_1 = 0$. Développons chaque P_i en une somme de fractions K-adiques :

$\dfrac{a_{i1}}{K} \quad \dfrac{a_{i2}}{K^2} + \ldots \dfrac{a_{ix}}{K^x} \ldots$ de dénominateur K^x . Pour la première observation nous grouperons ensemble les X_i ayant même valeur de a_{i2} pour les secondes ceux qui ont même a_{i1} et a_{i2} etc....

Si r_i est défini par $\operatorname{Log}_k 1/p_i \leqslant r_i \leqslant 1 + \operatorname{Log} 1/p_i$, on voit que P_i diffère de P_{i+1} au moins à la $r_i$ième place de son développement et que, par conséquent, si $\xi \in X_i$, le diagnostic est achevé en r_i observations au plus. En multipliant chacune des inégalités précédentes par p_i, puis en sommant, on retrouve bien le résultat annoncé.

Cette procédure est en général très près de l'optimum, Nous indiquerons cependant d'après D.A.HUFFMAN (41) une méthode qui permet d'obtenir celui-ci.

Par définition $L = \Sigma\, p_i\, L\,(X_i)$; si donc l'arbre définissant la procédure est donné, L sera le plus petit possible quand on aura permuté les X_i de telle sorte que $p_i < p_j$ entraîne $L\,(X_j) < L\,(X_i)$. Par conséquent, les K ensembles X_i ayant les probabilités les plus faibles doivent être distingués les uns des autres dans une seule et ultime observation.

Considérons maintenant la variable $\bar{\xi}$ qui a la même distribution que ξ sauf que l'on considère comme un seul état les K états X_i précédents. Le même raisonnement s'applique et l'on aboutit ainsi de proche en proche à n'avoir plus que $K' \leqslant K$ ensemble.

Traitons à titre d'exemple le cas suivant où les observations sont toutes dichotomiques et où l'on s'est donné les probabilités suivantes de six états possibles de l'aléatoire :

$p_1 = 0,26 \quad p_2 = 0,21 \quad p_3 = 0,19 \quad p_4 = 0,16 \quad p_5 = 0,11 \quad p_6 = 0,07$

La procédure qu'indique la méthode de SHANNON peut être symbolisée par $((X_1) (X_2 , X_3))((X_4), (X_5 , X_6))$ où les parenthèses superposées indiquent les observations successives. On a donc :

$L = 2 \times 0,26 + 3 \times 0,21 + 3 \times 0,19 + 2 \times 0,16 + 3 \times 0,11 + 3 \times 0.07 = 2,57$ observations.

Dans la méthode de HUFFMAN on considère successivement les ensembles suivants :

$0,26 \; / \; 0,21 \; / \; 0,19 \; / \; 0,16 \; / \; 0,11 \; / \; 0,07$

$0,26 \; / \; 0,21 \; / \; 0,19 \; / \; 0,18 = 0,11 + 0,07 \; / \; 0,16$

$0,34 = 0,18 + 0,16 \; / \; 0,26 \; / \; 0,21 \; / \; 0,19$

$0,40 = 0,21 + 0,19 \; / \; 0,34 \; / \; 0,26$

$0,60 = 0,34 + 0,26 \; / \; 0,40$

ce qui conduit à la procédure rigoureusement optimale :

$$((X_1) ((X_4) (X_5 \; X_6))) \quad (X_2 , X_3)$$

qui donne à peine un gain de 2 % puisque :

$L = 2 \times 0,26 + 3 \times 0,16 + 4 \times 0,07 + 2 \times 0,21 + 2 \times 0,18 = 2,53$

Signalons enfin la méthode de FANO (21) qui consiste à diviser chaque fois selon les probabilités aussi voisines que possible de 1/2. Nous aurons à revenir longuement sur ce point de vue dans la IIIème partie.

Cas d'une variable continue.

Nous ne ferons qu'énoncer les résultats suivants dus aussi à SHANNON et qui constituent l'essentiel de ce que l'on connait des propriétés générales de cette information pour une variable numérique répartie de façon continue.

Valeurs extrêmes : Si ξ varie entre a et b finis, H maximum correspond à une répartition uniforme de la probabilité et est alors égal à $\mathrm{Log}_2 (b - a)$ ce qui est le cas considéré par WIENER.

Si ξ varie entre 0 et l'infini, H maximum pour $\bar{\xi}$ fixé correspond à la répartition $a \exp(ax)$ où $\bar{\xi} = a^{-1}$ et est égal à Log. a e.

Si $\bar{\xi}$ varie entre plus et moins l'infini, H maximum pour une variance fixée correspond à la distribution de LAPLACE GAUSS et est égale à $\mathrm{Log} \sqrt{2 \pi e} . \sigma$.

Cas de plusieurs variables

Soit $\xi = (x_1 , x_2 \ldots . x_n)$ = un vecteur aléatoire de densité de probabilité $f(\xi)$

Soit $\eta(\xi) = y \ldots\ldots y$) un autre vecteur obtenu par une transformation continue (non aléatoire) de ξ . Les informations attachées respectivement à ξ et à η sont :

$H(\xi)$ et $H(\eta) = H(\xi) - E(J)$ où $E(J)$ désigne la valeur moyenne du jacobien de la transformation.

La distribution de LAPLACE GAUSS à n variables est doué des mêmes propriétés extrêmales que la distribution à une variable et l'on a : $H = \text{Log}\ (\ 2\Pi e)^{n/2} \mid a_{ij} \mid^{1/2}$ où $|a_{ij}|$ est le déterminant des a_{ij} définissant la forme ξ quadratique $\xi A \xi^*$ dont la distribution est laplacienne.

Addition de variables aléatoires indépendantes.

D'après le théorème général, la somme des informations relatives à deux variables ξ et ξ' indépendantes est égale à l'information relative au couple (ξ, ξ'). Si au lieu du couple on ne connait que la somme $\xi = \xi + \xi'$ on a seulement les inégalités :

$$\bar{N}(\xi) + \bar{N}(\xi') \leqslant \bar{N}(\zeta) \leq \text{var}\ \xi + \text{var}\ \xi'$$

ou $\bar{N}(x)$ désigne "l'entropy power" c'est-à-dire la fonction $1/2\Pi e . \exp 2\, H(x)$ et le remplacement de l'une des inégalités par une égalité n'est possible que dans le cas d'une distribution de LAPLACE GAUSS où les trois expressions sont d'ailleurs égales.

L'INFORMATION DE FISHER

C'est, comme nous l'avons déjà dit, le premier exemple d'information qui ait été étudié et c'est aussi peut être celui qui joue le rôle le plus important dans la statistique mathématique.

Nous nous bornerons à définir l'information de FISHER dans le cas où les $\Pr(X_i) = p_i$ dépendent continuement d'un système de paramètres θ_j ($j = 1,2,\ldots,m$) que l'on cherche à estimer (Cf. 1,20).

Définition : L'information de FISHER H est la valeur moyenne de la matrice hessienne de Log 1/pi par rapport aux θ_j .

Par conséquent H est définie dans le module des m x m matrices et l'opérateur linéaire S est symbolisé par la matrice d'élément générique $\dfrac{\partial}{\partial\theta_\alpha \partial\theta_\beta}$.

Il en résulte que si l'on fait une transformation fonctionnelle sur les θ_j, la nouvelle information est égale à l'ancienne multipliée par le hessien de cette transformation.

Dans le cas où nous nous sommes placés on a évidemment :

$$-\sum_i p_i \frac{\partial^2}{\partial\theta\,\partial\theta'} \text{Log}\ p_i = \sum_i \frac{1}{p_i^2}\left(\frac{\partial}{\partial\theta} p_i\right)\left(\frac{\partial}{\partial\theta'} p_i\right) - \sum_i \frac{\partial^2}{\partial\theta\,\partial\theta'} p_i$$

$$-\sum_i \frac{\partial^2}{\partial\theta\partial\theta'} p_i + \sum_i p_i \left(\frac{\partial}{\partial\theta} \text{Log}\ p_i\right)\left(\frac{\partial}{\partial\theta'} \text{Log}\ p_i\right)$$

Comme $\Sigma p_i = 1$, le premier terme du dernier membre est identiquement nul et l'on obtient le résultat classique :

L'information de FISHER est égale à la matrice des variances et co-variances des dérivées logarithmiques des p_i.

Le théorème fondamental est alors le théorème de FRECHET DARMOIS (30.18.19) qui associe 1/H à la variance d'échantillonnage de toute estimation possible des $\theta\alpha$. Nous nous bornerons au cas où il n'existe qu'un seul paramètre, et désignant par x_i une observation de ξ, nous supposerons que $\varphi(x)$ est une fonction d'estimation correcte c'est-à-dire telle que :

$$\Sigma p_i \quad \varphi (x_i) = \theta$$

Par dérivation par rapport à θ il vient :

$$\Sigma \varphi(x_i) \; p_i \quad \frac{\partial}{\partial \theta} \quad \text{Log } p_i \; = 1$$

et comme :

$$\Sigma \, p_i \frac{\partial}{\partial \theta} \quad \text{Log } p_i \; = 0$$

on peut écrire :

$$\sum_t \left(\varphi(x_i) - \theta\right) p_i \frac{\partial}{\partial \theta} \quad \text{Log } p_i \; = 1$$

d'où, par l'inégalité de SCHWARTZ :

$$\text{Var } (\theta - \varphi(x) \geqslant 1/\text{Var} \frac{\partial}{\partial \theta} \quad \text{Log } p_i \; = 1/H$$

Ce que nous énoncerons par :

L'inverse de l'information de FISHER est la borne inférieure de la variance de toute estimation correcte de θ.

On sait aussi, ce que nous ne démontrerons pas, que cette borne est atteinte asymptotiquement quand $\varphi(x)$ est la fonction d'estimation basée sur l'équation au maximum de vraisemblance. Pour un nombre fini d'observations, on ne peut avoir égalité que si $(\xi(x) - \theta) \sqrt{p_i}$ est proportionnel à $\sqrt{p_i} \frac{\partial}{\partial \theta}$ Log p_i. Dans le cas continu, ceci entraîne que Log f(x) soit de la forme :

$$a (\theta) \; b (x) + c (\theta) \; + d(x)$$

ce qui peut aussi s'écrire, moyennant un double changement de variable, sous la forme que nous retrouverons plus tard :

$$f (y) = \frac{f (y)}{g (t)} \quad \exp \sqrt{-1} \; ty$$

où g(t) est la transformée de FOURIER de f(y)

Les extensions du théorème fondamental sont nombreuses : D'abord au cas continu où la sommation est remplacée par une intégrale de HOLLINGER (15) et où l'on doit imposer en plus ou bien la condition que les limites de la variation de ξ ne dépendent pas de ξ ou bien que $f(\xi)$ s'y annule. Ensuite, au cas de plusieurs paramètres (Théorème de DARMOIS) où, de nouveau, apparaît une liaison certaine avec les propriétés de la trans-

formée de FOURIER de la distribution. Enfin, un théorème analogue (CRAMER-RAO) (15) vaut dans le cas où les estimations sont affectées d'une erreur systématique et dans celui de HAMMERSLEY(37) où θ est astreint à n'avoir ses valeurs que dans un espace discret. Nous ne pouvons naturellement pas songer à donner même un sommaire des travaux relatifs à cette question. ce qui serait écrire l'histoire de la théorie de l'estimation, soit, pratiquement, de la statistique mathématique.

Nous démontrerons pour terminer un théorème qui est en quelque sorte le converse du théorème de SHANNON sur le codage.

Bornons-nous encore au cas d'un seul paramètre. L'expression "information de FISHER relative à θ contenue dans ξ " a un sens bien clair et il est intéressant de savoir quelle fraction de cette information peut être gagnée en moyenne par une seule observation dichotomique.

Ces considerations semblent pouvoir présenter quelque intérêt pratique dans certaines applications de la méthode de MONTE-CARLO où il pourrait être relativement plus aisé de savoir si ξ appartient ou non à un certain intervalle pour un grand nombre de tirages que de déterminer complètement la valeur pour un nombre moindre d'épreuves.

Théorème : Si l'aléatoire considérée est susceptible de prendre N états, il en existe au moins un tel que l'information attachée au fait de savoir si ξ est ou non à cet état soit au moins égale à 1/N-1 de l'information totale contenue dans ξ .

En effet, en posant :

$$u_i = \left(\frac{d}{d\theta} p_i\right)^2 / p_i(1-p_i) = H/N-1 + x_i \text{ , on peut écrire :}$$

$$H = \sum_1^N u_i \left(1 - p_i\right) = \frac{H}{N-1} \sum_1^N (1-p_i) + \sum_1^N (1-p_i)\, x_i$$

d'où le théorème, puisque $\sum_1^N (1-p_i)x_i$ est nul et que, par conséquent, l'un des x au moins est plus grand que zéro s'ils ne sont pas tous nuls.

Le contre exemple suivant montre qu'il n'est pas toujours vrai que l'on puisse obtenir la moitié de l'information en une seule observation.

$$p_1 = a + t \; ; \; p_2 = a - t \; ; \; p_3 = 1/2 - a + t \; ; \; p_4 = 1/2 - a - t$$

avec a tel que $a(1-a) \leqslant 1/32$.

Cependant nous avancerions volontiers l'hypothèse que des conditions générales très simples peuvent assurer cette possibilité d'une seule observation préférable à elle seule au reste du diagnostic.

Voici maintenant un théorème qui facilite singulièrement la recherche de cette observation dichotomique optimale. Appelons "premier composant" celui des deux sous-ensembles X' qui la constituent dont la probabilité a priori Pr(X') est plus petite ou égale à 1/2. En changeant au besoin le signe de θ, nous pouvons faire en sorte que $\frac{d}{d\theta}$ Log Pr(X') soit non négatif.

Théorème : Le premier composant de l'observation optimale contient tous les états tels que la dérivée logarithmique de leur probabilité soit supérieure à une certaine valeur non négative K .

Pour établir ce résultat, nous étudierons d'abord les cas particuliers de 3 et de 4 états. Ceux-ci étant notés : A, B , C , D , soient p , q , r , s leurs probabilités et a , b , c , d les dérivées logarithmiques correspondantes.

Si trois états A , B et C tels que a b c sont en cause, la perte d'information associée au fait que l'observateur confond A et B est donnée par :

$$p\,a^2 + p\,b^2 - (p-q)\,((p\,a+q\,b)/p+q)^2 = (a-b)^2\,\frac{pq}{p+q}$$

Il en résulte que l'observation symbolisée par (AC)(B) et consistant à confondre les états A et C n'est jamais meilleure à la fois que les deux observations (AB) (C) et (A) (BC).

En effet, on devrait avoir :

$$(a-c)^2\,pr/p+r \leqslant (a-b)^2\,pq/p+q \quad \text{et} \quad (a-c)^2\,pr/p+r \leqslant (b-c)\,qr/q+r$$

soit encore :

$$(a-c)\sqrt{\frac{p\,r}{p+r}}\times\frac{p+q}{p\,q} + (a-c)\sqrt{\frac{p\,r}{p+r}}\times\frac{q+r}{q\,r} = a-b+b-c = a-c$$

ce qui est impossible puisque :

$$\left(\sqrt{\frac{p\,r}{p+r}}\times\frac{p+q}{p\,q} + \sqrt{\frac{p\,r}{p+r}}\times\frac{q+r}{q\,r}\right)^2 \geqslant \frac{r(p+q) + p(q+r)}{q^2(p+r)} \geqslant 1$$

Soient maintenant 4 états avec encore $a \geqslant b \geqslant c \geqslant d$.

Montrons que l'observation dichotomique (AC) (BD) ne saurait être la meilleure et pour cela posons A' = A + C et D' = B + D ; les nouveaux états A' et D' correspondent au regroupement de A et de C d'une part, et B et D d'autre part et leurs dérivées logarithmiques sont :

$$a' = (p\,a + r\,c)/p + r \quad \text{et} \quad d' = (q\,b + s\,d)/q + s.$$

Si (AC) (BD) était optimale, A'(BD) qui n'est qu'une autre manière d'écrire cette observation devrait être meilleure que (A'B)(D) et (A'D)(B) et de même (AC)(D') devrait être meilleure que (A)(CD') et (AD)(C).

Ceci entraîne :

$$(b-d)^2\,q\,s/q+s \leqslant (a'-b)^2\,(p-r)q\,/\,p+r+q$$

$$(a-c)^2\,p\,r/p+r \leqslant (c-d')^2\,r\,(q+s)\,/\,r+q+s$$

En outre, on doit avoir $a > a' > b$ et $c > d' > d$ car sinon, d'après ce qui vient d'être vu pour trois états, ni (BD)(A') ni (AC)(D') ne sauraient être optimales. Ceci conduit aux nouvelles inégalités :

$$a'-b = p(a-b) - r(b-c)\ /p+r \leqslant (a-b)p\ /p+r$$

$$\text{et}\quad c-d' \leqslant (c-d)s/s+q$$

Comme $c<d$ et que, par conséquent, a-b et c-d sont plus petites que a-c et b-d, on a :

$$(c-d)^2 \leqslant (b-d)^2 \leqslant (a'-b)^2 \quad (p+r)q(p+q+r)^{-1} \quad q^{-1}\ s^{-1}\ (q+s)$$

$$(a-b)^2 \quad p^2 \quad s^{-1}\ (p+r)\ \ (q+s)(p+q+r)^{-1}$$

De même :

$$(a-b)^2 \leqslant (c-d)^2 \quad s^2 \quad p^{-1}\ \ (p+r)(q+s)^{-1}\ (q+r+s)^{-1}$$

ce qui ne se peut pas puisque ps est sûrement plus petit que :

$$(p+q+r)\quad (s+q+r).$$

La démonstration est achevée puisque X' étant le premier composant de l'observation optimale, le résultat précédent montre que si les états Y et Z appartiennent à X' il doit en être de même de tous les états T tels que :

$$\frac{d}{d\ t}\ \text{Log Pr}(Y) \leqslant \frac{d}{d\ t}\ \text{Log Pr}\ T \leqslant \frac{d}{d\ t}\ \text{Log Pr}\ Z$$

C'est-à-dire enfin qu'il doit exister une valeur K telle que tous les états, dont la dérivée logarithmique de la probabilité l'excède, appartiennent à X'.

Il reste à prouver que cette limite K doit être non négative.

$$\text{Posons}\ \frac{d}{d\ \theta}\ \text{Log Pr}(X') = P'/P \qquad (P=\text{Pr}(X')$$

L'information associée à l'observation optimale peut être écrite $\frac{P'^2}{P(1-P)}$. Si X' contenait un état Y tel que la dérivée logarithmique correspondante soit non positive, l'observation (X'(Y) (X''+Y) serait meilleure que (X')(X'') car

$\frac{d}{d\theta}$ Pr (X'-Y) serait plus grande que $\frac{d}{d\theta}$ Pr(X') et

Pr (X'-Y) x P(X'+Y) serait plus petite que Pr(X') Pr(X'').

L'INFORMATION DE WALD

Nous pensons pouvoir donner ce nom à l'expression que nous allons étudier en raison de la place capitale que, sans la nommer d'ailleurs, A.WALD lui a fait jouer dans l'analyse séquentielle (85). Il en est de même des autres auteurs qui l'ont considérée : ceux-ci ou bien n'en ayant pas reconnu le caractère d'information (57) ou bien ayant préféré ne l'envisager que comme une expression accessoire (46,47) de l'information de SHANNON.

Définition : les $p_i(\theta)$ dépendant d'un système de paramètres symbolisé par l'information de WALD, $W(\theta)$ est définie par l'opérateur linéaire $[\ ]_{\theta=\theta_1}^{\theta=\theta_0}$

Par conséquent $W(\theta)$ est la valeur moyenne de la variable z_i qui représente le logarithme du rapport des vraisemblances de l'évènement x_i dans les deux hypothèses:

$$\chi_o = \theta = \theta_o \quad \text{et } \chi_1 : \quad \theta = \theta_1$$

On a la propriété très importante :

$W(\theta)$ n'est jamais négatif et ne s'annule que si pour tous les états $p_i(\theta_o)$ est égal à $p_i(\theta_1)$, ceci étant son seul minimum, à la fois comme fonction des $p_i(\theta_o)$ et comme fonction des $p_i(\theta_1)$

En effet, pour un système quelconque de variations des $p_i(\theta_1)$ on a :

$$d\,W(\theta_1) = \sum_i d\,p_i(\theta_1)\left\{1 + \mathrm{Log}\,\frac{p_i(\theta_o)}{p_i(\theta_i)}\right\}$$ d'où: $p_i(\theta_o)/p_i(\theta_i)$ = constante.

Et de même :

$$d\,W(\theta_o) = \sum_i \frac{p_i(\theta_o)}{p_i(\theta_i)}\,d\,p_i(\theta_1) \text{ avec } \Sigma\, d\,p_i(\theta_i) = 0$$

D'où le résultat annoncé.

