# INTRODUCTION A L'ANALYSE DES COMPARAISONS POUR LE TRAITEMENT DES DONNEES EXPERIMENTALES

par

Henri ROUANET

C.N.R.S. et Université René Descartes (U.E.R. de Mathématiques, Logique Formelle et Informatique) (*)

et

Dominique LEPINE

Laboratoire de Psychologie Expérimentale, E.P.H.E. 3ème section et Université René Descartes associé au C.N.R.S. (*)

(*) 12, rue Cujas 75005 Paris

(*) 28, rue Serpente 75006 Paris

*AVANT-PROPOS*

Depuis maintenant quelque dix années, nous avons poursuivi, en liaison constante avec des expérimentalistes (travaillant dans le domaine de la psychologie), des recherches statistiques d'ordre théorique et pratique (informatique).

Les travaux théoriques ont porté sur la construction d'une approche algébrique de l'analyse des données expérimentales, dont la notion centrale est celle de "comparaison" ; c'est pourquoi nous désignons cette approche, et les procédures qui s'ordonnent autour d'elle, sous le terme d'analyse des comparaisons.

Les travaux informatiques, qui prolongent directement les recherches théoriques, ont conduit à un ensemble de programmes-machine, dont les plus importants à l'heure actuelle sont ceux de la série VAR (VAR3 et VAR4 étant les derniers en date de la série), rédigés par M-O LEBEAUX ; d'autres programmes ont été rédigés par des collègues à double formation psychologique et informatique, notamment V. Duquenne et J-M Hoc.

La plupart des développements théoriques sont exposés dans des articles, certains encore sous presse (cf. bibliographie du présent texte). Par ailleurs, la Réf. 1976a (dite "brochure verte"), constitue pour les utilisateurs du programme VAR3 (à l'heure actuelle le plus utilisé de nos programmes), le "document technique de base" *(1)*.

Cependant, étant donné le nombre croissant des utilisateurs et leur plus grande diversité, la demande a été exprimée d'un texte de caractère introductif sur l'analyse des comparaisons, lequel, rassemblant et situant les idées essentielles, pourrait faciliter l'accès aux développements théoriques et à l'utilisation des programmes. Le présent texte constitue une tentative pour répondre à cette demande *(2)*.

Dans une entreprise de cette sorte, un obstacle majeur est la diversité des lecteurs potentiels, beaucoup moins cruciale à notre avis, du

---

*(1)* Toutefois la "notice technique" (p.46 sq. de cette brochure) devra être remplacée par la nouvelle "Notice d'utilisation de programme VAR3", en date de Novembre 1977 (document disponible sur demande). Par ailleurs, une nouvelle version de la "brochure verte" est en cours de rédaction ; elle tiendra compte des derniers aménagements de VAR3 et englobera le programme VAR4.

*(2)* Une tentative antérieure était la Réf. 1976, qui, malgré sa brièveté, pourrait déjà permettre un premier "survol" (le présent texte en constitue un élargissement direct ; nous avons notamment, repris le même exemple expérimental).

point de vue de leur "niveau mathématique", que de leur expérience préalable avec la statistique et avec la problématique expérimentale (sans compter la variété de leurs attentes et de leurs intérêts). Dans la rédaction du présent texte, on a cherché à tenir compte de cette diversité. Le "texte principal" est en principe autonome, et ses seuls "préalables" ne vont guère au-délà du langage ensembliste élémentaire et d'une certaine ouverture à l'"état d'esprit expérimental" *(1)* ; l'usage d'un style "littéraire" délibérément plus illustratif que démonstratif, vise à fournir, non un substitut "adouci" des développements formalisés, mais un éclairage de ces derniers. D'autre part, en marge du texte principal, on a inséré plusieurs "aperçus théoriques", lesquels, sans prétendre résumer les développements formalisés, donneront une idée de leur style. Enfin, une longue digression, au milieu du chapitre V, s'adresse plus particulièrement aux lecteurs déjà familiarisés avec les pratiques ou théories courante de l'analyse des données expérimentales.

Il va de soi que tous les commentaires suscités par ce texte seront les bienvenus.