Dans le cas dichotomique, on a l'inégalité suivante qui semble nouvelle. Ecrivons :

$$D = p(\theta_o) - p(\theta_i) = q(\theta_i) - q(\theta_o)$$

$$\boxed{W \geqslant 2D^2 + \tfrac{4}{3}D^4.}$$

Posons en effet $2\,p(\theta_o) = 1-x$ et $2\,p(\theta_1) = 1-y$ après avoir choisi p de telle sorte que x soit positif.

On peut développer W en série de puissance de x et de y :

$$2\,W = (1-x)\,\mathrm{Log}(1+x)/(1-y) + (1+x)\,\mathrm{Log}\,(1+x)(1+y).$$

On trouve :

$$W = \sum_{i=1}^{\infty} (4\,i^2 - 2i)^{-1}\,(x^{2i} - 2ixy^{2i-1} + (2i-1)y^{2i})$$

Tous les termes sont positifs car le polynome $t^{2i} - 2it + 2i - 1$ a un unique extremum pour $t = 1$ et prend en ce point la valeur 0 .

Bien plus :

$$x^{2i} - 2ixy^{2i-1} + (2i-1)y^{2i} = 4\,D^2\,(x^{2i-2} + 2x^{2i-3}y + 3x^{2i-4}y^2 + \ldots$$

$$\ldots\ldots + (2i-1)y^{2i-2}\ )$$

Par conséquent W est plus grand que la somme des deux premiers termes de son développement qui sont :

$4\,D^2/2$ et $4\,D^2/12\,(x^2 + 2x\,y + 3\,y^2)$ et la valeur de ce dernier polynome étant supérieure pour D fixe à $D^2/3$ on trouve bien le résultat.

Si u désigne le plus grand de x et de y on a enfin l'inégalité complémentaire suivante qui se déduit du développement précédent :

$$W \leqslant 2\ D^2 (1-u^2)^{-1}$$

On observera qu'en faisant $p(\theta_1) = 1/2$ les résultats précédents permettent de trouver un développement de l'information de SHANNON en série de puissances (paires) de :

$$x = p-q = (1-4\ p\ q)\ 1/2.$$

Le théorème fondamental qui donne un sens physique a l'information de WALD appartient à la théorie de l'analyse séquentielle.

Soit à choisir entre les deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ d'après une méthode telle que les probabilités d'erreur soient inférieures ou égales respectivement à a et b selon que la vraie valeur de θ est θ_0 ou θ_1 ; l'analyse séquentielle prescrit l'emploi du "sequential probability ratio test" qui consiste à effectuer une série de tirages successifs indépendants de $x_1, x_2 \ldots x_n$ de ξ jusqu'à ce que la somme $Z_n = \sum_1^n z_\alpha$ atteigne l'une des deux limites A et B calculées à l'avance.

Comme $z_\alpha = \text{Log} \dfrac{Pr\ (\xi_\alpha(t)\ X_i \mid \theta_0)}{Pr\ (\xi_\alpha(t) X_i \mid \theta_1)}$, ceci revient exactement a poursuivre le test jusqu'à ce que le rapport des vraisemblances des deux hypothèses ait atteint certaines limites.

Nous appellerons N_0 et N_1 le nombre moyen d'observations nécessaires pour achever le test selon que c'est $\mathcal{H}_0$ ou $\mathcal{H}_1$ qui est vraie.

WALD a montré (86) qu'aucune autre procédure ne saurait être uniformément meilleure que le "sequential probability ratio test".

Mais les démonstrations sont trop longues et font appel à trop de notions extérieures à ce travail pour que nous puissions songer à les reproduire ici. Nous nous contenterons en suivant WALD lui-même à établir des résultats approchés valables quand les limites A et B sont assez grandes par rapport à $|z|$ pour que l'on puisse considérer comme pratiquement nulles les différences $|Z_n - A|$ et $|Z_n - B|$.

Choisissons un entier N tel que soit négligeable la probabilité que la procédure ne soit pas achevée au $N^{ième}$ tirage.

Supposons que n soit le numéro d'ordre de tirage auquel l'une des limites a été atteinte, mais que l'on ait quand même continué les observations jusqu'à N. On a, en désignant par E l'opération "valeur moyenne pour un θ donné" :

$$E\ (z) = W \quad ; \quad \sum_1^N \eta_\alpha = \sum_1^n \eta_\alpha + \sum_n^N \eta_\alpha$$

d'où puisque n est une variable aléatoire :

$$N\ W = E\ (\sum_1^n z_\alpha) + E\ (N-n) W$$

$$\text{soit} : \quad E\ (\sum_1^n z_\alpha) = E\ (n)\ W$$

Soient a et b les probabilités que $\sum_1^n z_\alpha$ atteigne respectivement A et B, on a enfin :

$$E(n) = (a\,A + b\,B)\,W^{-1}$$

Dans les hypothèses d'approximation où nous nous sommes placés on peut donc énoncer en tenant compte du théorème d'optimalité de WALD :

K/W (θi) où K est une fonction de a et de b, limite inférieurement le nombre d'épreuves indépendantes nécessaires si $\theta = \theta_i'$ pour tester l'hypothèse $\mathcal{H}_0$ contre l'hypothèse $\mathcal{H}_1$ avec les probabilités d'erreur a et b.

On remarquera l'analogie de cette propriété avec le théorème de FRECHET-DARMOIS. Nous allons voir que la liaison n'est pas fortuite et que l'information de WALD établit bien une médiation entre les problèmes de diagnostic et les problèmes d'estimation.

On a souvent observé (notamment 4) que si les $p_i(t)$ sont des fonctions deux fois dérivables du paramètre inconnu, l'information de FISHER (que nous écrirons ici F (t)) apparaît dans le développement en série de TAYLOR de W(t).

De fait on peut établir le résultat plus fort suivant :

Si $t_0 = t + dt$ et $t_1 = t - dt$; $W(t_1)$ et $W(t_0)$ ont les mêmes parties principales 2 F(t) qui est aussi la moitié de la variance de z_i.

On a en effet les développements :

$$p_i(t \pm dt) = p_i(t) \pm p'_i(t)\,dt + p''(t)\,dt^2 \pm Q\,dt^3$$

$$z_i = \text{Log}\,\frac{p_i(t+dt)}{p_i(t-dt)} = 2\,p'_i(t)/p_i(t)\,dt + R dt^3$$

Donc en négligeant les termes d'ordre supérieur ou égal à dt^3

$$W(t_0) = W(t_1) = 2\sum_i (p'_i(t)^2/p_i(t))\,dt^2 = 2\,F(t)\,dt^2$$

$$\text{et } \text{Var}(z_0) = 4\sum_i p_i(t)\,(p'_i(t)/p_i(t))^2\,dt^2 = 4\,F(t)\,dt^2$$

Par conséquent, si l'on effectue un nombre N constant et très grand de tirages et si $\mathcal{H}_0$ et $\mathcal{H}_1$ diffèrent extrêmement peu, la variable réduite :

$$Z_N^* = \sum_1^n z_i \sqrt{(4\,F(t))}\,/\,dt$$

est distribuée comme une variable de LAPLACE GAUSS de variance unité et de moyenne $\pm$ (F(t) 1/2 Ndt selon que c'est $\mathcal{H}_0$ ou $\mathcal{H}_1$ qui est vérifiée ce qui montre que, comme on le soupçonnait, les probabilités a et b ne dépendent à la limite que de F(t)

CAPACITE.

Nous ne possédons aucun théorème suffisamment général qui donne, comme dans le cas de l'information de FISHER, une aide pour étudier l'efficacité d'une observation unique. Il y a là un domaine de recherches très important qui se rattache, semble-t-il, à la théorie des inégalités relatives aux fonctions convexes.

Plus généralement, il nous faudrait savoir résoudre le problème suivant qui contient les précédents et que nous reformulerons d'après SHANNON.

Soient X_i les états de ξ. Supposons que ξ soit elle-même inobservable mais que l'on puisse déterminer l'état d'une seconde aléatoire, dépendant de ξ par les probabilités conditionnelles $P_2(\eta \vdash y_j \mid \xi \vdash X_i) = p_{j|i}$.

Nous avons vu que l'on a quelque soit S :

$$H(\xi \times \eta) = H(\xi) + H(\eta|\xi) = H(\eta) + H(\xi|\eta)$$

SHANNON(72) appelle "capacité" le maximum de $H(\xi) - H(\xi|\eta)$ relativement aux $p_i = pr(\xi \leftarrow X_i)$ pour des $p_{j/i}$ fixés et donne la solution du problème consistant à trouver précisément les p qui assurent ce maximum dans le cas de l'information qu'il considère

Le problème plus général consiste, soit à déterminer les p pour une information quelconque soit - (ce qui serait plus important dans la théorie de l'estimation et du test où il est fréquent que $H(\eta|\xi)$ soit nul et où par conséquent comme on l'a vu $H(\xi) - H(\xi|\eta) = H(\eta)$)- de choisir les $p_{j/i}$ en tenant compte de certaines contraintes imposées par la technique d'observation.

Le théorème que nous avons donné plus haut est un premier pas dans cette voie et se traduit dans les notations de ce paragraphe par $j = 1$ ou 2 et $p_j|i = 0$ ou 1.

LES INFORMATIONS DE TRI

Considérons une famille d'évènements e_i que nous appellerons élémentaires et dont nous supposerons que les probabilités a priori s'annulant pour $t = 0$ sont développables en séries de puissance croissante d'un paramètre t. Soit E l'ensemble de tous les évènements composés qui peuvent être construits à partir des e_i par les procédés habituels de la théorie des évènements compatibles et dépendants (29). Il est commode d'associer à chaque e_i son indicatrice x_i prenant la valeur 0 ou 1 et au système des e_i la variable ξ dans l'espace produit.

$$X = x_1 \times x_2 \times \ldots\ldots \times x_i.$$

Chaque évènement de E est alors caractérisé par un certain domaine X_i de X et sa réalisation peut être notée $\xi \vdash X_i$. Supposons maintenant que des observations antérieures ayant ou non déjà restreint le champ E' des états possibles de ξ, nous effectuions l'observation O correspondant à la partition $(X_1)(X_2)\ldots(X_k)$ de E'.

L'information attachée à l'opérateur $W\left[t\frac{\partial}{\partial t}\right]_{t=0}$ livre le nombre moyen d'évènements élémentaires supplémentaires dont O nous permet d'apprendre qu'ils ont été réalisés. Nous appellerons cette information "information de tri de première espèce".

Considérons en effet un sous-ensemble quelconque X_i de X .

Les deux assertions :

"X_i n'est possible que si au moins n_i évènements élémentaires au moins sont réalisés", et

"$Pr(X_i)$ présente un zéro d'ordre n_i à l'origine"

sont équivalentes d'après nos hypothèses sur le développement des $Pr(e_i)$ en puissance de t.

Mais d'autre part si $P = t^{n_i}(a+bt+\ldots) = t^{n_i} Q$ la valeur pour $t = 0$ de $t \frac{d}{dt}$ Log $P = n_i + t\, Q'/Q$ est précisément n_i. D'où le théorème puisque, plus généralement, si on savait déjà que n' évènements e_i devaient être réalisés, on aurait :

$$P = \frac{t^{n_i}}{t^{n'_i}} \frac{Q}{R} \quad \text{et} \quad \left[t\frac{d}{dt} \text{ Log } P\right]_{t=0} = n_i - n'$$

Naturellement, si l'on avait choisi t de telle sorte que chacune des expressions $1-Pr(e_i)$ présente un zéro simple pour $t=1$, on aurait une information analogue concernant le <u>nombre</u> d'évènements élémentaires <u>non réalisés</u>.

Enfin à l'opérateur : $\left[t\frac{d}{dt}\right]_{t=0} + \left[(1-t)\frac{-d}{dt}\right]_{t=1}$ correspondrait le <u>nombre des évènements dont l'état est encore indéterminé</u>.

On voit que ces opérateurs diffèrent de celui associé à l'information de SHANNON par le fait qu'ils ne tiennent aucun compte de <u>l'identité</u> des évènements en cause mais seulement de leur nombre ou même plutôt du minimum de leur nombre.

Il serait donc peut être plus imagé de parler d'"information de comptage". La terminologie adoptée trouve sa justification dans la IIIème partie de ce travail(Chapitre 4"Problèmes de tri").

Avant d'en terminer, nous donnerons encore un autre exemple d'information associée à des problèmes analogues. Toujours dans le cadre de ce système d'évènements e_i, supposons que nous ayons pu restreindre ξ par des observations préalables, à appartenir à un sous-ensemble E' impliquant qu'au moins n évènements élémentaires sont réalisés. Limitons-nous aux observations qui ne changent pas ce minimum (c'est-à-dire à celles qui n'apportent <u>aucune</u> information de tri de première espèce) et faisons l'hypothèse supplémentaire que pour <u>tous</u> les e_i le coefficient du terme t dans leur développement est égal à un .

> <u>L'information de tri de seconde espèce</u> associée à l'opérateur $[\,]_{t=0}$ livre le logarithme du nombre de manières dont la réalisation de n évènements élémentaires permet la réalisation de l'évènement observé.

En effet, d'après les théorèmes classiques et les hypothèses faites, si $Pr(X_i) = a_i\, t^n + \ldots$. c'est qu'il y a a_i combinaisons de n évènements réalisés qui assurent que $\xi \in x_i$.

Comme Pr(E') admet aussi un zéro d'ordre n avec un coefficient que nous pouvons écrire a, chacun des $Pr(X_i \mid E')$ correspon-

dant à l'un des composants de l'observation est de la forme

$$\frac{a_i\ t^n + b_i\ t^{n+1} + \ldots\ldots}{a\ t^n + b t^{n+1} + \ldots\ldots} = a_i/a + ct + c't^2 + \ldots$$

et par conséquent, la valeur pour t = 0 de Log $Pr(\xi \in E')$ est bien Log a_i - Log a.

Il serait facile d'illustrer d'exemples l'emploi de cette nouvelle information et nous en trouverons aussi l'occasion dans la IIIème partie.

On voit également comment pourraient se faire diverses généralisations en introduisant par exemple plusieurs paramètres correspondant chacun à une sous-famille de l'ensemble des évènements élémentaires.

Il y a là un domaine de recherches qui semble nouveau et qui serait de nature à éclaircir la signification physique des opérateurs S caractérisant les informations. Dans l'état actuel de nos connaissances et de manière plus heuristique que mathématique leur rôle apparaît être le suivant :

Dans chaque domaine (diagnostic, estimation....) certains états de la variable aléatoire peuvent être considérés comme équivalents du point de vue spécial où l'on se place ; ainsi dans les problèmes d'estimation par exemple, est-il indifférent que les tirages successifs aient donné les valeurs x_1, $x_2 \ldots x_n$ ou toute autre permutation de ces n nombres. De même par hypothèse, lorsqu'il s'agira d'extraire d'une population infinie un objet présentant telle propriété spécifiée à l'avance ne nous soucierons-nous point d'apprendre qu'un ou plusieurs objets ne présentent pas cette propriété.

Ceci suggère inévitablement le recours a la notion fondamentale d'exhaustivité introduite par Mr. le Professeur G. DARMOIS ; pour chaque type de problème seuls certains invariants de certaines classes d'états sont à considérer et leur connaissance épuise pour le but particulier poursuivi tout ce que nous désirions savoir.

Il apparaît donc légitime que l'opérateur linéaire destiné à transformer en une information efficace - en une "information pour nous", - "l'information en soi", abstraite et universelle Σp_i S Log p_i, fasse apparaître précisément cette équivalence des états et, par là même, réintroduise dans le concept d'information la notion de sa valeur que nous avions volontairement feint d'ignorer lors de la définition axiomatique. En même temps les informations qui sont des grandeurs physiques recouvrent une dimension - celle-ci d'ailleurs pouvant être aussi bien un nombre pur (information de WALD) que l'inverse du carré d'une grandeur (information de FISHER) selon le modèle dont elles sont un invariant.

Dans cette perspective, on voit comment pourraient s'ordonner les différents types avec à la base l'information de SHANNON WIENER redondante par rapport à toute autre puisqu'impliquant une connaissance absolue de l'aléatoire. Ces remarques s'apparentent aux conceptions de MANDELBROT (55) sur le temps relatif des diverses stratégies liées aux théories physiques. Elles expliquent, nous semble-t-il, le succès de l'emploi de l'informa-

tion de SHANNON dans l'étude de la transmission des messages en présence de bruit bien qu'en toute rigueur il s'agisse là d'un problème d'analyse descriminative (Cf.89, 66, 63, 75) - donc d'une information de Wald. L'opérateur S = constante fournit en effet seulement une quantité approchée, qui est valable à la limite dans les cas particuliers envisagés mais n'est absolument correcte que si le bruit à une structure telle que l'on peut en principe achever le diagnostic (38) c'est-à-dire s'il n'est plus aléatoire quand l'observation dure un temps suffisamment long.

IV. - PSEUDO INFORMATIONS ET "CHI CARRÉ"

LES PSEUDO INFORMATIONS

Nous appellerons ainsi certaines expressions qui quoique pouvant être mises sous la forme Σp_i S Log P_i ne sont pourtant pas en général des informations. Le seul exemple que nous indiquerons sera celui des cumulants et celui des valeurs extrêmes de la variable aléatoire, qui dérivent, d'ailleurs de ces derniers.

ξ pouvant prendre les valeurs numériques x_i avec les probabilités p_i nous dirons que le paramètre t intervient de façon canonique dans la distribution de ζ si ζ prend la valeur de x_i avec la probabilité $q_i = p_i (\varphi(t))^{-1} \exp(\sqrt{-1}\ t\ x_i)$ ou $\varphi(t)$ est la transformée de FOURIER de la distribution initiale c'est-à-dire, à une constante près $\varphi(t) = \sum p_i \exp \sqrt{-1}\ t\ x_i$.

Dans le cas continu on définirait de même :

$$g(\zeta) = \left(\int_{-\infty}^{\infty} d\xi\, f(\xi) \exp \sqrt{-1} t \xi\right)^{-1} f(\zeta) \exp \sqrt{-1} t \zeta$$

Il est clair que si t intervient de manière canonique dans la distribution des variables indépendantes ξ et ξ' il en est de même pour la distribution produit symbolique f x f'

Par contre - et c'est pourquoi nous n'obtiendrons que des pseudo informations - si l'on considère la variable ζ déduite de ζ en confondant certains états, les termes correspondants dans l'expression de q'_i ne sont pas réductibles à une paramétrisation canonique puisque, par exemple :

$p_1 \exp \sqrt{-1}\ t\ x_i + p_2 \exp \sqrt{-1}\ t\ x_2$ ne peut être mis sous la forme :

$$(p_1 + p_2) \exp \sqrt{-1}\ t\ y.$$

Considérons maintenant les opérateurs :

$S_h = (\sqrt{1})^h \frac{d^h}{dt^h}$ et les informations K_h qui leur sont attachées. Par hypothèse :

$$K_h = (\sqrt{-1})^h \left(\frac{d}{dt}\right)^h \operatorname{Log} \varphi(t)$$

est le cumulent du h-ième ordre de ζ pour $h \geqslant 2$ et est égal à zéro pour h = 1.

Par conséquent :

les cumulants sont égaux à des informations quand le paramètre est canonique et dans ce cas seulement :

Nous dirons que les cumulants sont des pseudo-informations. On retrouve bien ainsi leurs propriétés classiques d'additivité et, dans une certaine mesure, on voit comment pourraient se rattacher à la théorie générale les quantités introduites par BHATTACHARYYA (7, 8, 9).

Sous des conditions très générales l'opérateur symbolisé par $(i-\sqrt{1}\,\frac{d}{d\,t})^{-1}$ a un sens et donc :

le logarithme de la transformée de FOURIER de f (la fonction génératrice des cumulants) est une pseudo-information.

Si l'on est dans le cas continu et si a et b finis sont respectivement la plus grande et la plus petite valeur de ξ telle que $f(\xi) \neq 0$, on sait d'après Polyà (65) que a et b peuvent être retrouvés par des opérations linéaires sur $\mathrm{Log}.\varphi(t)$, ceci resultant du théorème de PALEY-WIENER et du fait que $\varphi(t)$ est une fonction du type exponentiel. L'on a :

$$a = \lim_{n\to+\infty} r^{-1}\,\mathrm{Log}\,\varphi(\sqrt{-1}r) \quad \text{et} \quad b = \lim_{r\to-\infty} r^{-1}\,\mathrm{Log}\,\varphi(\sqrt{-1}\,r)$$

Par conséquent :

Les valeurs extrêmes de $\xi-\bar{\xi}$ sont des pseudo-informations.

Ceci était d'ailleurs vraisemblable a priori puisque ce sont des grandeurs additives pour la composition des variables indépendantes.

On observera cependant que l'ensemble des relations d'équivalence correspondant aux regroupements d'états qui n'affectent pas la valeur de cette pseudo-information est beaucoup plus vaste que dans le cas des cumulants.