---

*(1)* On constatera en particulier que les bases "statistiques" (par opposition à "mathématiques") se réduisent à peu de chose ; nous sommes de plus en plus convaincus qu'un accès authentique à bien des théories statistiques dépend davantage d'une "culture mathématique générale" préalable que de connaissances proprement statistiques ; conviction reflétée dans nos textes théoriques récents, accessibles à des mathématiciens même peu "statisticiens" au sens traditionnel du terme.

## *CHAPITRE I - DE L'ANALYSE DE VARIANCE A L'ANALYSE DES COMPARAISONS*

### *Spécificité des méthodes d'analyse statistique des données expérimentales.*

On entendra ici, par données expérimentales, des données qui mettent en jeu d'une part une intervention provoquée, ou "traitement" (condition expérimentale, médicament, innovation pédagogique, etc.) et d'autre part un plan d'expérience, qui spécifie les modalités de "contrôle" des facteurs essentiels susceptibles de se traduire par un effet. La notion de "données expérimentales" s'opposera donc ici, essentiellement, à celle de données de simple observation.

L'expérimentation intervient généralement à un moment où sont considérées comme suffisamment cernées les principales variables pertinentes relatives aux phénomènes étudiés. Le but premier d'une expérimentation sera donc rarement de partir en quête de "structures cachées", et dans l'analyse statistique, les méthodes d'analyse factorielle ou typologique pourront jouer un rôle, mais qui sera subordonné à l'objectif principal : celui d'examiner et d'évaluer les effets de l'intervention provoquée ; d'où une spécificité marquée des méthodes d'analyse statistique des données expérimentales.

Par ailleurs, presque toujours, les données recueillies au cours d'une expérimentation ne portent que sur une partie de la population sur laquelle le chercheur souhaite parvenir à des conclusions ; d'où il résulte que même lorsque les procédures statistiques mises en oeuvre seront seulement descriptives, la visée de l'analyse sera essentiellement inductive (c'est-à-dire généralisante) ; normalement, cette visée s'exprimera à tervers des procédures expressément conçues pour conduire à des conclusions inductives quantitatives, c'est-à-dire inférentielles (stricto sensu, c'est-à-dire reposant sur un modèle d'échantillonnage explicite).

La pertinence de l'inférence statistique n'est certainement pas réservée aux données expérimentales, et on peut l'envisager également pour certaines données sociologiques, archéologiques, etc. Mais pour les données expérimentales, le modèle d'échantillonnage nécessaire à la validité de ces procédures sera souvent bien davantage qu'une "conjecture raisonnable" ; en effet, il peut dans une certaines mesure être incorporé à l'organisation même de l'expérimentation (notamment au moyen de l'affectation au hasard des "uni-

tés expérimentales" aux divers "traitements"), ce qui confère aux procédures inférentielles alors utilisées un degré de légitimité vraiment privilégié.

Néanmoins, cette pertinence globale et cette légitimité de la problématique inférentielle n'entraîneront nullement que les diverses méthodes inférentielles soient toujours pertinentes, ni qu'elles soient toutes pertinentes au même degré ; les tests de signification, notamment, qui constituent à l'heure actuelle la procédure inférentielle de loin la plus répandue, apparaissent bien insuffisants lorsque les objectifs de recherche amènent à des questions un peu "fines" (ainsi que nous le développerons au chapitre VII).

*L'analyse de la variance classique*

Nous désignerons ici par "analyse de la variance classique" l'ensemble des méthodes conçues par R.A. Fisher pour l'analyse des données expérimentales ; ensemble qui fut rapidement porté, par lui et ses contemporains, à un degré d'"achèvement" resté insurpassé, notamment du point de vue de l'équilibre entre les préoccupations théoriques et l'élaboration pratique des procédures ; il suffira d'évoquer ici quelques grands textes comme "Statistical methods for research workers" et "the Design of experiments" de Fisher lui-même, ainsi que "Experimental designs" de Cochran et Cox.

Les caractéristiques essentielles d'une analyse de la variance classique peuvent être résumées en quelques lignes :

- à un plan d'expérience donné, l'analyse associe une décomposition particulière, la "décomposition standard", selon les sources de variation canoniquement associées au plan : effets globaux des différents facteurs, effets résiduels, effets d'interaction, etc. ; pour chaque source de variation, on calcule une somme de carrés (que nous appellerons également : inertie), un nombre de degrés de liberté (d.l.), et un carré-moyen (rapport de l'inertie au nombre de d.l.) ;

- à chaque source de variation dont on examine l'effet, on associe une source de variation, que nous appellerons source adjointe, susceptible de lui servir de "terme de référence", et traditionnellement utilisée pour procéder à un test de signification. Dans ce but, on constitue le rapport "F de Fisher" : rapport du carré-moyen de la source examinée au carré-moyen adjoint ;

si la valeur de ce rapport est "significative" (aux seuils conventionnels usuels), on conclut (inférentiellement) à un effet de la source de variation examinée.

L'ensemble des analyses de la décomposition standard est selon la coutume rassemblé dans le tableau d'analyse de la variance familier aux expérimentalistes. Dans la sortie n° 2 de l'Annexe, la partie intitulée : "Analyse standard" constitue un exemple d'un tel tableau. (Données H & B).

*"Pratiques établies" : tableau d'analyse de la variance et analyses complémentaires.*

L'analyse de la variance classique apportait des réponses élaborées, théoriques et pratiques, à certaines préoccupations restées longtemps insatisfaites : notamment, celle de pouvoir examiner les effets de "plusieurs facteurs variant à la fois". D'où l'immense succès de ces méthodes, et le fait que pendant longtemps, le tableau d'analyse de la variance classique a constitué la base essentielle à partir de laquelle l'expérimentaliste cherchait à formuler les conclusions de la recherche effectuée.

Cependant, progressivement, sont venues d'adjoindre des analyses diverses visant à "compléter" les informations fournies par le tableau ; aujourd'hui, ces "analyses complémentaires" ont proliféré à un point tel qu'elles tendent parfois à "étouffer" le tableau classique (sans le supplanter cependant, du moins en théorie, aucune "révision déchirante" de la "doctrine" n'étant intervenue) ; de sorte que pour schématiser les "pratiques établies" actuelles, il suffira d'évoquer la juxtaposition : tableau classique et analyses complémentaires.

La mise en oeuvre de ces pratiques fait montre, à l'occasion, d'un déploiement d'ingéniosité et d'érudition tellement considérable que d'aucuns pourraient imaginer que les méthodes d'analyse statistique des données expérimentales ne laissent désormais plus rien à désirer quant à leur adéquation aux objectifs courants des recherches expérimentales. Malheureusement, un examen attentif révèle, au-delà d'apparences parfois brillantes, certaines faiblesses inattendues : notamment, le fait que dans maint mémoire expérimental, l'énoncé des conclusions ne semble avoir qu'un lien lointain, ou artificiel, avec les indications fournies par les analyses statistiques ; parfois,

le décalage est tellement flagrant qu'un lecteur candide pourrait presque s'interroger sur l'intérêt, ou la véritable finalité, de tout l'appareil statistique déployé ...

Les raisons profondes d'un tel décalage entre les pratiques statistiques et les conclusions des chercheurs n'ont pas été pour nous si immédiates à cerner, même si nous nous sommes rapidement convaincus que la source du décalage, en tout état de cause, ne pouvait être imputée aux seules ignorances ou maladresses de "praticiens" qui auraient imparfaitement assimilé la "théorie", mais qu'il fallait la rechercher bel et bien dans les insuffisances de cette théorie elle-même (Sinon, comment s'expliquer, au niveau des pratiques, autant d'"accommodements au coup par coup", compromis entre les désirs de l'expérimentaliste et les apparences de l'objectivité scientifique ?).

*La "théorie reçue" : statistique probabiliste "orthodoxe" et modèle linéaire général de la régression.*

Parallèlement (ou plus exactement : avec un léger décalage dans le temps) aux développements de l'analyse de variance classique, se développaient les constructions théoriques générales de la statistique probabiliste sous l'auto-appellation de "statistique mathématique", et parmi elles, la construction de l'école "orthodoxe" de Neyman et Pearson, dont l'emprise devait être grande dans beaucoup de domaines, voire quasi-exclusive comme dans celui des données expérimentales.