Par exemple, si l'on considère comme concentrée à l'origine toute la probabilité sauf une fraction ε en a et une autre ε' en b, ces limites ne sont évidemment pas affectées.

Nous ne saurions terminer ces brèves indications sans attirer l'attention sur un problème de nature analytique qui ne semble pas avoir reçu encore de solution satisfaisante et qui est la réciproque des résultats précédents : celui de la détermination directe des informations à partir de la seconde caractéristique sans l'étape intermédiaire du calcul explicite des probabilités de chaque état.

LE "CHI CARRÉ"

Nous avons vu que l'information de WALD fournissait une sorte de pseudo-distance entre deux distributions. Les statisticiens font souvent usage depuis PEARSON (01,62) (Cf.40 et 56)

d'une autre grandeur qu'ils appellent communément le "chi carré" et qui est définie par :

$$\chi^2 = \sum_i \frac{p_i(\theta_o) - p_i(\theta_1)^2}{p_i(\theta_o)}$$

χ^2, outre certaines facilités de calcul, possède des propriétés très intéressantes pour de nombreuses applications (1) d'invariance par transformation par un certain groupe orthogonal et de décomposition subséquente. Manifestement χ^2 est une valuation du treillis des partitions des états de l'aléatoire considérée et l'on vérifie sans peine qu'il ne jouit pas d'additivité pour la composition des distributions indépendantes. Le "chi carré" n'est donc pas une information.

Nous allons voir comment il est cependant possible, comme pour celles-ci, d'en calculer la valeur en utilisant des restrictions du treillis de partition de X .

Soit $X = (X_1)\ (X_2)\ (X_3)$ et posons pour abréger :

$$p_i(\theta_o) = a_i \ ; \ p_i(\theta_1) = b_i$$

Si l'on considère que la partition $(X_1)\ (X_2 + X_3)$ on a :

$$\bar{\chi}^2 = \frac{(a_1 - b_1)^2}{a_1} + \frac{(a_2 + a_3 - b_2 - b_3)^2}{a_2 + a_3}$$

Si l'on se restreint au contraire à $(X_2)\ (X_3)$ on trouve :

$$\chi'^2 = \left(\frac{a_1}{a_2 + a_3} - \frac{b_2}{b_2 + b_3}\right)^2 \left(\frac{a_2}{a_2 + a_3}\right)^{-1} + \left(\frac{a_3}{a_2 + a_3} - \frac{b_3}{b_2 + b_3}\right)^2 \left(\frac{a_3}{a_2 + a_3}\right)^{-1}$$

En vertu de l'identité :

$$\frac{x^2}{y} + \frac{x'^2}{y} - \frac{(x + x')^2}{y + y'} = \frac{(x\,y' - x'\,y)^2}{y\,y'\,(y + y')}$$

il vient :

$$\chi^2 = \bar{\chi}^2 + (a_2 + a_3)\left(\frac{b_2 + b_3}{a_2 + a_3}\right)^2 \chi'^2$$

formule qui est très analogue, mais cependant différente, de celle que nous avons postulée pour les informations.

D'autre part, on peut écrire aussi χ^2 sous la forme :

$\sum \frac{p'^2_i(t)}{p_i(t)}$ en posant $p'_i(t) = p_i(\theta_1) - p_1(\theta_o)$ et comparer cette expression à l'information de FISHER qui a la même apparence formelle. Dans cette perspective on constate que la différence provient de ce que si p' est effectivement $\frac{d}{dt} p_i \frac{d}{d\theta} p_i$ il doit devenir $\frac{p'P - P'p}{2}$ quand on considère ξ comme restreint à une partie X' de X de probabilité totale P, alors qu'il devient $\frac{p'}{p}$ dans le calcul de χ^2.

(1) En général les pi(t) ne sont pas des distributions a priori mais sont des nombres de cas observés ce qui entraîne que leur somme n'est pas égale à 1. Ceci n'a aucune importance du point de vue où nous nous plaçons.

Nous arrêterons là ces quelques remarques qui n'avaient d'autre but que de relier en les contrastant la théorie du test de "chi carré" à la théorie de l'information.

RÉFÉRENCES BIBLIOGRAPHIQUES

I - AITKEN A.C. (1948) Proc.Soc Roy.Ed. A.62 P.362

2 - BARNARD G.A. (1951) Jour. Roy Stat. Soc. (138) P.46

3 - BARTLETT M.S. (1950) Proc. Lond. Symp P.81

4 - BARTLETT M.S. (1946) J. Roy. Stat. Soc 8 P.27

5 - BAVELAS (1950) J. Acoustical Joc. Am. 22. P.725

6 - BAVELAS (1951) Communication patterns;"Cybernetics" J. Macy. Conf)

7 - BHATTACHRYYA A. (1946) Sankya (8) P.1

8 - BHATTACHRYYA A. (1947) Sankya (8) P.201

9 - BHATTACHRYYA A. (1948) Sankya (8) P.315

10 - BLANC LAPIERRE A et PERROT M. (1950) C.R.Ac Sci (231) P.539

11 - BLUNDELL (1952) Proc. Lond. Symp

12 - CARNAP. R. and BAR HILLEL Y. (1952) M.I.T. Tech. Rep 247

13 - CHERRY E.C. (1951) Proc. Inst E.P. Engn (98.111) P.383-393

14 - CHERRY E.C. (1951) Proc. Symp. Lond. 161-168

15 - CRAMER (1946) Mathematical Methods. Of Stat. Princeton

16 - DARMOIS G. (1935) C.R.Ac. Sc. (200) P.1176 et 1265

17 - DARMOIS G. (1936) Inst. Int. Stat P. 288

18 - DARMOIS G. (1945) Rev. Inst. Int. Stat.

19 - DARMOIS G. (1942) C.R.Ac. Sci (222) P.164

20 - DARMOIS G. (1942) C.Ac.Sc.(221) P.266

21 - DUGUE O. (1939) Ecole Poly (3) P.305

22 - FANO R.M. (1949) M.I.T. Tech. Rep N° 65

23 - FERON R. (1950) CR.Ac.Sc.T.230 P.1495-97

24 - FERON R. et G. FOURGEAUD (1951) C.R.Ac.Sc.t:232 P.1636-1638

25 - FISHER R.A. (1944) Statistucal Methods.Edinburgh.

26 - FISHER R.A. (1942) The design of experiments.Edinbugh.

27 - FISHER R.A. (1921) Phil.Trans.Roy.Soc. A.22 P.309

28 - FISHER R.A. (1925) Proc. Camb. Phil. Soc. (22) P.700

29 - FORTET R. (1951) Processus stationnaires et entropie ; La Cybernétique ... Paris

30 - FRECHET M. (Paris 1943) Les probabilités associées à un système d'évènements compatibles et dépendants Ière partie.

31 - FRECHET M. Rev. Inst. Stat. 3/4/ P.182

32 - GABOR D. (1950) La théorie des communications et la physique. La Cybernétique

33 - GABOR D. (1946 Jour. I.E.E. (93) P.429

34 - GABOR D. Phil. Mag (1950) (41) P.1161

35 - GOOD I.J. (London 1950) Probability and the weighting of evidence

36 - GOOD I.J. (1952) J. Roy. Stat. Soc. (14) P.107

37 - HALMOS P.R. and L.J. SAVAGE (1949) Ann. Math. Stat.vol. 20. P.225

38 - HAMMERSLEY (1950) J. Roy. Stat. Soc. (12.B.) P.192

39 - HAMMING R.W. (1950) Bell Syst. Tech J. (29) P.147

40 - HARTLEY R.V.L. (1928) Bell Syst Technicol J. 7. P.535

41 - HELMERT P.R. (1875) Zeit F Math Phy 20.P.300

42 - HUFFMAN (1952) Proc. Lond Symp

43 - HUGGINS (1952) Proc. Lond. Symp

44 - KARHUNEN K. (1949) Arkiv.f.Mat.(I) P.141

45 - KOLMOGOROF A. (1942) Bull Ac. Sci.U.R.S.S. (série Math) (5)

46 - KOOPMAN B.C. Trans.Am. Math. Soc. (39) P.399

47 - KULLBACK S. and R.A. LEIBLER (1951) Ann. Math. Stat. Vol. 22 P.79

48 - KULLBACK S. (1952) Ann. Math. Stat. Vol. 23. P.88

49 - KUPFMULLER K. (1924) Elek Nachtechn I.P.141

50 - LEAVITT H.J. (1951) J. Ab. Soc. Psy. (46) P.38

51 - LEMOINE E. (1902) Géométrographie. Paris.

52 - MAC KAY D.M. (1950) Phi.Mag (41) P.284

53 - MAC KAY D.M. (1950) Proc. Symp.Lond. P.9.

54 - MAC KAY D.M. (1950) Proc. Symp. Lond. P.162

55 - MAC KAY D.M. (1951) Basic Symbols "Cybernetics" (J. Macy Conference)

56 - MANDELBROT B. (1952) Thèse (Paris 1952)

57 - MATHERS K. (London 1932) Measurement of linkage in heredity.

58 - MOURIER E. (1946) C.R. Acad. Sci. Vol. 223 P.712

59 - NEYMAN J. and E.S. PERSON (1933) Proc. Camb.Phil.Soc. 29 P. 492

60 - NEYMAN J. and E.S. PERSON (1933) Phil. Trans. A.231-289

61 - NEYMAN J. (1937) Phil. Trans. A.236. P.333

62.- PEARSON K. (1900) Phil. Mag. (50) P. 157

63 - PEARSON K. (1922) Biometrika (14) P.186

64 - PENROSE L.S. (1947) Ann. Eug (3) P.228

65 - PITMAN E.J.G. (1937 Proc.Camb.Phil.Soc (33) P.212

66 - POLYA G. (1949) Proceedings of the Berkeley symposion Math Stat. and Prob P. 115-123

67 - RAO C.R. (1948) J. Roy. Stat. Soc. (10B) P.159

68 - RICE S.O. (1950) Bell Syst Techn J. 29. P.60

69 - RIGUET J. (Thèse Paris 1951) Théorie des relations binaires

70 - SCHRODER (1871) Mar. Ann. (3) P.296

71 - SCHUTZENBERGER M.P. (1951) C.R.Ac.Sc. 233 P.925

72 - SCHUTZENBERGER M.P.(1951) C.A.Ac.Sc.233. P.1805

73 - SHANNON C.E. (1948) Bell Syst Techn J. Vol.27 P.379-423 et 623-656

74 - SHANNON C.E. (1950) Proc. Lond. Sym P.102

75 - SHANNON C.E. (1950) Proc. Lond. Sym P.105

76 - SMITH C.A.B. (1947) Ann. Eug. (3) P.272

77 - SZILARD L. (1929) Zeitschrift f. Phg 53. P.840

78 - TULLER W.G. (1949) Proc. Ins. Rad. Eng. 37 P/468

79 - VALLEE R. (1951) C.R. Ac.Sc. 233 P.1580

80 - VILLE J. (1948) Cable et Trans (2) P.61

81 - VILLE J. (1950) Cable et Trans (4) P. 9

82 - VILLE J. (1951) Cable et Trans (4) P.76

83 - VILLE J. (1951) Cable et Trans (5) P.126

84 VILLE J. (1951) Cable et Trans (5) P.189

85 - VILLE J. et SCHUTZENBERGER M.P. (1951) C.R. Ac. Sc. (232) P. 206

86 - WALD A. (N.Y. 1947) Sequential analysis

87 - WALD A. et WOLFOWITZ (1948) Ann. Math. Stat. 19

88 - WALD A. (N.Y.1950) Statistical decision fonctions.

89 - WEAVER (1949) Sc. Am. 181 P.I

90 - WELCH B.L. (1939) Ann. Math. Stat. (31) P.218

91 - WIENER N. Cybernetics (Paris 1948).

92 - WIENER N. The extrapolation ..... etc (New-York 1949).

93 - KHINTCHIN A.J. Ucpexi Mat. Nauk. (VII). 1953. p. 320.

TROISIÈME PARTIE

LES MÉTHODES DE GROUPAGE

I. - PROBLÈMES ET MÉTHODES

LES MÉTHODES DE GROUPAGE

Les méthodes que nous aborderons dans ce travail semblent n'avoir jamais été l'objet d'une étude systématique malgré la diversité des cas où l'on en fait un usage plus ou moins empirique (1). Nous nous proposons donc de définir exactement les hypothèses de base, de classer les problèmes possibles et finalement soit d'indiquer des solutions, soit de développer des méthodes générales permettant de faciliter leur recherche.

Quelques exemples permettront de rendre plus clair la discussion des hypothèses.

1° - Circuits électriques : Un appareillage électrique est hors d'état de marche. Pour localiser la "panne" on subdivise en sous-circuits que l'on "sonne" successivement. Le résultat de chaque épreuve permet de conclure, soit que tous les éléments constituant le circuit "sonné" sont en état, soit que l'un d'eux, au moins, est défectueux.

2° - Calculs numériques : Une série de calculs numériques est telle que l'on possède une méthode permettant de décider pour certaines suites d'opération si elles sont toutes correctes ou si au moins l'une d'elle est erronée.

3° - Epreuves biologiques : Ayant fait une série de prélèvements, on les homogénéise et on effectue un seul ensemencement Moyennant diverses conditions de nature biologique un résultat négatif de l'ensemencement permet de conclure à la stérilité de tous les prélèvements.

4° - Test chimiques ou sérologiques : La même possibilité existe si l'on est dans un cas tel que la sensibilité du réactif soit suffisante pour donner une réaction positive, même si un seul des composants du mélange était "marqué".

---

(1) La seule référence que nous ayons pu retrouver dans la littérature est un travail de R. DORFMANN, (the detection of defective members of large population (1943). Ann. Math. Stats.(14) p.436-440)

5° - Observation de phénotypes en génétique mendélienne : étant donné une paire d'allèles (A, a) d'un même gène à dominance et pénétrance complètes, l'observation pure et simple d'un individu permet, si son phénotype correspond à l'allèle récessif a, d'être sûr que son génotype est $\frac{a}{a}$ (qu'il est "homozygote récessif") ou au contraire, si son phénotype correspond à l'allèle dominant A, de savoir seulement que son génotype est $\frac{A}{a}$ ou $\frac{A}{A}$ sans pouvoir trancher entre ces deux dernières éventualités.

LES HYPOTHÈSES DE BASE

Dans tous ces exemples se retrouvent à divers degrés les mêmes particularités que nous discuterons d'abord ici avant de les formuler rigoureusement dans le chapitre suivant sous forme d'un modèle mathématique.

1° - Les observations sont d'une nature telle qu'une observation unique peut éventuellement apporter une certaine information (1) sur plusieurs objets à la fois.

2° - Cette information est par nature fournie de manière dissymétrique en ce sens que les divers résultats d'une même épreuve en apportent des "quantités" différentes voire même incomparables au niveau purement intuitif où nous nous plaçons ici : pour le médecin par exemple, savoir que tous les membres d'un groupe ont un B.W. négatif ce qui assure qu'ils n'ont pas de syphilis secondaire est un renseignement clinique utile. Ce serait pratiquement ne rien apprendre que d'avoir seulement la preuve que l'un d'eux a un B.W.positif si l'on devait continuer à ignorer qui d'entre eux est malade.

3° - Le renseignement que l'on sollicite de l'expérience est de nature dichotomique ou dépend de caractères dichotomiques Ce calcul est-il juste ou non ? - Ces prélèvements sont-ils septiques ou non ? Ou bien, mais ce n'est là,comme nous le verrons, qu'un stade supérieur de complexité du même phénomène : combien de ces appareils électriques sont hors d'état ? - Quelle est la fréquence de cet allèle dans la population ?

4° - L'observation ne fournit à chaque fois qu'une réponse par oui ou par non, donc dichotomique elle aussi.

5° - Le facteur aléatoire n'intervient pas dans le processus d'observation lui-même mais seulement, par définition, en tant que caractérisant notre ignorance partielle ou complète des objets sur lesquels portent nos investigations.

6° - Pour autant qu'elles s'adressent à des objets différents, les observations sont indépendantes ou plutôt ne sont liées entre elles que par les paramètres qui ou bien,sont donnés explicitement (par exemple quand on connaît déjà la proportion

(1) Nous prenons ici le mot "information" dans son acception la plus vague. On verra ultérieurement comment la spécification de concept joue un rôle essentiel dans la solution de ces problèmes.

des appareils électriques défectueux dans un lot qu'il s'agit de trier) ou bien, constituent le but même de la recherche comme dans le cas de l'estimation de la fréquence inconnue d'un gène dans une population.

Nous allons maintenant reprendre un à un ces points et tout en donnant une formulation mathématique des hypothèses, discuter les raisons de notre choix.

La propriété décrite en I° revient à dire qu'au lieu d'observer des variables, on observe des fonctions de plusieurs de ces variables ce qui est une démarche fréquente en statistique mathématique comme en témoignent, par exemple, les applications de l'analyse de variance à la technique des pesées. (Cf. Hotelling (1944) Ann. Math Stat (15) p 297-307).

Cependant la seconde particularité introduit un élément nouveau qui n'apparaît pas dans les méthodes auxquelles on vient de faire allusion puisque celles-ci au contraire reposent sur l'homogénéité des différentes mesures et, partant, sur la possibilité de leur substituer des formes linéaires qui leur soient équivalentes. Cette dissymétrie est donc une caractéristique essentielle de notre problème et c'est précisément à elle qu'est due la possibilité d'économiser éventuellement des observations. Il y a là une analogie qui n'est peut être pas que superficielle avec les problèmes mathématiques de "type réel" où la constatation que la valeur d'une forme quadratique définie positive est plus petite ou égale à zéro, permet de conclure que toutes les variables qui y figurent sont nulles alors qu'on ne saurait que fort peu de choses sur elles si l'on connaissait seulement la valeur numérique (différente de zéro) de cette forme.

Les conditions 3° et 4° s'imposent tout naturellement dans les exemples que nous avons donnés ; il importe peu en effet qu'un calcul soit faux de telle ou telle manière une fois que son inexactitude est établie. Ce n'est de même qu'à un stade ultérieur de l'étude clinique que l'on aura besoin d'évaluer la quantité exacte de réagine contenue dans le sang d'un malade soupçonné de tréponématose. La pratique médicale courante dans ce cas est d'ailleurs de faire d'abord un test dit "qualitatif" puis, seulement si ce dernier est "positif" de procéder aux opérations sensiblement plus coûteuses en temps et en argent de ce qu'on appelle la "quantitation". D'autre part remarquons que dans certains cas il existe simultanément plusieurs classifications dichotomiques. Par exemple dans une étude par agglutination des groupes sanguins dits "classiques" de LANDSTEINER on a les diverses possibilités suivantes; avec un sérum contenant les agglutinines "alpha" et "béta" on peut classer les sujets en " O " (dont les hématies ne contiennent ni l'agglutinogène A ni l'agglutinogène B) et "non O" (dont les hématies contiennent A, B, ou A et B).

Dans ce cas mettant à part diverses considérations sérologiques extérieures à notre propos, il est théoriquement possible de vérifier d'un seul coup qu'un lot de sujets est constitué exclusivement de donneurs universels O (pas d'agglutination observée en testant un mélange de leurs sangs) ou contient au moins un sujet (A) ou (B) ou (A B).

La dichotomie est donc (O)/ (A) (B) (AB).

Par contre avec un serum contenant seulement de l'agglutinine "alpha" la dichotomie serait : (O) (B)/ (A) (AB) c'est-à-dire que l'observation d'un mélange de sang permettrait de conclure :

- soit que tous les sujets sont O ou B (pas d'agglutination)

- soit qu'un sujet au moins est A ou AB (agglutination). De même un serum "béta" conduirait à la dichotomie:

(O) (A) / (B), (AB)

Nous ferons toujours l'hypothèse qu'il s'agit de la même dichotomie à l'intérieur de chacun des problèmes que nous étudierons. Le cas où des systèmes plus complexes d'observations existeraient est d'ailleurs justiciables sinon des mêmes méthodes tout au moins,dans certaines limites, de leur extension. Il est cependant nécessaire d'introduire une nomenclature systématique qui caractérise, pour le mode d'observation employé, les alternatives qui sont en considération.

Selon la suggestion qui nous a été faite par notre Maître Monsieur le Professeur R. TURPIN, nous dirons qu'un objet est "marqué" s'il est possible de dépister en une seule observation sa présence dans un lot et qu'il est "neutre" si, au contraire, une seule observation permet seulement de le prouver tel quand tous les autres objets du lot sont "neutres" aussi.

En ce qui concerne les observations, nous dirons que l'une d'elles a fourni un résultat "positif" quand elle a permis de prouver l'existence d'au moins un objet "marqué"dans le lot; son résultat serait dit "négatif" dans le cas contraire.

Le tableau suivant résume et explique cette terminologie dans les exemples que nous avons donnés plus haut :

Type d'observation	Objets "marqués"	Objets "neutres"
I° Vérification électrique	Défectueux	Normaux
2° Vérification des calculs	Faux	Corrects
3° Ensemencements	Septiques	Aseptiques
4° Tests sérologiques ou chimiques	Présence du réacteur	Absence du réacteur
Groupes sanguins O, A, B, AB		
Sérum α, β	A,B,AB	O
Sérum α	A, AB	O, B
Sérum B	B, AB	
5° Phénotype	Allèle A (dominant)	Allèle B (Récessif)

Une formulation mathématique achèvera de préciser s'il en était besoin, ce que nous entendons par "marqué", "neutre", " positif", et "négatif" bien que malheureusement, l'emploi de ces

deux derniers termes, qu'impose à peu près nécessairement la pratique chimique et sérologique, soit ici fort mal adaptée.

Etant donné un ensemble fini de variables finies, supposons que l'on se limite à l'observation de leur produit; s'il est nul, nous dirons que le résultat est "positif" (!) parce qu'au moins l'une des variables est nulle (c'est-à-dire "marquée"), s'il ne l'est pas, nous dirons que le résultat est "négatif" (!) parce que ceci prouve que toutes les variables sont différentes de zéro (c'est-à-dire toutes "neutres").

Insistons enfin sur le fait que théoriquement on peut concevoir des cas où deux systèmes d'observations permettraient de faire jouer un rôle symétrique aux catégories "marquées" et "neutres". Par exemple, si les variables x pouvaient prendre exclusivement les valeurs 0 et 1, l'observation du produit des (x_i-x) au lieu de celui des x_i échangerait les deux possibilités. En effet si le premier produit n'était pas nul, ceci impliquerait que tous les x sont nuls. Toutefois, dans la pratique, il en est rarement ainsi puisque presque toujours la dichotomie qui sert de base aux observations résulte en quelque sorte de l'isolement d'un point critique dans le champ continu où varient les propriétés physiques des objets. Nous n'envisagerons donc pas cette possibilité, qui n'est en définitive qu'un cas très particulier de cette autre plus générale, que nous discuterons plus loin et qui est l'existence d'une méthode d'observation permettant de déterminer en une seule observation le nombre d'éléments "marqués et "neutres" composant un lot.

Il nous faut donc discuter les raisons pour lesquelles nous n'avons pas envisagé ce mode d'observation. Pour cela considérons de quelle manière celles-ci pourraient être effectivement réalisées. Pratiquement la seule possibilité serait de mesurer d'un seul coup la valeur d'une certaine fonction des variables x_i caractérisant chacun des objets du lot. Le choix de cette fonction impliquerait donc d'abord la nécessité de se donner un type d'additivité. Mais en outre il faudrait que chacune des variables x ne fut susceptible que de prendre deux valeurs ce qui est une hypothèse très artificielle, car sinon l'on se trouverait ramené à un problème classique d'estimation de variables numériques par la donnée numérique aussi de certaines fonctions en ces mêmes variables, problème qui est entièrement hors de notre sujet.

Les mêmes remarques s'appliquent également à la propriété annoncée dans 5° : outre le fait qu'il est difficile de concevoir un modèle statistique assez général pour couvrir une classe suffisamment vaste de cas où une aléatoire peut perturber les relations déterministes strictes :

Un objet "marqué" équivalent à un résultat positif de l'observation

Tous les objets "neutres" équivalents à un résultat négatif de l'observation

il semble bien que le problème ne devienne alors le problème général de la statistique mathématique dans un cas très particulier.

Nous nous bornerons donc à ce schéma qui est, lui aussi, suffisamment vraisemblable dans la plupart des situations pratiques. Nous formulerons cependant pour quelques cas précis des résultats qui généralisent à des modèles aléatoires bien fixés, certains des énoncés valables pour le modèle déterministe.

Enfin, il n'est pas inutile de souligner que plus généralement, il est possible de poser un problème avec une observation dépendant stochastiquement des particularités des objets comme un problème de type déterministe mais muni d'une structure plus compliquée de l'ensemble des objets ou, plus précisément, de l'ensemble des observations que l'on peut effectuer sur eux. Ceci très typiquement est le cas des observations de la génétique mendelienne même quand on inclut dans les schémas d'observation le processus complexe qui consiste à croiser deux individus et à observer leur descendance. Puisque l'observation des phénomènes dans le cas limite, mais pratiquement très général, où nous nous sommes placés permet seulement de savoir si des individus sont $\frac{a}{a}$ ou s'ils sont $\frac{A}{a}$ ou $\frac{A}{A}$, tout croisement ultétieur permettra seulement de découvrir que certains qui présentaient un phénotype correspondant à l'allèle dominant **A**, étaient en réalité hétérozygotes (c'est-à-dire étaient $\frac{A}{a}$). Rigoureusement parlant, il ne sera jamais possible de faire la preuve au sens de la logique formelle qu'un sujet est $\frac{A}{A}$ mais seulement de formuler une assertion statistique du type courant "il y aurait au moins x chances sur cent, mille, dix mille... que ces phénotypes aient été observés si tel ancètre avait eu le génotype $\frac{A}{a}$". C'est ce qui se rencontre en particulier dans l'analyse des problèmes d'exclusion de paternité qui relèvent du même schéma logique et où la génétique peut au plus exclure un progéniteur présumé sans pouvoir formellement conclure en faveur d'une paternité certaine.

En général, il faudrait remplacer les objets élémentaires, par des urnes pour réaliser par un artifice un modèle déterministe ayant le même comportement que le modèle aléatoire que l'on désire étudier. Ici, comme on s'en convraincra sans peine, il ne peut s'agir que de cas d'espèces car le problème fondamental est de préciser la nature de l'additivité de caractères "marqué". En effet si l'on admet qu'il y a seulement une certaine probabilité pour que la réaction soit "positive" quand le lot contient un objet marqué, il est peu plausible de supposer que celle-ci soit la même quand tous les objets le sont, d'où l'introduction d'une expression décrivant cette probabilité en fonction de la composition du lot. Nous espérons que les méthodes générales décrites ici permettront d'aborder les cas précis où il serait à la fois nécessaire d'introduire cette fonction et possible de spécifier son expression.

Pour en terminer avec la signification et le rôle des probabilités dans notre problème il nous faut signaler qu'on supposera presque toujours, sinon connue à l'avance, tout au moins grossièrement estimée, la fréquence des objets "marqués". Comme dans l'analyse séquentielle de WALD avec laquelle nos problèmes présentent beaucoup d'analogies, c'est justement cette connaissance préalable qui permet d'augmenter l'efficacité d'un système d'observation. Un problème typique dans ce sens est celui de l'estimation de la fréquence inconnue des objets marqués dans

une population infinie. Il est évident que si l'on ignore totalement son ordre de grandeur entre 1/2 et $1/10^6$, la seule chose à faire est d'obtenir une estimation préalable sur un premier échantillon qui servira ensuite à choisir les paramètres fixant la deuxième série d'observations qui pourra celle-ci s'effectuer selon un "design" optimal.

Reste enfin la dernière caractéristique qui se justifie comme les précédentes par le fait :

- qu'elle est pratiquement réalisée dans les cas importants
- qu'elle contient les autres possibilités en ce sens qu'un modèle simple du type "dépendant" peut être ramené à un modèle "indépendant" à structure complexe.
- que son abandon impliquerait le choix d'une loi de dépendance qui ne saurait être que plus arbitraire encore.

Signalons enfin parmi les problèmes analogues aux nôtres et justiciables sans doute de méthodes parallèles celui où au lieu d'objets "neutres" et "marqués" existeraient des objets disons de type I et de type II et où l'observation serait positive quand et seulement quand le lot contiendrait simultanément au moins un objet I et un objet II. Une illustration d'un tel cas pourrait encore être fournie par la recherche de nouveaux groupes sanguins quand manquent à la fois des hématies porteuses de l'agglutinogène cherché et une réserve de sérum contenant l'agglutinine correspondante.

LES TYPES DE PROBLÈMES

Toujours sur le plan de la discussion intuitive, il est nécessaire d'indiquer les principaux problèmes qui peuvent se poser en présence d'un système du type décrit; nous en isolerons quatre, relevant deux par deux de domaines différents du calcul des probabilités.

Tout d'abord les deux problèmes classiques de la statistique mathématique :

l'estimation d'un paramètre inconnu, typiquement, la fréquence des objets "marqués" dans une population infinie

le test d'une hypothèse assignant à cette probabilité de se trouver dans un certain intervalle.

Naturellement, dans ces deux cas, il ne s'agira ici que d'appliquer les méthodes générales connues en profitant des particularités structurales envisagées plus haut pour diminuer à précision ou risque d'erreur égaux, le nombre des observations nécessaires. Dans certains cas on arrivera à des résultats extrêmement substantiels.

Les deux autres problèmes par contre sont en réalité distincts de ceux de la statistique mathématique habituelle quoi qu'il soit possible de les y ramener par un biais très artificiel à vrai dire.

Ce sont :

les problèmes de diagnostic : étant donné un ensemble fini d'objets, déterminer pour chacun d'eux s'il est "neutre" ou "marqué".

les problèmes de tri : étant donné un ensemble fini ou non, en extraire un nombre m. fixé à l'avance d'objets "marqués" ou bien n objets "neutres" ou bien encore à la fois m objets "marqués" et n objets "neutres".

Un lien entre ces deux derniers exemples et les problèmes relevant de la statistique mathématique serait le comptage des objets "marqués" dans un lot fini, problème statistique si l'on se contente d'une réponse "en probabilité", mais problème du second type si l'on veut une réponse catégorique. De fait, le problème de comptage ne se pose pas parce que, comme on le verra plus loin, il est impossible, du fait même de la structure logique du modèle d'observation, de "compter" les objets "marqués" sans les repérer individuellement. Ce problème se résume donc exactement au problème de diagnostic. Cet exemple cependant fait apparaître assez nettement le contraste entre les deux groupes de problèmes qui résulte à la fois de la nature déterministe des observations et du caractère discontinu (dichotomique même !) à la fois des grandeurs observées et des états inconnus des objets. Il apparaît donc que les problèmes du deuxième groupe rentrent plutôt dans le cadre des questions étudiées par la théorie de l'information. Le cas de diagnostic est d'ailleurs exactement, comme nous l'avons déjà dit dans la deuxième partie, une modalité particulière du problème du codage; les problèmes de tri sont, semble-t-il, nouveaux.

LES TACTIQUES

Ayant ainsi délimité les problèmes que nous aborderons, il nous faut enfin dire quelques mots sur la nature des solutions auxquelles nous nous efforcerons d'aboutir.

L'idéal serait de pouvoir, dans chaque cas, donner explicitement l'expression de la "décision fonction" de WALD ou de la "stratégie minimax" au sens de VON NEUMANN qui garantit à l'expérimentateur, les meilleurs résultats pour le coût en moyenne le plus faible.

De fait, la complexité combinatoire des problèmes et la nécessité de rester dans certaines limites de simplicité pour les procédures d'expérimentation nous conduira, selon une démarche très fréquente d'ailleurs dans les mathématiques appliquées, à substituer au concept de "stratégie optimale" celui de "tactique optimale" que nous allons définir.

Soit un certain but à atteindre, qui est aussi bien ici le diagnostic ou le tri d'un ensemble d'objets que la réduction en dessous d'un certain niveau de la variance d'estimation ou des probabilités d'erreur de première ou de deuxième espèce. Nous symboliserons ce but par un point, O, dans un certain espace B.

Quel que soit le problème, une procédure est une suite, menant au but proposé, d'opérations élémentaires, dont chacune a un coût fixé à l'avance. La question qui se pose est celle de trou-

ver la où les procédures telles que la somme correspondante des coûts soit minimum.

Du point de vue qui nous intéresse nous pouvons encore schématiser ceci par une feuille de chemins Ci menant de l'origine I au but O et une fonction f (P) des points de B. Le coût (ou plutôt sa valeur moyenne, mais il est équivalent d'employer ou non un langage déterministe), correspond alors simplement à l'intégrale de f (P) prise le long de C_i et le problème est de trouver le ou les chemins qui rendent celle-ci minimum.

Les méthodes générales du calcul des variations permettent alors le raisonnement suivant ; supposons que pour chaque point P la valeur minimum h (P) de l'intégrale précédente prise de P à O soit connue. Les surfaces h (P) = C^{te} constituent une famille dépendant d'un paramètre (ce sont les "transversales"), et les chemins C optimaux en sont les trajectoires orthogonales ce qui permettrait de les déterminer.

Ceci est d'ailleurs intuitif puisque le fait de suivre ces trajectoires orthogonales revient à rendre maximum pour chaque déplacement infinitésimal la décroissance du coût total correspondant au chemin qui reste à parcourir.

La méthode reste valable quand, au lieu d'une trajectoire continue, on a affaire à une suite discrète de choix et nous avons pu montrer que dans certains cas particuliers elle présentait l'avantage de nécessiter le plus petit nombre d'opérations arithmétiques. Pour l'appliquer, on part des positions voisines du point O et on calcule pour chacune d'elles le coût minimum nécessaire pour atteindre le but ; puis successivement on effectue cette détermination pour toutes les positions qui sont voisines des précédentes, puis pour celles qui sont voisines de ces dernières ... etc ; jusqu'à ce qu'on soit remonté au point de départ I.

L'économie de calcul résulte évidemment du fait que l'on a jamais à tenir compte pour les comparer entre eux que des chemins qui sont déjà optimaux pour leur portion qui avoisine O ce qui réduit grandement le nombre des opérations.

Une semblable méthode a probablement été employée de façon implicite dans de nombreux cas et nous nous bornerons à citer celui du codage binaire d'un système quelconque de messages (Cf. D. HUFFMAN - 1952 - A method for the construction of minimum rédundancy codes - Symp on comm. application. London).

Mais, en général, la détermination de la fonction h (P) entraîne un travail considérable et l'on peut essayer de lui substituer une autre fonction k (P) qui soit plus facile à calculer.

Revenons au cas discret et relativement a un certain choix d'une fonction k (P) nous aurons alors ce que nous appellerons une tactique (optimale) d'ordre zéro ; celle-ci consistera pour chaque point P à passer au point P' tel que la décroissance de k (P) soit la plus grande possible, tactique qui serait rigoureusement optimale, comme on l'a vu, si k (P) était précisément h (P).

On définirait de même des tactiques d'ordre un par le choix à chaque position P du point suivant P' tel qu'il permette lui-même de passer en un deuxième temps à P'' avec k (P) - k (P'')

maximum, des tactiques d'ordre deux où P' serait choisi en fonction de k (P) - k (P''') où P''' peut être atteint à partir de P'' ..... Naturellement cette méthode serait rigoureusement optimale quel que soit k (P) si une tactique d'ordre assez élevé pouvait être utilisée.

L'analogie de ces méthodes, qui semblent d'ailleurs appeler de nombreux travaux, avec la conduite effective des individus dans des situations telles que le jeu d'échecs, est évidente : à chaque position le joueur calcule le mouvement qu'il va effectuer en fonction du gain tactique qu'il lui assurera un, deux, trois, ...n coups plus tard.

Sans pouvoir insister sur cette question, que nous avons développée ailleurs d'un autre point de vue (A tentative classification of goal seeking beaviours. J of Ment.Sci.-Octobre 1953) indiquons rapidement quelques exemples d'emploi d'une tactique d'ordre zéro dans divers domaines.

Problème de construction d'un circuit à contacts approximativement le plus économique pour réaliser une fonction logique donnée (Cf GAVRILOV. Teoria Releinokontaktnyx cxem. Moskva 1950. p. 185).

Problème des diaphonies de J.VILLE -(les (a_i) étant donnés, il s'agit de choisir les signes + ou - de manière à minimiser $(\Sigma \pm a_i)$ - (Cf Variables aléatoires equiparties (1949) Cables et transmissions - p 262-274).

Problème de détermination de la permutation des colonnes d'une matrice donnée rendant maximum la somme des éléments de la diagonale.

"Problème du livreur" : c'est-à-dire choix de l'itinéraire le plus court qui passe par un ensemble de points dont les distances mutuelles sont connues.

Enfin il semble que dans le domaine des calculs numériques de nombreuses méthodes d'itération reposent aussi sur le même principe que l'on pourrait appeler d'optimalité locale.

Certaines fonctions k (P) correspondent à ce que serait effectivement une stratégie optimale au sens strict si les paramètres caractéristiques du problème étaient choisis au fur et à mesure par un adversaire tel que le définit la théorie des jeux, au lieu d'être donnés à l'avance.

Ainsi, dans le problème classique généralisé des courbes de poursuite, la tactique d'ordre zéro est basée sur la fonction k (P) qui est simplement la distance du poursuivant à l'objectif et elle consiste à choisir une trajectoire dont la tangente passe à chaque instant par l'objectif. Cette tactique serait optimale si celui-ci, dépourvu d'inertie, voyait sa position soumise à des fluctuations aléatoires et indépendantes. La tactique d'ordre un tiendrait compte de l'existence de l'inertie et interprèterait les mouvements de l'objectif comme résultant d'une chaîne de MARKOV d'ordre un elle aussi ... etc ...

En définitive, le problème se trouve ramené au choix de la fonction k (P); mises à part des propriétés qualitatives assez peu caractéristiques : (par exemple k (0) = 0), nous ne savons pour ainsi dire rien à ce sujet dans le cas général puisque le critère essentiel qui conduit à préférer k (P) à h (0) est celui

de la simplicité du calcul, qu'il est difficile d'expliciter en dehors des cas particuliers. Notons cependant qu'on a employé souvent, de manière implicite d'ailleurs, la méthode qui consiste à remplacer h (P) par une fonction simple qui lui soit équivalente dans certains cas où les paramètres du problème ont des valeurs particulières, correspondant le plus souvent à des cas extrémaux.

Revenons à notre problème après cette longue parenthèse. Pour les mêmes raisons de simplicité qui nous ont guidé dans le choix des hypothèses, nous prendrons pour coût le nombre moyen d'observations élémentaires effectuées quel que soit le nombre des objets sur lesquelles elles portent. Nous pouvons déjà indiquer que pour chaque type de problème considéré, les fonctions k (P) seront les informations appropriées : en effet d'après ce que nous avons vu dans la deuxième partie, celles-ci sont susceptibles d'être calculées simplement et pour certaines valeurs, des paramètres elles coïncident avec les fonctions h (P) en vertu des théorèmes d'optimalité. De plus grâce à ses mêmes théorèmes limitatifs on sera souvent en mesure de prouver que les tactiques d'ordre zéro conduisent dans l'ensemble à des résultats qui ne s'écartent pas trop des résultats strictement optimaux.

II. - PROPRIÉTÉS GÉNÉRALES DU MODÈLE

DÉFINITIONS ET NOTATIONS

L'ensemble (fini ou infini) des objets sera désigné par E et les objets eux-mêmes par a, b, c ..... ou bien selon les cas par a_1, a_2, ... a_n.

Le fait qu'un objet "a" soit à l'état marqué (respectivement neutre) sera noté par a^+ (respectivement : a^-).

Nous désignerons par A = (a, b, c .....)

B = (x, y, z...) ou bien A_i les ensembles d'objets susceptibles d'une observation élémentaire permettant de savoir s'ils sont tous "neutres" ou si l'un deux au moins est "marqué".

Ainsi A désignera à la fois un ensemble (a, b,...) et l'observation élémentaire portant sur A. Le plus souvent, tous les sous-ensembles de E seront susceptibles d'être observés; dans d'autres cas, certains seulement qui seront dits "admissibles". Par exemple en génétique mendelienne, les ensembles admissibles sont certaines paires d'allèles, celles appartenant à un individu de l'échantillon dont elles constituent le génotype. Il sera commode d'utiliser la notation $A^{(n)}$ pour désigner un ensemble quelconque formé de n objets et on dira qu'un ensemble A est "libre" si aucun des objets qui y figure n'a encore appartenu à un ensemble observé. Enfin, l'évènement constitué par le fait que l'observation de A a donné un résultat positif (respectivement : négatif) sera simplement écrit A^+ (respectivement A^-).

Pour tout objet a, b, c ou x_i la probabilité qu'il soit à l'état marqué sera représentée par la lettre grecque correspondante sauf, pour des raisons typographiques évidentes, dans le cas où toutes ces probabilités élémentaires ayant la même valeur commune celle-ci sera notée par p = 1 - q.

LES TREILLIS DES POSITIONS

Après ces définitions, nous pouvons passer à l'étude des principales propriétés mathématiques du modèle et tout d'abord au calcul de la probabilité du résultat d'une observation X quand on connaît déjà les résultats d'un système $\mathcal{A}$ d'observations A,B..

Pour cette étude, le résultat suivant est essentiel.

Théorème I - Etant donné un système d'observations, A, B ... C ayant porté sur les objets d'un ensemble E, les résultats de celles-ci sont décrits intégralement par les données :

I° - D'un élément $\mathcal{A}^+$ du <u>treillis distributif</u> libre T dont les générateurs sont les objets de E.