Rapidement, les procédures (*)léguées par les "pères fondateurs" se virent intégrées, "récupérées" dans une construction qui n'allait pas tarder à devenir la "théorie reçue" de l'analyse de la variance. Parmi les nombreux textes exposant cette construction, on se bornera ici à mentionner le traité de Scheffé : "The Analysis of Variance," paru en 1959, car il constitue l'entreprise d'intégration non seulement la plus complète mais sans doute aussi la plus lucide (l'auteur ne dissimulant guère les obstacles sur lesquels, malgré un arsenal technique impressionnant et impeccable, butait l'entreprise) ; la plupart des textes parus depuis ont d'ailleurs repris la conception d'ensemble de Scheffé, certains cherchant à en "simplifier la présentation", mais sans toujours parvenir à en égaler la richesse et la solidité.

---

(+) ou plus précisément : la plupart des procédures, l'exception des méthodes fiduciaires (que nous évoquerons au chapitre 8) étant vraiment bien là pour confirmer la règle.

Du point de vue de son contenu, ce qui sans doute est le plus frappant dans cette construction, c'est la portion congrue à laquelle est réduite la spécificité de la problématique expérimentale. La principale notion de base authentiquement expérimentale, celle de "matrice de plan" ("design matrix") est plongée d'emblée dans le cadre du "modèle linéaire général de la régression". Quant aux autres notions de base, elles ne renvoient pas à une construction formalisée qui serait propre à l'expérimentation, mais à des résultats généraux de statistique probabiliste (de l'école "orthodoxe"). Enfin le "style" de raisonnement est fondé sur un usage intensif du calcul matriciel, ponctué par quelques représentations "géométriques" (dans des espaces multidimensionnels), au statut analogique et toujours marginal.

Certes il est incontestablement important de montrer comment des procédures applicables à un domaine particulier peuvent être dérivées de théories plus générales ; et il serait déraisonnable de nier tous les acquis qui ont pu être obtenus par cette voie. Mais il serait non moins aventuré, à l'inverse, de dissimuler que certains concepts que cette voie a amené à mettre en avant prêtent encore largement à discussion, lorsqu'on les envisage selon une problématique centrée sur l'expérimentation.

Bien révélatrice, à cet égard, fut la discussion qui suivit un exposé de J. Nelder (Journal of the Royal Statistical Society, 1977) sur les recherches actuellement en cours chez les statisticiens de la station agronomique de Rothamsted, naguère rendue illustre par Fisher ; cette discussion opposa les statisticiens les plus éminents sur le statut de notions aussi fondamentales, pour l'interprétation des données expérimentales, que celle d'"effet global" d'un facteur en présence d'interaction, ou encore celle de "facteur aléatoire" (notion post-fishérienne, mais qu'on aurait pu croire définitivement acquise), etc.

*Un "diagnostic" ; vers une "approche algébrique de l'analyse des données expérimentales".*

Peut-on concevoir, sur le plan théorique et pratique, des procédures statistiques qui seraient davantage intégrées à la démarche expérimentale, depuis l'explicitation des objectifs, jusqu'à la formulation des conclusions à tirer des données recueillies ? Pour aborder ce problème, il nous est ap-

paru indispensable de prendre un certain recul vis-à-vis aussi bien des "pratiques établies" que de la "théorie reçue". Dans la Réf. 1968, nous formulions le "diagnostic" suivant :
"Une source essentielle des difficultés réside dans la distance entre les intentions du chercheur et les procédures de calcul. Or, nous avons constaté que les structures mathématiques sous-jacentes aux procédures se trouvent en réalité plus proches des intentions du chercheur que les procédures de calcul elles-mêmes : d'où l'intérêt d'expliciter ces structures".

Par "structures mathématiques", dans ce texte, nous entendions les structures algébriques de l'algèbre ensembliste et de l'algèbre linéaire (à quoi nous ajouterions, maintenant, les structures de la géométrie affine multidimensionnelle), c'est-à-dire, en gros, ce que les mathématiciens désignent par "structures abstraites" et le grand public par "algèbre moderne".