2° - D'un sous-ensemble A^- de E, $\mathcal{A}^+$ et A^- ne contenant <u>aucun objet en commun</u>.

Considérons en effet d'abord l'ensemble $\mathcal{A}^-$ de toutes les observations de $\mathcal{A}$ qui ont donné un résultat négatif. D'après les propriétés déterministes du modèle, un objet ne figure dans une observation de $\mathcal{A}^-$ que s'il est à l'état "neutre". Réciproquement si tous les objets x, y, z...sont à l'état "neutre" toute observation sur un sous-ensemble d'entre eux donne un résultat négatif. Par conséquent, a posteriori, $\mathcal{A}^-$ est équivalent à une seule observation élémentaire A^- ayant donné un résultat négatif.

Toujours en raison du caractère déterministe du modèle, si l'observation X = (a b c...) a donné un résultat positif, elle est a posteriori équivalente à l'observation X' = (a b c ...) où X' se déduit de X par suppression des objets appartenant à A .

Nous pouvons donc réduire $\mathcal{A}$ à $\bar{A}$ et à $\mathcal{A}^+$ tels qu'aucun objet ne figure à la fois dans $\bar{A}$ et dans une observation de l'ensemble d'observations $\mathcal{A}^+$.

Pour achever la démonstration, il nous suffit maintenant de montrer que si $\mathcal{A}^+$ contient deux observations A et B telles que l'ensemble A soit un sous-ensemble de B on peut sans rien changer supprimer B de $\mathcal{A}^+$.

Mais ceci est évident, puisque le fait que A soit positif entraîne qu'il contient au moins un objet "marqué" et, par conséquent que toute observation portant sur un ensemble contenant A est positive et ne saurait rien apprendre de plus sur aucun objet si elle est effectuée après A.

Par une extension bien naturelle d'une terminologie introduite par Monsieur le Professeur G. DARMOIS, on pourrait dire que, dans notre modèle, le <u>résultat positif</u> de l'observation d'un sous-ensemble (et non pas l'ensemble des résultats comme en général en statistique) est un <u>résumé exhaustif</u> du résultat de l'observation d'un ensemble.

Une conséquence immédiate de ce résultat est que si l'observation B où B est un sous-ensemble de A, est effectuée après A , les objets de A-B redeviennent des objets libres c'est-à-dire que l'on ne sait rien de plus sur eux que si A n'avait jamais été faite.

Appelons "position" d'un observateur le couple $\mathcal{A}^+A^-$ qui résume les résultats qu'il a déjà obtenus. D'après ce que nous venons de dire, l'ensemble de ces positions constitue un treillis T_E (le treillis de position) si l'on introduit un ordre "$\prec$" gouverné par les conventions suivantes :

I° - $(\mathcal{A}^+A^-) \prec (\mathcal{A}'^+ A'^-)$ si et seulement si $\mathcal{A}^+ \leqslant \mathcal{A}'^+$ dans T_E et A^- A'^- dans le treillis booléen des parties de E.

2° - Il existe une position ($\emptyset$) antérieure à toutes les autres : c'est celle de l'observateur avant toute observation.

3° - Il existe une position (purement virtuelle d'ailleurs) qui est celle d'un observateur qui aurait obtenu des résultats contradictoires (par exemple A^+ et B^- si A était un sous-ensemble de B).

La relation d'ordre "antérieur à" dans T, coïncide avec l'ordre temporel, physique, des observations.

Enfin, il est important de noter que si $E' \subset E$ et si tous les objets figurant dans $\mathcal{A}^+A^-$ et $\mathcal{A}'^+A'^-$ appartiennent à E', la relation $\mathcal{A}^+$ $A^- \prec \mathcal{A}'^+ A'^-$ dans $\bar{T}_E$ est équivalente à la même assertion dans $\bar{T}_{E'}$.

A titre d'exemple, nous allons montrer comment la dissymétrie des états "marqués"et "neutres" entraîne qu'un résultat positif apporte moins de renseignements qu'un résultat négatif.

Plus exactement :

soit $\mathcal{A} = \mathcal{A}^+ A^-$ la position d'un observateur avant qu'il effectue l'observation élémentaire X. Soit $\mathcal{A}'^+$ A'^- sa position si X^-; $\mathcal{A}''^+A''^-$ sa position si X^+ ; enfin $\mathcal{B}^+$ B^-, sa position s'il apprenait que tous les objets de X sont à l'état marqué (ce que l'on peut noter X^{++}).

C'est-à-dire que $\mathcal{B}^+$ B^- peut s'écrire $\mathcal{B}'^+$ X^{++} B^- où le mot $\mathcal{B}'^+$ ne contient aucun des objets de X . On a :

Dans T_E : $\mathcal{A}'^+$ $A'^- \prec \mathcal{B}^+$ $\mathcal{B}^-$

Dans T_{E-X}: $\mathcal{B}'^+ \prec \mathcal{A}'^-$

ce qui signifie donc que mis à part les objets de X , X^+ conduit à une position antérieure à celle qui résulte de X^-.

En effet :

I° - A'^- est la réunion de A^- et de X^- et $\mathcal{A}'^+$ est obtenue à partir de $\mathcal{A}^+$ en y effacant purement et simplement les lettres qui symbolisent les objets de X.

2° - A''^- et B^- sont identiques à B^- ce qui établit la première relation puisque $\mathcal{B}^+$est la réunion de A''^+ et de X^{++}.

3° - La deuxième relation découle de ce que $\mathcal{A}''^+$est la réunion de $\mathcal{B}'^+$ avec ce qui reste des monomes de $\mathcal{A}^+$après effacement des objets de X .

Il est intéressant à ce stade de notre discussion de dire un mot des affaiblissements du théorème I qu'impliquerait le remplacement du modèle déterministe par certains modèles aléatoires. Nous considèrerons successivement les deux possibilités limites qui se traduisent par une liaison stochastique au lieu d'une implication formelle pour l'hypothèse :

I° - au moins un objet marqué entraîne un résultat positif

2° - tous les objets "neutres" entraînent un résultat négatif

Dans le premier cas, le théorème I se réduit, d'une part à la simple possibilité de "résumer" une série d'observations sur un seul objet par une observation quand au moins une fois un résultat positif a été obtenu. D'autre part, il reste encore la faculté de négliger l'observation d'un lot d'objets quand tous les objets testés individuellement ont fourni un résultat positif.

En effet soit r le paramètre $\Pr(A^+ \mid a^+)$ où A est l'observation de l'ensemble A réduit à a et soit p_n la probabilité que la n-ième observation du même objet a donné un résultat négatif quand les n-1 premières ont été négatives aussi, on a :

$$p_n = \frac{\alpha(1-r)^n + 1 - \alpha}{\alpha(1-r)^{n-1} + 1 - \alpha} \qquad (\alpha = \Pr(a^+))$$

Par conséquent, sur le plan opérationnel où nous nous plaçons c'est-à-dire même si α et r sont connus (et a fortiori si α est inconnu) chaque observation négative nouvelle de a apporte un changement sur la probabilité des résultats d'une autre observation ultérieure.

Au contraire, en vertu du déterminisme (unilatéral) que nous avons supposé conservé il est clair que $A^+ B^+ C^+$ entraîne $(A + B + C \ldots)^+$ avec une certaine probabilité r' qui ne dépend que de la structure stochastique du modèle à condition que A B C .... soient des objets et non groupes d'objets puisque par exemple la connaissance de A^+ et $(B+C)^+$ n'est pas équivalente à celle de A^+ ; $(B + C)^+$ et $(A + B + C)^+$ pour prévoir le résultat de l'observation C .

De façon duale, dans le deuxième cas, seule subsiste la première partie du théorème I, c'est-à-dire que l'on peut seulement remplacer α^- par une observation unique portant sur un certain ensemble A .

Montrons pour finir que dans le modèle déterministe le comptage du nombre des objets marqués d'un ensemble E implique le diagnostic de tous les objets.

Soit en effet $\alpha^+ A^-$ la position d'un observateur sachant que E contient n objets marqués et m neutres.

Par hypothèse A^- est équivalent à une observation élémentaire $A^{(m)}$ ayant donné un résultat négatif et garantissant donc le caractère neutre des m objets qui y figurent.

D'autre part α^+ est équivalent à un système d'observations positives et ne peut que fixer une limite inférieure au nombre n d'objets marqués à moins que toutes les opérations élémentaires qui le composent n'aient porté que sur un seul objet.

LES PROBABILITÉS DES POSITIONS

Nous sommes maintenant en mesure de calculer les probabilités qui nous intéressent. Nous désignerons par $Pr(X|\alpha^+ A^-)$ la probabilité que l'observation élémentaire X donne un résultat positif quand on a déjà obtenu les résultats résumés par le couple α^+ et A^-.

D'après la théorie des probabilités composées on a :

$$Pr\ (X^+|\alpha^+ A^-) \times Pr\ (\alpha^+ A^-) = Pr\ (X^+\ \alpha^+\ A^-)$$

Comme l'ensemble $X^+\ \alpha^+$ est aussi un mot α'^+ de T, le problème est donc ramené au calcul des Pr (A) c'est-à-dire au calcul des Pr (α^+) puisque le fait que α^+ et A^- n'aient pas d'éléments communs entraîne :

$$Pr\ (\alpha^+ A^-) = Pr\ (\alpha^+) \times Pr\ (A^-)$$

Le calcul de Pr (A^-) est immédiat :

$$Pr\ (A^-) = (1-\alpha)\ (1-\beta)\ (1-\gamma)....(1-\xi)$$

si a, b, c ....x $\in$ A.

Pour calculer les Pr (α^+) on dispose du théorème suivant :

Théorème II - Les probabilités Pr (α^+) sont une valuation du treillis T_E.

Le théorème n'est encore une fois que la transcription en langage des treillis des notions fondamentales du calcul des probabilités. Toutefois cette transcription nous permet d'obtenir directement les résultats pratiques suivants qui découlent de la possibilité de calculer les valuations de tous les mots à partir de la seule donnée des valuations des éléments . et + irréductibles du treillis T_E.

1ère méthode :

Soit $\alpha^+ = \sum_{i\in I} A_i^+$ une représentation réduite de α^+ en les observations élémentaires A_i $(i\in I)$ qui sont les éléments + irréductibles du treillis.

D'après les méthodes générales de la théorie des évènements compatibles et dépendants ou - ce qui revient au même - d'après les méthodes de la théorie des valuations dans les treillis distributifs on a :

$$Pr\ (\alpha^+) = \sum_{i\in I} Pr(A_i^+) - \sum_{i,i'} Pr(A_i^+\ A_{i'}^+) + \sum_{i,i',i''} Pr\ (A_i\ A_{i'}\ A_{i''}) + ..$$

ce que l'on peut encore, formuler de la façon suivante :

Corollaire I - Pour chaque sous-ensemble I' de puissance n de I soit B_I' l'observation élémentaire portant sur les objets qui appartiennent au moins à l'un des A_i $(i\in I')$, on a :

$Pr\ (\alpha^+) = (-1)^n\ Pr\ (B_I^+)$ où la sommation est étendue à tous les sous-ensembles I' de I.

Chacun des B_1' étant équivalent à une observation élémentaire on a enfin :

$$Pr\ (B_1^{+\prime}) = 1 - Pr\ (B_1^{-\prime}) = 1 - (1-\alpha)\ (1-\beta)\ldots(1-\xi)$$

si a, b, ... $x \in B_1'$

2ème méthode :

Nous partons maintenant de la représentation de $\mathcal{A}^+$ comme intersection des mots irréductibles.

Soit C_j $(j \in J)$ les ensembles minimaux d'objets tels que $\mathcal{A}^+$ ne sont possibles que si tous les objets de l'un des C sont à l'état marqué ce que nous noterons par $(C_j)^{++}$. En langage de probabilité on a donc :

$$Pr\ (\mathcal{A}^+) = Pr\ (C_1^{++} \text{ ou } C_2^{++} \text{ ou } \ldots\ C_J^{++})$$

d'où :

$$Pr\ (\mathcal{A}^+) = \sum_j Pr\ C_j^{++} - \sum_{jj'} Pr\ (C_j^{++}\ C_{j'}^{++}) + \ldots$$

De manière formelle les ensembles C_j seront obtenus simplement en permutant dans l'expression de $\mathcal{A}^+$ les opérations . et + puis, grâce à la loi distributive, en développant en somme de monomes ces derniers étant précisément les C_j .

La formule précédente s'énonce alors de manière rigoureuse:

Corollaire II - Pour chaque sous-ensemble de puissance n, J' de J soit $C_{j'}$ l'ensemble des objets qui figurent dans au moins un C_j , $(j' \in J')$, on a :

$$Pr\ (\mathcal{A}^+) = \sum (-1)^n\ Pr\ (G_{j'}^{++})$$

où la sommation est étendue à tous les sous-ensembles J' de J.

Enfin si a, b ..... $x \in G_j$, on a par définition :

$$Pr\ (G_{j'}^{++}) = Pr\ (a^+\ b^+\ \ldots\ x^+) = \alpha\beta\ \ldots\xi$$

Exemple : soit E = (abcd) et soient les observations élémentaires $A = (a,b,c)^+$ $B = (a,d)^+$ $C = (b,d)^+$

Nous représenterons simplement le fait que $A^+B^+C^+$ par le mot abc + ad + bd et nous avons, d'après le corollaire I : [Pr $(A^+B^+C^+)$ = Pr (abc + ad + bd) = Pr (abc) + Pr (ad) + Pr (bd) - Pr (abd) - Pr (abcd).

Donc $Pr\ (A^+B^+C^+) = 1 - \bar{\alpha}\bar{\beta}\bar{\gamma} - \bar{\alpha}\bar{\delta} - \bar{\beta}\bar{\delta} + \bar{\alpha}\bar{\beta}\bar{\delta} + \bar{\alpha}\bar{\beta}\bar{\gamma}\bar{\delta}$

en posant $\bar{\xi} = 1 - \xi$ = probabilité que x soit neutre.

De même en employant le corollaire II :

Les combinaisons minimales d'objets "marqués" qui garantissent que $A^+B^+C^+$ sont, comme on le calcule sans peine :

(a et b) ; (a et d) ; (b et d) ; (c et d). On a donc :

A_1 = ab ; A_2 = ad ; A_3 = bd ; A_4 = cd ; et finalement :

$$Pr\ (A^+B^+C^+) = \alpha\beta + \alpha\delta + \beta\delta + \gamma\delta - 2\alpha\beta\delta - \alpha\gamma\delta - \beta\gamma\delta + \alpha\beta\gamma\delta$$

valeur qui est bien égale à celle trouvée à l'aide du corollaire I.

3ème méthode.

Les deux méthodes précédentes conduisent à des expressions contenant des signes négatifs. La théorie des évènements compatibles et dépendants nous livre une troisième représentation qui ne présente pas cet inconvénient éventuel.

Considérons en effet les évènements D_i ("les diagnostics complets") consistant en le fait qu'est connu l'état de chacun des N objets de l'ensemble étudié E. La probabilité de l'un quelconque de ces 2^N évènements est de la forme :

$$Pr(D_i) = \alpha_1 \alpha_2 \ldots (1-\alpha j)\,(1-\alpha j') \ldots$$

et tout α^+ pouvant être représenté de manière unique comme somme logique de ces évènements incompatibles, il s'en déduit bien l'expression cherchée.

On obtient ainsi pour l'exemple précédent :

$$Pr\,(A^+ B^+ C^+) : \alpha\beta\gamma\delta + \alpha\beta\gamma\bar{\delta} + \alpha\beta\bar{\gamma}\delta + \alpha\beta\bar{\gamma}\bar{\delta} + \alpha\bar{\beta}\gamma\delta + \alpha\bar{\beta}\bar{\gamma}\delta + \bar{\alpha}\beta\gamma\delta + \bar{\alpha}\beta\bar{\gamma}\delta + \bar{\alpha}\bar{\beta}\gamma\delta$$

LE CAS DES PROBABILITÉS ÉLÉMENTAIRES ÉGALES

Dans la suite, nous étudierons surtout le cas particulier correspondant à une probabilité d'être marqué p égale pour tous les objets.

A chaque mot α correspond alors un polynome $P(p) = Pr(\alpha^+)$ prenant ses valeurs dans (0,1).

Les propriétés suivantes quoique très simples méritent d'être relevées.

1° - P(p) est une fonction croissante de p (plus généralement $Pr(\alpha^+)$ est fonction croissante des αi).

2° - P(p) admet la racine p = 0 avec un ordre de multiplicité égale au nombre minimum n_0 d'objets dont le caractère marqué assure que α^+ est réalisé.

En effet, d'après le corollaire II, $Pr(\alpha^+)$ est une somme de monomes en les αi correspondant chacun à un ensemble C_j, d'objets tous marqués et les ensembles minimaux de cette famille sont les C_j eux-mêmes. Le coefficient du terme p^{n_0} est donc en outre le nombre de ces C_j de puissance n_0.

3° - 1-P(p) admet la racine p = 1 avec un ordre égal au minimum de la puissance des observations élémentaires qui constituent α^+.

Ceci est immédiat d'après le corollaire I.

4° - P(p) est d'ordre au plus égal au nombre n des objets qui figurent dans α^+. Cet ordre s'abaisse chaque fois que n étant pair α^+ est un mot identique à son dual. $\tilde{\alpha}^+$

Considérons en effet α^+ et l'évènement $\check{\alpha}^-$ défini par le fait que pour qu'aucun des c l'on a (c) . La double implication :

α^+ entraîne non $\check{\alpha}^-$ et $\check{\alpha}^-$ entraîne non α^+

permet d'écrire :

$$Pr\ (\alpha^+) = 1 - Pr\ (\check{\alpha}^-)$$

Mais comme $Pr(\check{\alpha}^-)$ est équivalente d'après les corollaires I et II à $Pr(\check{\alpha}^+)$ où l'on aurait remplacé les i par les $\bar{i}$, nous pouvons écrire si $\alpha^+ = \check{\alpha}^+$, $Pr\alpha(p) + Pr\check{\alpha}(q) = 1$ et le coefficient du terme de plus haut degré doit bien être nul si ce degré est pair.

Enfin, pour les cas où des dispositifs expérimentaux particuliers (circuits électriques par exemple) permettraient de réaliser de façon relativement aisée certains mots complexes, nous rappellerons le résultat suivant bien connu en théorie des treillis :

La condition nécessaire et suffisante pour que E contienne au moins h objets marqués est que l'on ait α_h^+ ou α_h désigne le mot symétrique $\Sigma A^{(N-h+1)}$ avec la sommation étendue à toutes les combinaisons (N-h+1) à (N-h+1) des N objets de E.

En effet le calcul montre que le dual de α_h est précisément α_{N-h+1}. D'après la troisième méthode de calcul, les fonctions correspondantes sont :

$$P_h(p) = \sum_0^{N-h} \left[\begin{matrix} N \\ h+i \end{matrix}\right] p^{h+i} (1-p)^{N-h-i}$$

La fonction $P_h(p)$ a la propriété remarquable de présenter un point d'inflexion unique pour $p = (h-1) / (N-1)$ comme on peut le vérifier par le calcul direct. Quand N et h tendent vers l'infini de telle manière que (h-1) / (N-1) et (N-h) / (N-1) aient des limites finies, $P_h(p)$ tend vers une fonction de saut en ce point ce qui d'ailleurs est intuitif.

III. - PROBLÈMES DE DIAGNOSTIC

GÉNÉRALITÉS

Dans ces problèmes, il s'agit, comme on l'a vu, de déterminer celui qui s'est réalisé des évènements incompatibles D_i consistant en le fait que tous les objets de E ont un état donné

Nous désignerons par Π une procédure, c'est-à-dire une suite déterminée d'observations à effectuer en fonction des résultats des observations antérieures; c'est-à-dire, encore, formellement un arbre à chaque point de ramification duquel est attaché le sous-ensemble X de E qui doit être observé.

Les α_i étant donnés, nous désignerons par $L(\alpha)$ la somme $\Sigma Pr(D_j)\lambda(D_j)$ étendue à tous les D_j postérieurs à α (c'est-à-dire auxquels on peut encore aboutir quand on est en position α) et où $\lambda(D_j)$ désigne le nombre d'observations élémentaires conduisant de α à D_j, en suivant la procédure Π.

Ainsi, par exemple, $L(\emptyset)$ est le coût total moyen de la procédure et inversement $L(D_j) = 0$ pour tout D_i par hypothèse.

Si π prescrit l'observation élémentaire X quand on est en α, on a la relation :

$L(\alpha) = Pr(\alpha) + L(\alpha X^+) + L(\alpha X^-)$ ce qui montre que L est une <u>fonction décroissante</u> sur le treillis de position.

Théorème : Une condition nécessaire pour que π soit optimale est que l'on ait quelque soit α :

$$L(\alpha X^+) \leqslant L(\alpha X^-).$$

Supposons en effet que pour un certain α cette condition ne soit pas vérifiée. Pour la partie de la procédure qui concerne les observations postérieures à αX^-, remplaçons les observations prescrites par π celles que prescrit π quand αX^+. D'après ce qui a été vu dans le chapitre précédent sur la dissymétrie des résultats qu'entraînent respectivement les résultats X^+ et X^- on aura en appelant π' la procédure π ainsi modifiée :

$$L'(\alpha X^-) \leqslant L(\alpha X^+) \leqslant L(\alpha X^-)$$

Donc π' est uniformément préférable à π.

On en déduit la conséquence importante suivante :

Une condition <u>nécessaire</u> pour que π soit optimale est que $\lambda(D_j) \leqslant \lambda(D_{j'})$ pour tout $\alpha = \emptyset$ et pour toute paire $D_j, D_{j'}$, où le second diagnostic ne diffère du premier que par remplacement de l'état neutre par l'état marqué pour certains objets. Dans ces conditions $L(\emptyset)$ est une fonction croissante de chacun des α_i.

En effet π' étant un arbre, il existe pour toute semblable paire D_j, $D_{j'}$ une position α_j et une observation X telles que D_j soit postérieur à $\alpha_j X^-$ et $D_{j'}$ à $_jX^+$. Il en résulte la première partie de l'énoncé puisque π est supposé optimale. D'autre part, dans la somme :

$$L(\emptyset) = \Sigma Pr(D_j)\,\lambda(D_j),$$

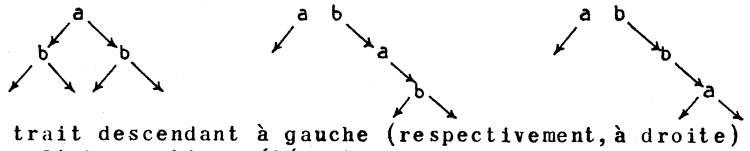
et relativement à chacun des a_i, on peut associer les paires D_j $D_{j'}$ telles que dans D_j on ait a_i^- et dans $D_{j'}$ a^+i, les états des autres objets étant les mêmes.

Il en résulte la deuxième partie de l'énoncé puisque les $Pr(D_j)$ sont des fonctions linéaires en chacun des α_i.

<u>Représentation des procédures</u>.

Il nous faut maintenant donner une représentation commode des procédures.

Considérons les trois schémas suivants :



où un trait descendant à gauche (respectivement, à droite) signifie que l'observation élémentaire dont ils partent a fourni un

résultat négatif (respectivement positif). Ainsi la seconde procédure consiste en les étapes suivantes :

I° - Observer a et b ensemble.

Si le résultat est négatif c'est que a^- et b^-

Si le résultat est positif : on passe à :

2° - Observer a

Si le résultat est négatif c'est que a^- et b^+

Si le résultat est positif : on passe à :

3° - Observer b ce qui conduit finalement à l'un des deux diagnostics $a^+ b^-$ ou $a^+ b^+$.

Dans les problèmes que nous traiterons, il apparaîtra qu'un diagnostic complet est presque toujours obtenu quand deux résultats négatifs ont été observés. Pour des raisons de commodité typographique, nous utiliserons cette particularité pour noter une procédure quelconque de la façon suivante :

Les ensembles constituant les observations élémentaires successives sont écrits de gauche à droite et celles-ci doivent être effectuées dans cet ordre quand la précédente a donné un résultat positif. Quand il est nécessaire d'effectuer encore des observations après un résultat négatif, la séquence correspondante complète est notée entre parenthèses immédiatement après l'observation en cause.

Par exemple, la première procédure du tableau précédent s'écrirait : a, (b), b et la seconde ab, a, b.

CAS D'UN PETIT NOMBRE D'OBJETS

Etudions d'abord le cas de $n = 2$. Les schémas donnés plus haut correspondent évidemment aux seules procédures possibles et les fonctions de coût sont :

$L\ I = 2$

$L\ II = 1 \times Pr(a^- b^-) + 2 \times Pr(a^- b^+) + 2 \times Pr\ (a^+) = 1 + 2\alpha + \beta - \alpha\beta$

$L\ III = 1 + 2\alpha + \beta - \alpha\beta$

Il convient donc si l'on n'adopte pas I de choisir II ou III selon que α est plus petit ou plus grand que β

On aura alors (en supposant $\alpha \leqslant \beta$) :

$1 + 2\alpha + \beta - \alpha\beta \leqslant 2$; équivalent à $\alpha \leqslant \beta \leqslant 1-2\alpha\ /\ 1-\alpha$ ce qui montre que le groupage n'est efficace que si l'une au moins des deux probabilités est inférieure à $3-\sqrt{5}\ /\ 2 = 0.381966...$

Nous ferons désormais l'hypothèse que $\alpha i = p$ pour tout i.

Pour trois objets déjà, il devient nécessaire de procéder selon une méthode rigoureuse, si l'on veut aboutir, sans trop de difficultés au résultat. Conformément aux idées que nous

avons développées dans le premier chapitre, nous avons fait usage d'une méthode récurrente qui consiste à calculer la fonction minimale L (α) pour toutes les positions en partant des positions terminales c'est-à-dire en étudiant d'abord la procédure à partir de

$((ab)^+(ac)^+)$ puis de $(ab)^+$ puis de $(abc)^+$ puis de $\emptyset$.

On obtient ainsi sept procédures dont la fonction n'est minorée par aucune autre, Ce sont :

I - abc, ab, ac, bc, a, b, c

$L_I = 1 + 6p + 3p^2 - 3p^3$

II - abc, a, (b, c) bc, b, c

$L_{II} = 1 + 7p - 4p^2 + p^3$

III - abc, a, (bc), b, (c), c.

$L_{III} = 1 + 8p - 7p^2 + 2p^3$

IV - ab, (c) ac, bc, a, b, c

$L_{IV} = 2 + p + 7p^2 - 4p^3$

V - ab, (c), a, (c) bc, b, c.

$L_V = 2 + 2p + 2p^2 + p^3$

VI - ab, (c), a, (c), b, (c), c.

$L_{VI} = 2 + 3p - p^2$

VII - Observation objet par objet

$L_{VII} = 3$

Le calcul montre que les procédures suivantes sont seules à retenir :

I pour $0 \leqslant p \leqslant 1/8\ (7 - \sqrt{35}) = 0.15628$ .. avec un coût toujours inférieur à 2,004 .. dans cet intervalle

II pour $1/8\ (7 - \sqrt{35}) \leqslant p \leqslant 2 - \sqrt{2} = 0.29289$ .. (coût inférieur à 2.722 ...)

V $2 - \sqrt{2} \leqslant p \leqslant 1/2\ (3 - \sqrt{5}) = 0.38187$ .. (coût inférieur à 3)

VII $1/2\ (3 - \sqrt{5}) \leqslant p$ coût toujours égal à 3.

Pour quatre objets la même technique conduit aux procédures suivantes. (Nous ne donnons ici que les résultats définitifs).

I - abcd, ab, (c, d) acd, bcd, c, (a, b, d) a d, bd, a, b, d.

$L_I = 1 + 10 + p + 2p^2 - 3p^3$ (à utiliser pour $p \leqslant 1/8\ (7 - \sqrt{35})$

II - abcd, ab, (c,d) a, (cd, c, d) bcd; b, (c,d) cd,c,d

$L_{II} = 1 + 11p - 5p^2 + p^3$ (à utiliser pour

$1/8\ (7 - \sqrt{35}) \leqslant p \leqslant p_1 = 0.21385 \ldots)$

(p_1 = racine de $1 - 6x + 7x^2 - 4x^3 + x^4 = 0$

III - ab (cd,c,d), a, (cd,c,d,)bcd, b(c,d) cd, c, d

$L_{III} = 2 + 5\,p + 2\,p^2 - 3\,p^3 + p^4$ (à utiliser pour $p_1 \leqslant p \leqslant 2-\sqrt{2}$)

IV - ab (cd,cd,) cd, (a,b) a, (c,d) c, (b) bd, b, d.

$L_{IV} = 2 + 6\,p - 3\,p^2 + 3\,p^3 - p^4$ (à utiliser pour $e-\sqrt{2} \leqslant p \leqslant 1/2\ (3-\sqrt{5})$)

V - observation objet par objet :

$L_V = 4$ (à utiliser pour $p \geqslant 1/2\ (3-\sqrt{5})$)

On obtient donc déjà ici un gain assez important dès que p est faible puisque par exemple si p = 0.10 on économise à peu près 50 % du nombre des observations en utilisant I au lieu de V.

Il est à remarquer que certaines positions (par exemple $(abc)^+$) ne sont jamais atteintes si l'on suit une procédure optimale.

Nous arrêterons là la discussion de ces procédures.

En effet, si l'on veut bien se souvenir que le treillis libre a cinq générateurs à 7581 éléments, on comprendra que la méthode d'exhaustion que nous avons suivie devient inapplicable. Il paraît sûr que de nombreuses règles combinatoires pourraient être trouvées qui faciliteraient ces recherches. Nous comptons revenir ultérieurement sur cette question qui peut présenter un certain intérêt théorique.

EMPLOI DE L'INFORMATION DE SHANNON WIENER

Par définition, l'information de SHANNON WIENER associée à la position α est nulle quand et seulement quand plus rien d'aléatoire ne reste dans le processus éventuel d'observation; c'est à-dire quand cette position est un diagnostic complet. Les autres propriétés que nous avons étudiées plus en détail dans la première partie confirment encore dans cette idée que l'information de SHANNON WIENER peut être une approximation raisonnable de la fonction $L(\alpha)$ qui décrirait le nombre moyen minimum d'observations restant à effectuer et qui, par conséquent, permettrait de choisir à chaque position l'observation x optimale. Nous utiliserons donc cette information pour construire des tactiques d'ordre zéro.

Avant toute observation, l'information que nous avons sur un objet a_i est par définition :

$$\alpha i \log_2 \alpha i + (1 - \alpha i) \log_2 (1 - \alpha i) = H_i$$

et pour l'ensemble des N objets qu'il nous faut diagnostiquer :

$\sum_i Hi$ puisque ces objets sont indépendants. Enfin on a :

$$H = N\,(p \log_2 p + (1 - p) \log_2 (1 - p))$$

si l'on suppose, comme nous le ferons désormais toujours que toutes les probabilités élémentaires sont égales à p.

La première observation peut porter sur un lot A de n objets et le gain moyen d'information qui lui est attaché est égal à

$$(\Pr A^+ \operatorname{Log}_2 \Pr A^+ + \Pr A^- \operatorname{Log}_2 \Pr (A^-)$$

soit

$$(1-q^n) \operatorname{Log}_2 (1 - q^n) + q^n \operatorname{Log}_2 q^n$$

En vertu du caractère monotone de cette fonction de n il revient au même pour fixer la première observation optimale selon la tactique d'ordre zéro de rechercher n tel que q^n soit le plus voisin possible de 1/2, c'est-à-dire de prendre

$n_0 = \left[\frac{\operatorname{Log} 1/2}{\operatorname{Log} (1-p)}\right]$ où la notation $[x]$ signifie l'entier le plus voisin de x. On en déduit le tableau suivant qui donne pour les premières valeurs de n les valeurs critiques pour lesquelles sont les mêmes valeurs absolues de la différence entre la valeur optimale vraie Log. 2/Log et celles correspondant aux deux entiers n et n + 1.

Ces valeurs sont les racines de l'équation $q^n + q^{n+1} = 1$

n		n		n	
1	0.312	5	0.118	9	0.070
2	0.245	6	0.101	10	0.064
3	0.181	7	0.088	11	0.05^
4	0.143	8	0,079	12	
5		9			

On pourra comparer ce tableau aux résultats antérieurs et vérifier que l'écart est faible entre ces solutions approchées et les procédures optimales.

La deuxième observation consiste alors, si un résultat positif a été obtenu, à observer un lot B formé de n objets de telle sorte que

$\Pr(B^+|A^+) = \frac{\Pr(B^+ A^+)}{\Pr(A^+)}$ soit le plus voisin possible de 1/2.

Il faudrait donc trouver les valeurs critiques correspondantes aux solutions de

$q^n + q^{n'} - q^{n+n'-m} = 1/2 (1-q^n)$ pour n fixé par l'équation précédente.

On aurait ensuite à chosir (si le résultat était encore positif) n'' objets nouveaux, m' objets appartenant au lot qui a servi à la première observation m"" appartenant au lot de la deuxième observation, m''' appartenant à la fois aux deux .... et l'on serait ramené dès les premières observations a un problème pratiquement aussi compliqué que celui de la détermination des procédures rigoureusement optimales.

Nous restreindrons donc le champ des tactiques admissibles à ce que nous appellerons les "tactiques simples".

Par définition, une "tactique simple" sera une tactique dans laquelle seront considérés comme admissibles :

soit des observations portant sur un lot d'objets dont on ne connaît rien d'autre que leur probabilité a priori (objets "libres").

soit des observations portant sur un sous-ensemble d'un lot ayant déjà donné un résultat positif.

Ces deux conditions sont bien compatibles puisque si B est un sous-ensemble de A^+ et si on a obtenu successivement A^+ et B^+, on ne sait rien de plus sur les objets de A-B que si aucune observation n'avait été faite sur eux,

Par conséquent, toute position dans le domaine des tactiques simples est résumé par :

1° - l'ensemble des objets sur lesquels l'on ne sait rien.

2° - l'ensemble des objets que l'on sait être "neutres".

3° - une famille d'ensembles disjoints contenant chacun au moins un objet "marqué".

Nous n'insisterons pas sur les avantages pratiques de ces tactiques tels que la diminution du risque des erreurs que pourrait faire craindre l'emploi de schémas opératoires trop compliqués.

Nous sommes donc amenés à choisir m objets formant le lot B dans le lot initial A de n objets de telle sorte que :

$$\Pr\,(B^+ \mid A^+) = \frac{\Pr\,(B^+)}{\Pr\,(A^+)} = \frac{1 - q^m}{1 - p^n}$$

soit le plus voisin de 1/2.

Par un développement en série, on obtient la valeur approchée $m = n/2 - \frac{pn^2}{8} + \ldots$.

L'évaluation précise du coût de cette tactique est un problème combinatoire assez compliqué et nous nous bornerons à donner une limite supérieure d'ailleurs très satisfaisante.

Chaque observation élémentaire est d'autant moins bonne que les probabilités correspondantes s'éloignent plus de 1/2; nous allons montrer que le pire cas est celui d'un lot de trois objets ayant donné un résultat positif et où les probabilités sont pourtant sûrement comprises entre 1/3 et 2/3.

En effet, un lot de n objets n'a donné un résultat que si on l'a testé et par conséquent que si p est inférieur à la valeur donnée par la table I.

Mais alors, pour n = 2, la probabilité conditionnelle que l'un de ces objets soit positif quand le lot a donné un résultat est évidemment plus grande que 1/2 mais inférieure à

$$\frac{p}{1 - q^2} \leqslant 0{,}62 \leqslant 2/3.$$

Pour $n > 3$ le choix de $m = (n/2)$ objets assure déjà que la probabilité du résultat est de $\frac{1 - q^m}{1 - q^n} \leqslant 2/3$ et pour $n>4$ cet écart maximum ne peut que décroître avec n.

Par conséquent, le gain moyen d'information par observation est donc toujours plus grand que :

$$\frac{1}{3} \text{Log}_2 \frac{1}{3} + \frac{2}{3} \text{Log}_2 2/3 = 0.9128 \ldots$$

alors que le maximum théorique est 1.000.

Dès que le nombre N d'objets à diagnostiquer est suffisamment grand, il en résulte que l'on aurait en moyenne N observations en procédant objet par objet et au plus à peu près 1.085 x N H_p en utilisant la tactique que l'on vient de décrire, résultat qui est bien voisin du minimum N H_p qu'impose le théorème fondamental de la théorie de l'information.

Rappelons pour terminer que la procédure proposée par DORFMANN qui a le premier étudié les possibilités des méthodes de groupage, mais en se limitant aux problèmes de diagnostic, est ce que l'on pourrait appeler une procédure ultra simple.

En effet DORFMANN ne considère qu'une seule observation de groupe suivie, - si elle est positive - d'un diagnostic objet par objet.

Cette procédure ne réalise évidemment pas l'économie la plus grande possible mais peut présenter un intérêt certain dans la pratique quand les propriétés déterministes du modèle ne sont pas rigoureusement valables (en particulier si la liaison "neutre" entraîne négatif est de type stochastique).

UN CAS PARTICULIER DE MODÈLE A STRUCTURE PLUS COMPLEXE

VÉRIFICATION D'UNE TABLE DE COVARIANCE

Toujours à titre d'exemple d'application des principes généraux, nous étudierons un cas pratique où l'ensemble E des ensembles admissibles est restreint par des considérations structurelles imposées à l'avance.

Soit x_{ij} $(i = 1, 2 \ldots\ldots k)$; $j = 1,2,\ldots,n)$ les k séries de n valeurs numériques dont on a calculé les $\left[\frac{k}{2}\right]$ covariances. Il est d'usage de vérifier l'ensemble des calculs (1) en formant la quantité $Z_j = \Sigma_i x_{ij}$ et en comparant les deux expressions Var Z et Σi var x_i + $2\Sigma ii'$ cov $(x_i x_{i'})$ qui doivent être égales. On admet que le résultat "négatif" de cette observation, c'est-à-dire l'égalité des deux expressions garantit que tous les objets sont "neutres" (c'est-à-dire que chacun des covariances a été calculée exactement).

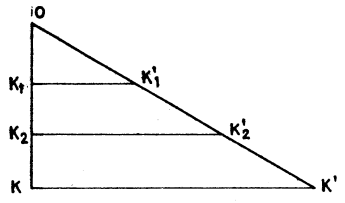
Voyons comment procéder pour retrouver le plus rapidement possible une erreur qui a été décelée par la méthode précédente.

---

(1) A moins que l'on ait employé des méthodes telles que celles de JOWETT, par exemple, qui donnent automatiquement une vérification des calculs. (Cf.J.Roy.Stat.Soc. (B) IX.(1949) p 89-90- et HAMMERSLEY (1952) Biometrics-(8) p.156-168.

Naturellement il est difficile de chiffrer exactement le coût réel en temps des diverses opérations possibles. Nous ferons donc l'hypothèse très simple et assez rapprochée de la réalité que le coût est le même pour toute observation élémentaire consistant à comparer la variance de la somme d'un certain nombre de variables x_i obtenu par calcul direct à sa valeur déduite de la table des variances et covariances que l'on vérifie. D'autre part, il est souvent raisonnable de penser que la probabilité d'une erreur est assez faible pour que l'on puisse considérer comme négligeables les chances d'en rencontrer plusieurs dans le tableau.

Enfin, le problème ne se pose seulement que si k est grand (disons plusieurs dizaines comme cela se présente typiquement en psychométrie). La représentation graphique ci-jointe permet de rendre intuitifs les raisonnements.



Le triangle O K K' symbolise la moitié inférieure gauche de la table des covariances et l'on fait correspondre les ordonnées aux indices k spécifiant la série des variables. Tout test basé sur le double calcul de la variance de la somme des k_i premières variables revient donc à savoir s'il y a ou non une erreur dans le triangle O K_i K'_i .

La procédure consiste donc d'abord à choisir successivement les valeurs de k_1 , k_2 .... etc et à effectuer les vérifications correspondantes.

Conformément au principe des tactiques d'ordre zéro, on commencera donc par vérifier les covariances entre les $k_1 \# \frac{\sqrt{-2}}{2}$ k premières variables ce qui signifie que l'on choisit K tel que la probabilité d'avoir un résultat négatif soit la plus voisine de 1/2.

Si le résultat est positif c'est qu'il y a sûrement une erreur parmi les covariances de ces variables et l'on est ramené au problème initial, mais avec k_1 variables seulement. Sinon on vérifierait les covariances entre les k + k" = k premières variables en prenant k'' de telle sorte que soit minimum la quantité :

$$\left(\begin{bmatrix} k_2 \\ 2 \end{bmatrix} - \begin{bmatrix} k_1 \\ 2 \end{bmatrix}\right) : \left(\begin{bmatrix} k \\ 2 \end{bmatrix} - \begin{bmatrix} k \\ 2 \end{bmatrix}\right)$$

qui détermine l'écart entre les probabilités de cette nouvelle observation et la valeur optimale 1/2. Si ce nouveau résultat était encore négatif on choisirait une nouvelle valeur k''' à l'aide d'une équation analogue et il en serait de même si le résultat était encore positif.

Après une suite de semblables observations, l'erreur se trouve localisée parmi les covariances entre une certaine variable x_i et les x'_i ($i \leqslant i'$) et il est alors probablement préférable de la rechercher directement.