Le diagnostic précédent nous conduisit d'abord à dégager les fondements algébriques des méthodes existantes ; mais très vite, des perspectives apparurent : la remise en place des "fondements" amenait, non pas précisément à éliminer, mais à ramener au second plan certaines notions "de base" de la théorie reçue, et à découvrir que les structures "abstraites" pouvaient être directement efficaces, non seulement pour éclaircir la "raison d'être" de procédures existantes, mais aussi pour susciter, concrètement, l'élaboration de procédures nouvelles. Dès lors, la souplesse et la puissance des notions mises en place à partir des structures abstraites nous conduisit à aborder certaines situations expérimentales telles que celles mettant en jeu des plans non-équilibrés, ou des plans à mesures répétées, situations tout à fait courantes dans la pratique expérimentale, mais qui se montrent peu accessibles, à partir d'un certain niveau de complexité, à l'aide des seules ressources de la théorie "orthodoxe". Pour certaines de ces situations, nous avons pu proposer des procédures franchement nouvelles (cf. notamment la Réf. 1977b), lesquelles reposent aussi directement que possible sur les structures algébriques. C'est pourquoi, de plus en plus, c'est autour de la formalisation algébrique qu'il nous paraît judicieux d'ordonner tout un ensemble de procédures pertinentes à l'analyse des données expérimentales, dont la plupart, bien sûr, sont déjà dans l'analyse de variance classique (fishérienne), ou dans les apports de la statistique " orthodoxe" , mais peuvent recevoir des statuts renouvelés (comme on le verra, par exemple, au chapitre 6, pour

l'"analyse standard") dans le cadre d'une construction qui tend à se constituer en une "approche algébrique", de plus en plus compréhensive et autonome, de l'analyse des données expérimentales.

*L'analyse des comparaisons ; vue d'ensemble*

Les procédures élaborées à partir de l'approche algébrique, ou utilisées dans la perspective de celle-ci nous paraissent désormais gagner à être démarquées de l'"analyse de la variance" ; c'est pourquoi pour les désigner, nous utilisons le terme d'analyse des comparaisons ; en effet :

(1) l'idée de comparaison, avant même toute formalisation, est centrale à la problématique expérimentale, et même plus généralement à toute la statistique (*) ;
(2) nos travaux nous ont amenés à proposer une formalisation de la notion de comparaison et à placer la notion formalisée au centre de l'"approche algébrique".

Les principaux développements de l'"analyse des comparaisons" peuvent être regroupés selon les rubriques suivantes, correspondant à des niveaux de complexité emboîtés :

| | | |
|---|---|---|
| (1) formalisation ensembliste | (cf.chap.3) | qui constituent les"bases algébriques" proprement dites ; |
| (2) formalisation linéaire | (cf.chap.6) | |
| (3) formalisation statistique et structures d'échantillonnage | (cf.chap.5) | conduisant aux procédures statistiques effectives ; |
| (4) formalisations bayésienne et fiducialre | (cf.chap.7) | |

à quoi il conviendrait d'adjoindre les développements multidimensionnels, (cf.chap.8), qui "recoupe" les niveaux précédents.

(*) Cette idée forme couple avec celle de corrélation : ne pourrait-on pas regrouper sous le terme d'"analyse des corrélations" les diverses méthodes d'analyse factorielle et canonique ? Mais alors que l'idée de corrélation a fait l'objet de formalisations nombreuses, il n'en allait pas de même que celle de comparaison.

*Quelques commentaires sur l'"analyse des données"*

Dans notre entreprise d'"algébrisation", nous avons très certainement été encouragés, ne serait-ce qu'"objectivement", par le contexte "algébrisant" de nombreux travaux statistiques, ou aux implications statistiques, qui ont été effectués récemment en France, quoique la plupart dans des domaines (ou selon des perspectives) extra-expérimentalistes ; parmi ces travaux nous songeons aussi bien à nombre de textes publiés dans la revue "Mathématiques et Sciences Humaines" qu'aux divers développements des méthodes multivariées que le terme d'"Analyse des données" évoque immédiatement, à l'heure actuelle, dans le contexte français.

Mais à ce propos, et afin d'écarter certains malentendus, notamment quant à l'usage que nous faisons du terme "analyse de données", quelques commentaires s'imposent.