A titre d'exemple de l'efficacité d'une telle procédure, on peut calculer que s'il n'y a effectivement qu'une seule erreur parmi les 45 covariances calculées entre dix variables, celle-ci peut être localisée en moyenne en 9,0 observations élémentaires seulement.

IV. - PROBLÈMES DE TRI

GÉNÉRALITÉS

Il est intéressant d'envisager les problèmes de tri en relation avec des probabilités d'absorbtion ; considérons en effet un observateur dont la situation sur le treillis des positions T se modifie en fonction du résultat de ses observations. Dans le cas des problèmes de diagnostic, T était fini et le processus se terminait quand l'observateur était parvenu à une position D. Dans les problèmes de tri, au contraire, T est supposé avoir un nombre infini de générateurs et le processus ne s'arrête que quand il atteint une position caractérisée par le fait que l'état de N^- objets est neutre et celui de N^+ autres marqué.

Il nous faudrait donc trouver une procédure telle que la somme infinie ; $\Sigma \Pr(\alpha_i) L(\alpha_i)$ soit minimum et nous indiquerons des méthodes pour obtenir des solutions approchées dans le cas où l'un des deux nombres N^- ou N^+ est nul, c'est-à-dire où l'on se propose seulement de trier des objets neutres ou des objets marqués. Nous ferons en outre l'hypothèse que toutes les probabilités élémentaires sont égales.

TRI D'OBJETS NEUTRES

Remarquons d'abord que si une nouvelle observation élémentaire a donné un résultat positif, il faut évidemment laisser de côté les objets sur lesquels elle a porté puisque toute nouvelle observation comprenant l'un d'eux aurait plus de chance d'être positive que s'il n'y figurait que des objets libres.

Une procédure optimale consistera donc à faire des observations $X_i^{(n-1)}$ sur des lots de n_1 objets libres jusqu'à l'obtention d'un résultat négatif; puis à recommencer sur des lots $X_i^{(n_2)}$ de n_2 puis sur des lots de $n_3 \ldots\ldots n_k$; les nombres n_1, n_2, $n_3 \ldots$ n_k ayant pour somme N.

A chaque fois il s'agira donc d'une distribution binomiale négative (distribution de PASCAL) avec le paramètre $(1-p)^{n_i}$. Le nombre moyen d'observations élémentaires sera :

$$L = \Sigma_{i=1}^{k} \left(\Pr\left(X^{n_i -}\right) \right)^{-1} = \Sigma_{i=1}^{k} (1-,p)^{-n_i} .$$

Pour minimiser cette expression on observera d'abord si le nombre k est déjà fixé, la plus petite valeur de la somme correspond à :

$$r\,(1-p)^{-n_o} + r'\,(1-p)^{-n_o-1}$$

où n_o est le plus grand entier contenu dans k, r' est le résidu de N, modulo k et enfin, $r = (N-r')k$ puisque l'on a toujours :

$$\frac{1}{(1-p)^{n_o-v}} + \frac{1}{(1-p)^{n_o+v}} \geqslant \frac{2}{(1-p)^{n_o}}$$

Pour déterminer maintenant l'ordre de grandeur de n_o, nous supposerons r' négligeable par rapport à k . On a alors :

$L \# \frac{N}{n_o}(1-p)^{-n_o}$ qui est minimum pour $n_o = -1/\log(1-p)$

On obtient ainsi pour p petit et N très grand, la valeur :

$L \# N_e^{-1}$ Log $(1/1-p)$ (où e désigne la base des logarithmes naturels) du nombre moyen d'observations élémentaires nécessitées par la procédure optimale.

TRI D'OBJETS MARQUÉS

Nos connaissances sur ce sujet sont beaucoup moins avancées et nous nous restreindrons au cas où l'on veut trier un seul objet marqué et ceci en employant exclusivement des procédures simples (c'est-à-dire qui ne comportent que des observations élémentaires X_i ordonnées par inclusion).

Il est vraisemblable que les procédures auxquelles nous aboutirons sont rigoureusement optimales, mais nous ne possédons pas encore la démonstration de cette hypothèse.

L' (n) représentera le nombre moyen d'observations à effectuer quand on sait déjà que $X^{(n)+}$ où $X^{(n)}$ est un ensemble de puissance n.

Dans les hypothèses où nous nous plaçons, on effectuera donc d'abord des observations sur des $X_i^{(n_o)}$ jusqu'à obtention d'un résultat positif, puis à ce moment on recherchera un objet marqué dans le dernier ensemble ainsi observé.

On a donc :

$$L(n) = \left(Pr(X^{(n_o)+})\right)^{-1} + L'(n_o) = 1/(1-q^{n_o}) + L'(n_o)$$

en particulier (Cf. plus bas) :

$$L(1) = p^{-1}$$

$$L(2) = \frac{2-q^2}{1-q^2}$$

$$L(3) = \frac{2+q-2q^3}{1-q^3}$$

$$L(4) = \frac{2 + q + q^2 + 3q^4}{1 - q^4}$$

ou suivant la nature de la seconde observation : $\frac{3 - 2q^4}{1 - q^4}$

Le problème se trouve maintenant reporté sur le calcul de $L'(n)$ pour une valeur donnée de q.

La règle suivante permet d'abréger les calculs :

Si X_n^+ et si $n' < N/2$, alors l'observation d'un sous-ensemble $Y=Y^{(n')}$ de X_n est plus efficace que celle de l'ensemble complémentaire $Z = Y^{(n-n')}$.

On a en effet (puisque nous nous limitons aux procédures simples) :

Si on effectue Y :

$$L''(n) = 1 + Pr\ (Y^-|X^+)\ x\ L'\ (n-n') + Pr\ (Y^+|X^+)\ x\ L'(n')$$

Si on effectue Z :

$$L'''(n) = 1 + Pr\ (Z^-|X^+)\ x\ L'\ (n') + Pr\ (Z^+|X^+)\ x\ L'\ (n-n')$$

soit encore :

$$Pr\ (X^+)\ x\left(L''\ (n) - L'''(n)\right) = L'\ (n')\ x\ Pr(Y^+)x\ (1-Pr(Z^-))$$
$$-Pr\ (Z^+)\ x\ (1-Pr\ (Y))L(n-n').$$

d'où le résultat puis $n' \leqslant n-n'$ entraîne évidemment :

$$L\ (n') \leqslant L\ (n-n')$$

Pour les premières valeurs de n on trouve alors les fonctions suivantes :

$L(n) = 3\ p(1+2q+2q^2)(1-q^3)^{-1}$ (Procédure : a,(b)Cf. notation du chapitre précédent).

$n = 4$: si $q \geqslant 0.7549$.. (racine de $x^3 + x^2 - 1 = 0$)

Procédure ab, (c), a.

$$L = p\ (2 + 2\ q + 2\ q^2 + 2\ q^3)\ (1 - q^4)^{-1}$$

si $q \leqslant 0.7549$ ....

Procédure a, (b), (c)

$$p\ (1 + 2\ q + 2\ q^3 + 3\ q^2 + 3\ q^3)\ (1-q^4)^{-1}$$

$n = 5$: si $q \geqslant 0.6823$ .. (racine de $x^4 + x^3 + x^2 - 1 = 0$)

Procédure ab, (c, d), a

$$p\ (2 + 2\ q + 2\ q^2 + 3\ q^3 + 2\ q^4)\ (1 - q^5)^{-1}$$

si $q \leqslant 0.6823$ ...

Procédure a,(b,c,d).

$$p\ (1 + 2\ q + 3\ q^2 + 4\ q^3 + q^4)\ (1-q^5)^{-1}$$

Ce calcul devient rapidement inextricable quand n croit et nous nous bornerons aux indications suivantes :

si $p > \frac{3 - \sqrt{5}}{2} = 0{,}382 \ldots$: trier objet par objet.

si $x_0 = 0.2451 \leqslant p \leqslant 0.382$: examiner des lots de 2 jusqu'à un résultat positif.

si $x_1 = 0{,}1809 \leqslant p \leqslant x_0$ examiner des lots de 3 jusqu'à un résultat positif.

$(1 - x_0)$ et $(1 - x_1)$ sont respectivement les racines de

$$x^3 + x^2 = 1 \text{ et de } x^4 + x^3 = 1).$$

La situation est donc très semblable à celle où nous nous sommes trouvés dans l'étude des problèmes de diagnostic et nous sommes conduits à faire appel à des méthodes d'approximation.

POSSIBILITÉ D'EMPLOI D'UNE INFORMATION

A - Cas des objets neutres.

Nous avons vu dans la première partie que l'information H_0 attachée à l'opérateur $[q\ d/\ d\ q]_{q=0}$ livrait le nombre moyen de fois où l'évènement de probabilité q était réalisé. L'information $H_0(X)$ associée à l'observation X nous donne donc exactement ce que nous cherchons puisque, comme on l'a vu, il est impossible dans notre modèle d'apprendre qu'il existe un objet à l'état neutre sans l'identifier du même coup.

On vérifie facilement que $H_0(X)$ est maximum quand les n objets de X sont libres, enfin n est déterminé par la condition que $H_0(X) = q^n \times n + (1 - q^n) \times 0 = n\ q^n$ soit maximum.

Ceci redonne bien la règle : $n_0 = -1/\log q$ et la même limite inférieure asympotique du nombre des observations.

B - Cas des objets marqués.

Considérons maintenant ce que donnerait l'usage de l'information H_1 associée à $[p\ d/\ d\ p]_{p=0}$ dans le cas des objets marqués.

Tout d'abord $H_1(X)$ pourrait être nul alors que manifestement un pas a été fait vers la solution : par exemple si on sait déjà que $(ab)^+$ l'observation de a termine la procédure.

Pourtant :

$$H_1\ (a \mid ab) = p(1-q^2)^{-1} \left(\frac{pd}{dp} \operatorname{Log} \frac{p}{1-q^2} \right)_{p=0} + p(1-p)(1-q^2)^{-1} \left(\frac{pd}{dp} \operatorname{Log} \frac{pq}{1\ q^2} \right)_{p=0}$$

est nul puisque l'on savait déjà que UN ob-

jet au moins était marqué et que l'observation de a n'a pas changé ce minimum.

Inversement H_1 peut être assez grande bien que l'utilité de l'observation soit fort restreinte du point de vue du tri. Par exemple si l'on sait déjà que :

$(abcd)^+$, $(abef)^+$, $(abgh)^+$ à l'observation (c e g) correspond un gain d'information non nul puisque $(ceg)^+$ entraînerait qu'au moins DEUX objets au lieu d'UN seul soient marqués.

La raison de cette insuffisance est évidemment due au fait que - contrairement à ce qui se passe pour l'état neutre - l'identification d'un objet marqué n'est pas équivalente à la preuve de son existence.

Nous nous limiterons donc, comme nous l'avons déjà fait plus haut, aux procédures simples et nous supposerons que nous savons déjà que l'ensemble Y_n^+ de puissance n contient un objet marqué. Le gain d'information $H_1^!$ associée à $[\ ]_{p=o}$ aura alors une valeur finie pour toute observation d'un sous-ensemble X_n de puissance m de Y. En effet

$$H_1^! (X_m | Y_n^+) = Pr(X_m^+ | Y_n^+) \text{Log } m - Pr(X_m^- | Y_n^+) \text{ Log } (n-m) + \text{Log } n$$

parce que d'une part Y_n^+ entraîne seulement qu'un seul objet soit marqué et ceci peut être de n manières différentes et, d'autre part, après l'observation de X_m ce nombre de possibilités est restreint à m ou n-m.

Ce résultat est d'ailleurs intuitif puisque, à la constante 1/Log 2 près, Log n donne le nombre moyen d'observations nécessaires pour achever le diagnostic.

Montrons qu'ici encore il vaut mieux si $m < n/2$ observer X_m que l'ensemble complémentaire X_{n-m}. On a en effet à comparer :

$$(1-q^m) \text{ Log } m + (q^m - q^n) \text{ Log}_2 (n-m)$$

avec :

$$(1 - q^{n-m}) \text{ Log}_2 (n-m) + (q^{n-m} - q^n) \text{ Log}_2 m$$

qui est certainement plus grand en vertu des inégalités :

$$\text{Log } m \leqslant \text{Log } (n-m) \text{ et } (q^{n-m} - q^n) = q^{n-m}(1-q^m) \leqslant 1 - q^m$$

Théoriquement il faudrait résoudre en m l'équation: $[H_1^! = \text{maximum}$. Dans la pratique il semble suffisant de considérer le problème comme un problème de diagnostic et de procéder comme il a été indiqué dans le chapitre précédent une fois que l'ensemble initial positif Y_{no} a été obtenu. Le coût total en moyenne de la procédure est alors :

$$L = 1/Pr\ (Y_{no}^+) + \text{Log}_2\ n_o$$

ce qui conduit, pour p petit, à déterminer n_o par l'équation: $\frac{dL}{dn} = 0$ c'est-à-dire à choisir n_o de telle sorte que $Q = q^{n_o}$ satisfasse à :

$\frac{\text{Log } Q}{(1-Q)^2} + \frac{1}{\text{Log } 2} = 0$ c'est-à-dire encore à prendre :

$$n_0 = \frac{1.71006}{\text{Log}_{10} q} \cdots \quad \text{puisque } Q = 0.51293...$$

Nous pouvons donc considérer comme pratiquement résolus les deux problèmes de tri que nous nous étions posés. Il faut cependant remarquer que dans le second cas, la question du choix de la valeur n_0 qui caractérise les premières observations Y ne peut être résolue autrement que par un recours à un raisonnement extérieur au type d'information employé. C'est là une déficience de l'information H' que nous avons du utiliser puisqu'elle ne s'applique qu'à partir du moment où l'on a obtenu un résultat positif. A son tour ceci résulte du caractère assez grossier de l'opérateur associé à l'information. Il est certain que si l'on voulait résoudre des problèmes de tri moins éléméntaires, il faudrait recourir à des considérations plus profondes sur les zéros des polynomes P (p) car c'est d'eux que dépend en définitive toute information de ce type. Des calculs encore fragmentaires semblent indiquer que des règles simples pourraient gouverner la distribution de ces valeurs suivant la position α. Il y a là un domaine de recherches intéressant en lui même du point de vue algébrique, mais aussi extrêmement prometteur par les possibilités qu'il offre de mieux comprendre la signification et le rôle des opérateurs linéaires définissant les informations.

V. - PROBLÈMES DE TEST D'HYPOTHÈSE

TEST DE L'HYPOTHÈSE $q_1 \leqslant q \leqslant q_0$

Nous nous limiterons au cas très particulier où la fréquence q des objets neutres dans une population infinie étant inconnue on se propose de choisir entre les deux hypothèses.

H_0 : q est supérieur à une certaine valeur donnée q_0

H_1 : q est inférieur à une certaine valeur donnée $q_1 < q_0$

Si l'on se fixe en outre les probabilités maxima α et β d'accepter H_0 quand $q = q_1$ et d'accepter H_1 quand inversement $q = q_0$, on sait qu'une procédure optimale consiste à utiliser le test séquentiel de WALD; mais ici de nouvelles possibilités sont ouvertes à cette méthode du fait que l'on peut remplacer l'observation élémentaire (x) : l'objet x pris au hasard est-il ou non marqué ? par l'observation d'un évènement α dont la probabilité est une fonction P(q) et l'on peut se demander s'il n'est pas possible d'abréger ainsi le nombre moyen des observations nécessitées par le test.

Comme on l'a vu dans la première partie, le nombre moyen d'observations indépendantes de α est de la forme :

$\frac{K}{W(0)}$ (si $q = q_0$) et $\frac{K'}{W(1)}$ si $q = q_1$ où K et K' ne dépendent que

de α et β et où W est "l'information de WALD", attachée à α c'est-à-dire :

$$W(i) = P(qi) \text{ Log } \frac{P(qi)}{P(qj)} + (1-P(qi)) \text{ Log } \frac{1-P(qi)}{1-P(qj)}$$

Il semble difficile de discuter directement W d'autant d'ailleurs qu'il faudrait considérer à la fois W (0) et W (1).

En nous basant sur les inégalités obtenues dans la deuxième partie nous remplacerons ce problème par celui de rendre maximum P (qi) - P (qj).

Supposons donc qu'un certain nombre d'observations élémentaires résumées dans le mot α^+ aient été effectuées et comparons la différence :

$\Pr(X^i|\alpha^+)_{q=qo} - \Pr(X^-|\alpha^+)_{q=q1}$ où X est une observation de n objets pouvant appartenir au mot α à la différence $[\Pr(X^-)_{q=qo} - \Pr(X^-)_{q=q1}$ relative au cas où les n objets sont libres.

On a $\Pr(X^-|\alpha^+) = q^n Q R^{-1}$ où Q et R sont les polynomes en q définis par $R = \Pr(\alpha^+)$ et $Q = \Pr(\beta^+)$ avec $\beta^+ X^+ = \alpha^+ X^+$; d'après ce qui a été vu dans le deuxième chapitre on a $Q \leqslant R$ et la deuxième différence ci-dessus est plus grande que la première quand :

$$q^{n_o} - q^{n_1} \geqslant q^{n_o} \frac{Q(q_o)}{R(q_o)} - q^{n_1} \frac{Q(q_1)}{R(q_1)}$$

Une condition suffisante pour que cette inégalité soit satisfaite est donc que $q^n (1 - Q(q) / R(q))$ soit fonction croissante de q ou encore que l'on ait identiquement :

$$\frac{d}{dq} \frac{Q(q)}{R(q)} \leqslant 0.$$

Dans ce cas (dont nous étudierons la signification au chapitre suivant), il est préférable d'observer n objets libres que les n objets de α.

En particulier si, α' était simplement l'observation Y^+ d'unlot de m objets, il y aurait avantage à choisir un X disjoint de Y.

Nous sommes ainsi conduits à définir notre tactique optimale d'ordre zéro comme consistant à remplacer grâce au groupage, le test de ($q \geqslant q_o$ ou $q \leqslant q_1$) par celui de ($q' \geqslant q_o^{n_o}$ ou $q' \leqslant q_1^{n_o}$) avec n_o tel que W(i) soit maximum et à procéder ensuite comme dans un problème ordinaire d'analyse séquentielle.

En ce qui concerne la détermination de n_o nous n'avons obtenu aucun résultat simple; il sera donc recommandé de chercher d'abord la valeur n_i de n telle que $q_o^{n_1} - q_1^{n_1}$ soit maximum puis par tâtonnement de rendre les plus grandes possibles W(0) et W(1). Le tableau (page 107) donnant pour des valeurs typiques de q_o et de q_1 la valeur qui rend maximum la différence $q_o^n - q_1^n$ permet de faciliter ces calculs. On remarquera que la méthode a un domaine d'application très large puisqu'il n'est préférable de faire les observations objet par objet que si $q_o + q_1$

est plus petit que l'unité. Enfin, on pourra noter que si certaines conditions matérielles rendent aisée l'observation des mots symétriques auxquels il a été fait allusion à la fin du chapitre III, ceux-ci fournissent une transformation $q \to P(q)$ particulièrement efficace pour le test de $q \lesseqgtr m/n$.

TEST DE L'HYPOTHÈSE: q_1 PLUS PETIT OU PLUS GRAND QUE q_2

Le problème, ici, est de savoir si la fréquence des objets neutres dans deux populations infinies E_1 et E_2, est la même ou non. Nous supposerons que l'on a décidé d'employer la méthode de A. WALD qui consiste, après avoir effectué un tirage aléatoire dans E_1 et dans E_2, à observer selon laquelle des quatre modalités suivantes s'est réalisé un certain évènement α construit à partir d ces objets.

α_1^+ et α_2^+ ou bien α_1^- et α_2^- c'est-à-dire α réalisé ou non à la fois dans les deux populations.

α_1^+ et α_2^- ou bien α_1^- et α_2^+ c'est-à-dire α réalisé seulement dans l'une des populations.

Pr (α^+) étant une fonction croissante de $p = 1 - q$, la comparaison du nombre des cas où α_1^+ et α_2^- à ceux où α_1^- et α_2^+ donne un test commode et d'une efficacité très suffisante de l'hypothèse $q_1 = q_2$.

Nous devrions donc considérer comme précédemment quels sont les mots qui rendent une semblable procédure la moins coûteuse en observations élémentaires. De fait, il ne peut en être ainsi, tout au moins dans le cas général, car le test basé sur α est un test de nature chaque fois différente selon le choix de ce mot. En effet, si l'on prend le cas considéré par WALD où α est un seul objet x, ce que l'on testera c'est la présence du point $(q_1\ q_2)$ à l'intérieur du domaine où est satisfaite l'inégalité :

$u_o \leqslant \dfrac{(q_1\ (1-q_2)}{q_2\ (1-q_1)} \leqslant u$, domaine qui n'est en aucune manière équivalent, même si l'on aboutit aux mêmes probabilités d'erreur de première et deuxième espèce, au domaine :

$u'_o \leqslant \dfrac{q_1^n\ (1-q^n{}_2)}{q_2^n\ (1-q^n{}_1)} \leqslant u'_1$ correspond à un test où α est une observation élémentaire sur un lot de n éléments. On pourrait arguer qu'il en était de même dans le chapitre précédent et que là aussi les "operating curves" étaient modifiées par l'utilisation du groupage. Mais du moins conservait-on ce qui paraissait l'essentiel du test à savoir des valeurs données des probabilités d'erreur α et β quand l'une des deux hypothèses H_o ou H_1 était vraie alors que rien de semblable n'existe dans le cas présent.

Nous nous placerons donc à un point de vue en quelque sorte préliminaire et nous indiquerons seulement les améliorations manifestes que peut produire le remplacement de q par q^n.

Soit donc q_1 q_2 et soit λ le rapport $\frac{q^n_1 \ (1-q^n_2)}{q^n_2 \ (1-q^n_1)}$; λ est toujours plus petit que l'unité et nous pouvons admettre que l'on aura avantage à remplacer n par n' au moins quand les deux conditions suivantes sont remplies :

I° - $\lambda_n \geqslant \lambda_{n'}$ puisque dans ce cas les probabilités $Pr(A_1^+ A_2^-|$ et $Pr\ (A_1^- A_2^+)$ seront plus éloignées de 1/2.

2° - La probabilité P que $(A_1^+ A_2^+)$ ou que $(A_1^- A_2^-)$ n'aura pas augmenté.

Les résultats suivants facilitent la détermination de la valeur optimale de n quand on a déjà choisi le type "d'operating curve" que l'on désire réaliser.

I° - λ est une fonction décroissante de n

En effet l'inégalité $\lambda n \geqslant \lambda h+1$ se ramène après quelques transformations simples à l'inégalité classique :

$$\frac{1 - x_2}{1 - x_1} \leqslant \frac{1 - x^m_2}{1 - x^m_1}$$

2° - Si $q_1 = q_2 = q$ la valeur minimum de P correspond à n tel que $q^h = 1/2$ ce qui est immédiat par calcul direct/

VALEURS OPTIMALES DE n

pour certaines valeurs courantes de q_o et q_1

Valeurs de q_1 en pourcentage	Valeurs de q_o en pourcentage								
	50	55	60	65	70	75	80	85	90
55	2								
60	2	2							
65	2	2	2						
70	2	2	2	3					
75	2	2	3	3	3				
80	2	3	3	3	3	5			
85	3	3	3	4	4	5	5		
90	3	4	4	4	5	6	6	8	
95	4	5	5	6	6	7	9	11	14

VI. - PROBLÈMES D'ESTIMATION

FRÉQUENCE DES OBJETS MARQUÉS DANS UNE POPULATION INFINIE

Les problèmes d'estimation au sens classique de la statistique mathématique offrent un large domaine aux méthodes de groupage que nous préconisons. Nous nous bornerons à traiter ici deux exemples typiques. Faisons d'abord observer que l'usage de l'information de FISHER est à peu près indispensable dans ce chapitre, car le calcul de la variance exacte d'une estimation est presque toujours extrêmement laborieux. Nous considèrerons donc que le but à atteindre par une procédure quelconque est de choisir les observations successives de telle sorte qu'après N d'entre elles la limite supérieure de la variance de la valeur estimée du paramètre soit le plus faible possible. Il est classique que cette limite supérieure soit atteinte asymptotiquement quand N tend vers l'infini.

Dans ce premier chapitre, nous traiterons le cas où le paramètre inconnu $p = 1 - q$ est la fréquence des objets marqués dans une population infinie et où toutes les observations élémentaires sont possibles a priori.

Supposons que nous ayons déjà effectué un certain nombre d'observations résumées par l'ensemble des deux mots $\mathcal{A}^+A^-$.

Evidemment nous devons laisser de côté les objets qui figurent dans A^- puisque leur état est connu. Considérons maintenant un lot X de m objets dont certains peuvent figurer dans $\mathcal{A}^+$. La quantité d'information attachée à l'observation de X dépend des probabilités : $\Pr(X^+|\mathcal{A}^+)$ et $\Pr(X^-|\mathcal{A}^+)$ qui sont respectivement :

$$\Pr\frac{(X^+\,\mathcal{A}^+)}{\Pr\,(\mathcal{A}^+)} \quad \text{et} \quad \frac{\Pr\,(A^-\,\mathcal{A}^+)}{\Pr\,(\mathcal{A}^+)}$$

Nous écrirons $\Pr(\mathcal{A}) = P$. Comme nous l'avons vu dans le second chapitre, l'évènement $X^-\mathcal{A}^+$ est résumé par X^- et un autre mot $\mathcal{B}^+$ qui se déduit de $\mathcal{A}^+$ par suppression pure et simple des objets de $\mathcal{A}^+$ qui appartiennent à X . On a donc $\Pr(X^-\mathcal{A}^+) = q^m\,Q$ en posant $Q=\Pr(\mathcal{B}^+)$ et, par définition même, comme $\mathcal{B}^+$ entraîne $\mathcal{A}^+$ on a $Q \leqslant P$, l'égalité ne pouvant avoir lieu que si $\mathcal{A}^+$ et X n'ont aucun objet en commun.

Dans ces conditions, l'information attachée à l'observation de X est :

$$(P\,\frac{d}{dq}\,(q^m\,Q) - q^m\,Q\,\frac{d}{dq}\,P)^2\;P^{-2}\;Q^{-1}\,(P-q^m\,Q)^{-1}\;\;q^{-m}$$

et peut se mettre sous la forme :

$$q^{m-2}\left[m + q\frac{P}{Q} \times \frac{d}{dq}\frac{Q}{P}\right]^2\left[\frac{P}{Q} - q^m\right]^{-1}$$

qui se réduit à : $\boxed{q^{m-2}\;\;m^2\;(1-q^m)^{-1}}$

quand $\mathcal{A}$ et Q sont disjoints.

Comme $\frac{P}{Q}$ est plus grand que 1, il ressort de cette formule que l'information apportée par ces n objets est sûrement plus petite que l'information apportée par n objets libres (c'est-à-dire ne figurant pas dans $\mathfrak{A}$).

Nous allons donner quelques cas où cette condition est remplie :

I° - Soit $\mathfrak{A}^+ = C^+$ consistant en un lot de n objets dont on sait qu'il contient au moins un objet "marqué", soit m le nombre d'objets de C^+ qui appartiennent aussi à A; on a : $P = 1 - q^n$; $Q = 1 - q^{n-m}$

$$\frac{d}{dq}\frac{P}{Q} = Q^{-2}\ q^{n-m-1}\ (n-m-nq^m + mq^n) \gg 0$$

d'après l'inégalité classique : $n > m$ entraîne $\frac{1-x^n}{1-x^m} \leqslant n/m$

En particulier si m = n c'est-à-dire si, sachant que C^+, on effectue l'observation d'un lot d'objets pris parmi ceux-ci l'information apportée est sûrement inférieure à $(1-q^{n-m})$ fois l'information que procureraient ces objets si l'on ne savait pas à l'avance qu'ils font partie d'un lot ayant donné déjà un résultat positif.

2° - Soit q très voisin de 1. D'après la deuxième méthode indiquée au chapitre II, on peut écrire :

$P = p^r\ (a_1 + a_2 p \ldots\ldots)$ et $Q = p^{r'}\ (b_1 + b_2\ p + \ldots)$ où r (respectivement r') est le nombre minimum d'objets qui doivent être marqués pour que l'on puisse avoir $\mathfrak{A}^+$ (respectivement $\mathfrak{B}^+$) et où a_1 (respectivement b_1) est le nombre de systèmes différents de r (respectivement r') objets marqués qui permettent $\mathfrak{A}^+$ (respectivement $\mathfrak{B}^+$). Puisque $\mathfrak{B}^+$ se déduit de $\mathfrak{A}^+$ par suppression de certains objets, on a certainement, soit $r < r'$, soit $r = r'$ et $a_1 > b_1$

Dans le premier cas on a donc au moins pour p tendant vers zéro: $\frac{d}{dq}\frac{P}{Q} \geqslant 0$. Par conséquent, l'observation de X apporte à la limite moins d'information que l'observation de n objets libres quand X^- et $\mathfrak{A}^+$ entraîne l'existence de plus d'objets marqués que n'entraîne $\mathfrak{A}^+$ seulement ce qui est un résultat assez surprenant a priori.

A titre de contre exemple, nous citerons enfin le cas de $\mathfrak{A}^+ = (a\ b)^+\ (b\ c)^+$; X = a ; B = b et où $\frac{d}{dq} P/Q = 1 - 2q$ n'est donc pas entièrement positif.

On a d'ailleurs ici pour valeur de l'information :

$(1 + q^2)^2\ q^{-1}\ (1-q^2)^{-1}\ (1+q-q^2)^{-1}$ et le rapport de cette quantité à 1/qp qui est la valeur classique de l'information apportée par l'observation d'un seul objet libre, tend vers 2 quand q tend vers 1 ce qui montre que les résultats précédents ne pourraient

pas être étendus sans précaution à toutes les positions possibles.

La tactique optimale : D'après ce que l'on vient de voir, quels que soient les nombres n et m la deuxième observation que l'on a à effectuer après avoir obtenu un résultat positif sur un lot de n objets n'apporte jamais tant d'information que quand elle porte sur m objets tous distincts des précédents.

Il sera donc conforme aux principes des tactiques optimales d'ordre zéro que de choisir un nombre n tel que l'observation de n_o éléments soit la meilleure possible et de répéter cette observation autant de fois qu'il sera nécessaire sur des objets toujours nouveaux pour obtenir la précision voulue.

Les tableaux I et II donnent les valeurs n_o pour un certain nombre de valeurs de q.

Pour p voisin de zéro, il est intuitif que n_o doit être grand. En dérivant par rapport à n l'expression

$$q^{n-2}\, n^2\, (1-q^n)^1$$

on vérifie que, pour q donné, la courbe représentant la quantité d'information en fonction de n a bien un seul maximum et que celui-ci correspond à n tel que $q^{n_o} = Q_o$; où Q_o est la racine unique entre zéro et un de l'équation :

$$1 - Q + 1/2 \text{ Log } Q_n = 0$$

Ce résultat conduit à la règle pratique suivante :

n doit être tel que la probabilité d'une observation négative soit la plus voisine possible de 1/5

En effet, le calcul donne $Q_o = 0.2031882...$

($\log_{10} Q_o = 1.3078985...$) et il est assez curieux tout à la fois que cette valeur soit si éloignée de la valeur 1/2 qui apparaissait dans tous les problèmes de diagnostic et si près d'une autre valeur simple 1/5. Il est à noter que cette règle fournit un moyen pratique de s'approcher de n optimal même si initialement on ignorait la valeur de q avec une approximation suffisante et si l'on ne voulait pas effectuer une estimation intermédiaire.

La tactique que nous proposons revient donc à remplacer l'estimation directe de q par celle de q^{n_o} puis à en déduire q. L'équation au maximum de vraisemblance donne la formule très simple $\hat{q} = (N^-/N)^{1/n_o}$ où N est le nombre de lots observés et N^- le nombre de ceux d'entre eux qui ont donné un résultat négatif.

Malgré l'importance de la réduction de variance qui résulte d'un groupage optimal quand p est petit (elle s'exprime par un coefficient de l'ordre de $1.54\ (p-p^3/12)$ il est nécessaire de souligner que l'emploi de cette méthode reste probablement limitée au cas où l'on envisage d'effectuer un nombre N d'observations assez élevé en raison des distorsions qu'inflige l'opération "racine n-ième" à la distribution de la valeur estimée.

Pour N modérément grand, il est préférable de ne pas négliger le fait que $\hat{q}$ présente une erreur systématique. [r étant la fréquence des résultats positifs dans les N observations nous développons en puissance de r la valeur estimée $\hat{q} = (1-r)^{1/n_o}$.

D'autre part, les expressions classiques des moments autour de zéro d'une variable binomiale donnent :

$$\text{valeur moyenne de } r^k = -(1-q^{n_o})^k + \frac{k}{2}\ \frac{q^{n_o}(1-q^{n_o})^{k-1}}{N} + \cdots$$

en négligeant les termes dont l'ordre en N est inférieur à moins un.

Il s'en déduit que, toujours avec la même approximation, la valeur moyenne de $\hat{q}$ est donnée par :

$$q + \frac{n_o^{-2}(n_o-1)}{2N}\ q^{1-n_o}\ (1-q^{n_o}) + \cdots \qquad \text{soit à peu près}$$

$q\left(1 + \frac{2}{n_o\ N}\right)$ ce qui permet de corriger la partie principale de l'erreur systématique.

Signalons enfin, pour terminer, que - toujours pour p petit - la seule connaissance du fait qu'un lot de n objets contient au moins un objet "marqué" apporte environ 1/3 (exactement (1+p/2) 0.3288...) de toute l'information relative au paramètre contenu dans ce lot.

UN PROBLÈME DE GÉNÉTIQUE MENDELIENNE

Nous rappellerons sommairement qu'un individu peut -relativement à une paire d'allèles a/A - présenter l'un des trois génotypes (a a), (a A), ou (A A), les deux derniers étant indistinguables en raison du caractère "marqué" de l'allèle dominant. Donc quand nous dirons qu'un individu est phénotypiquement dominant nous employerons seulement une locution équivalente à "l'observation de (x x) donne un résultat positif "en raison de la correspondance

allèle dominant ⟷ objet "marqué"

allèle récessif ⟷ objet "neutre"

Nous étudierons ici un problème d'estimation de la fréquence p d'un allèle dominant dans une population monogame, isogamique panmixique et en équilibre génétique (c'est-à-dire telle que les fréquences respectives des trois génotypes (a a) (a A) (A A) soient respectivement q^2, 2 p q, p^2) quand l'échantillon prélevé contient à la fois des individus indépendants que nous appellerons "parents" et d'autres individus que nous appellerons "descendants" et qui résultent du croisement des précédents.

En effet, d'après la génétique mendelienne, les descendants d'un croisement (x x') X (y y') où x x' y y' symbolisent quatre allèles quelconques ont, indépendamment et avec des probabilités égales a priori, l'un des quatre génotypes (x y), (x' y), (x y') ou (x' y').

Par conséquent, si parmi les descendants d'un individu phénotypiquement dominant on observe un individu récessif on peut en conclure que ce parent était hétérozygote (c'est-à-dire (a A) et non homozygote (A A).

Le problème que nous traiterons sommairement sera celui de l'efficacité relative de la détermination du phénotype soit d'un parent, soit d'un descendant pour l'estimation de la valeur de q dans la population où a été prélevé l'échantillon, en supposant qu'on a déjà examiné un nombre assez grand de parents pour qu'il soit toujours possible de choisir des descendants de parents dont le phénotype est connu. Pour cela, nous calculerons les quantités d'information I_x attachée à la détermination de divers phénotypes/

I° - PARENTS : Les formules du chapitre précédent (cas de deux objets libres) redonnent le résultat classique $1 = 4(1-q)$

2° - DESCENDANTS d'un croisement (récessif x récessif). L'information est nulle puisque le descendant a, nécessairement, lui aussi, le phénotype récessif.

3° - DESCENDANTS d'un croisement (récessif x dominant). Ici nous devons introduire une notion supplémentaire : étant donné un couple (récessif x dominant) le fait qu'un seul de leurs descendants soit récessif permet de conclure que le parent dominant était hétérozygote et à partir de ce moment l'observation des autres descendants n'apporte rien. Nous préciserons donc en ajoutant : "quand" m descendants déjà examinés avaient montré un phénotype dominant .

On a alors :

Probabilité a priori pour que le parent soit hétérozygote et qu'il ait m descendants dominant puis un récessif $2^{-m}pq$.

Probabilité a priori pour que les m + 1 descendants soient dominants : $p^2 + 2^{-m}pq$.

$$\text{D'où } I_{rm} = 2^{-m} q^{-1} (1-q-q2^{-m})^{-1} (1-q+q2^{1-m})^{-2}$$

4°- DESCENDANTS d'un croisement (dominant x dominant). Des remarques analogues doivent être faites puisque le croisement ne donnera un individu récessif que si les deux parents sont hétérozygotes et dans ce cas la probabilité sera égale à 1/4. Les probabilités résultantes sont : $(3/4)^m p^2 q^2$.

Et $4 p^2 q^2 (3/4)^m + 4 q p^3 + p^4$ d'où enfin :

$$I_{rm} = 4(3/4)^m (1+q)^2 (1+2q-3q^2+q^2(3/4)^m)^{-1}(1+2q-3q^2+q^2(3/4)^m)^{-2}$$

5° - Nous calculerons finalement l'information qu'apporterait la connaissance du génotype exact d'un individu dominant (cette connaissance pourrait résulter par exemple de l'examen d'un très grand nombre de descendants et d'une inférence statistique). On trouve :

$$I_g = 2 q^{-1}(1 + q)^{-1} (1-q^2)^{-1} .$$

Après ces calculs préliminaires, nous pouvons aborder la discussion des tactiques possibles :

I° - Puisque l'examen d'un descendant ne peut nous apporter que de façon aléatoire une connaissance sur le génotype des parents, on a certainement $I_{rm} \leq I_g$ et $I_{dm} \leq 2 I_g$ mais en comparant I_g à I_p on s'aperçoit que si $q(q+1)$ est plus grand que 1/2

(c'est-à-dire si q n'est pas inférieur à $\frac{1}{2}(\sqrt{3}-1) = 0.366...$),la dichotomie (récessif/dominant) apporte plus d'information que la dichotomie (homozygote/hétérozygote) parmi les individus à phénotype dominant. Il ne peut donc être intéressant de considérer les descendants des croisements (récessifs/dominants) que si q est inférieur à cette valeur puisque sinon il vaudra toujours mieux examiner le phénotype d'un nouveau parent.

2° - Sous l'hypothèse que $q < 1/2$, on vérifie par calcul direct pour les premières valeurs de m puis, en formant le rapport $Ir'm/Ii'm+1$ qui tend vers :

$2(1-q)(1-q/2)(1-q/4)^{-2} \geqslant 1$ que I_{rm} est une fonction décroissante de m.

Si donc on a déjà examiné un descendant d'un croisement (récessif/dominant) il est toujours préférable d'examiner un descendant d'un autre couple qu'un second descendant de celui-ci.

3° - L'information I_{rd} apportée par l'examen de ce seul descendant est $q^{-1}(1-q)^{-2}$. Elle n'est supérieure à $4(1-q^2)^{-1}$ que si q est plus petit que $\frac{1}{8}(\sqrt{41}-5) = 0,178...$,ce qui constitue la valeur critique pour l'examen des descendants de ce croisement.

4° - De la même manière, on peut calculer l'information I_h qui serait attachée à l'observation d'une paire d'individus dominants et qui permettrait de savoir s'ils sont ou non tous les deux hétérozygotes. On trouve :

$I_h = 16(1+q)^{-2}(1-q)^{-1}(1+3q)^{-1}$ et cette quantité est sûrement inférieure à $4(1-q^2)^{-2}$ quand $q \geqslant 1/2$.

5° - D'autre part I_{dm} est sûrement inférieur [$I' = 4(1+q)^2 p^{-2}(1+3q)^{-2}$; puisque cette valeur dérive de I_{dm} en faisant m=0 au numérateur et $m = \infty$ au dénominateur.

Mais I' n'est plus grand que $4(1-q^2)^{-2}$ que pour $q \geqslant \frac{\sqrt{10}}{5} = 0.63...$ et par conséquent, l'observation d'un parent nouveau apportera toujours plus d'information que l'observation d'un descendant d'un croisement (dominant x dominant).

La tactique que nous venons de décrire en fonction de q conduit donc à distinguer éventuellement trois catégories de parents :

Les récessifs,

les dominants dont le descendant examiné est récessif

les dominants dont le descendant examiné est dominant.

Soient N_1, N_2, N_3 le nombre d'individus de l'échantillon rentrant dans chacune de ces catégories : la valeur estimée de q donnée par l'équation au maximum de vraisemblance est alors :

$$\hat{q} = (2N_1 + N_2)(2N_1 + 2N_2 + 2N_3)^{-1}$$

La valeur asymptotique de sa variance est :

$$q(1-q)(1+q)^{-1}(N_1 + N_2 + N_3)^{-1}$$

TABLEAU I

No	Valeurs limites de q	No	Valeurs limites de q
1		11	0.859
	0.333		0.871
2		12	
	0.525		0.881
3		13	
	0.633		0.889
4		14	
	0.701		0.896
5		15	
	0.748		0.902
6		16	
	0.782		0.908
7		17	
	0.808		0.913
8		18	
	0.830		0.917
9		19	
	0.845		0.921
10		20	
	0.859		

TABLEAU II

Valeur de q	Valeur optimale de n
0.900	15
0.910	17
0.920	19
0.930	22
0.940	26
0.950	31
0.960	39
0.970	52
0.975	63
0.980	78
0.985	105
0.990	151
0.995	320
0.998	798
0.999	1592