L'objectif de toute procédure statistique appliquée à des données est de condenser ces données en une représentation simplifiée acceptable, susceptible de conduire à des conclusions "signifiantes", interprétables ; si cette représentation est suffisamment intelligible, simple, économique (trois qualificatifs volontairement bien subjectifs !) on pourra dire que cette représentation constitue un modèle (*) de ces données. Pour atteindre cet objectif, deux démarches extrêmes peuvent être adoptées :

- dans la démarche hypothético-déductive, on choisit au départ un certain modèle (a priori plausible) et l'objectif essentiel de l'analyse statistique sera de permettre aux données de se prononcer soit en faveur de ce modèle ("validation du modèle") soit en sa défaveur ;

- dans la démarche de l'"analyse des données", on préfère laisser la représentation simplifiée des données se constituer, "émerger" au cours de l'analyse (+).

S'il fallait résumer chaque démarche en un schéma, on pourrait risquer d'écrire ceci :

Démarche { hypothético-déductive : Modèle ⟶ Données
d'Analyse des données : Données ⟶ Modèle

---

(*) le sens donné dans cette discussion au terme de "modèle" ne sera pas repris ultérieurement dans ce texte.

(+) C'est de cette façon qu'on doit, semble-t-il, interpréter le "principe" énoncé par Benzécri (in"L'analyse des données," tome 2) : "la statistique précède le modèle et non l'inverse".

Bien entendu, dans la pratique, l'opposition précédente n'est jamais aussi tranchée, ne serait-ce que parce qu'il est possible, à l'intérieur de l'une ou l'autre des deux démarches, d'user de variantes qui en amendent profondément le sens.

Ainsi, dans la démarche hypothético-déductive, on envisagera généralement non pas un seul modèle (qu'on ne pourrait que déclarer validé ou invalidé) mais un ensemble de modèles, l'objectif de l'analyse statistique devenant de choisir parmi les modèles - objectif qui rapproche incontestablement la démarche de celle de l'analyse des données.

De son côté, adopter la démarche de l'analyse des données n'implique nullement qu'on se condamne à un nombre restreint de procédures "aveugles", fixées une fois pour toutes (*), mais peut comporter, pour faire émerger la représentation des données souhaitée, des procédures visant à répondre à des questions judicieusement choisies en fonction de la situation examinée. Cette conception souple de l'analyse des données pourrait être illustrée par la déclaration suivante de Tukey : "Il faut se concentrer sur les questions et non sur les modèles" ; c'est elle dont il sera question dans le présent texte ; elle revient bien à introduire dans l'analyse des données quelque chose de la démarche hypothético-déductive, mais au niveau local de chaque question, ce qui est tout de même très différent de l'idée d'un examen d'ensemble de modèles posés a pricri.

Dans une situation expérimentale concrète, il est souvent utile d'utiliser conjointement, de manière complémentaire, l'une et l'autre démarches. Ce point ne devra pas être perdu de vue au cours de la lecture du présent texte, dans la mesure où nous adopterons systématiquement, ne serait-ce que pour la commodité de la présentation, la démarche de l'analyse des données (·)

Un dernier commentaire visant, comme les précédents, à éclairer le texte qui va suivre. Comme on sait, le terme d'"analyse des données", évoque irrésistiblement pour certains, la "statistique descriptive" ; cependant que d'autres, et non des moindres, abusent ostensiblement du langage probabiliste au cours de la démarche d'analyse des données (cf. Réf. 1976e). En réalité,

---

(*) Ainsi que pourraient tendre à l'accréditer, dans le cas de méthodes de Benzecri, les prospectus publicitaires distribués par son éditeur.

(·) Dans le texte : H. Rouanet : *Les modèles stochastiques d'apprentissage*, Mouton -Gauthier-Villars, 1967, nous avions, à l'inverse, adopté systématiquement la démarche hypothético-déductive.

selon nous, adopter la démarche d'analyse des données (ou faire jouer un rôle central à une formalisation de nature algébrique) ne préjuge en aucune façon du statut, descriptif ou inférentiel, des procédures statistiques suscitées par cette démarche.

## *CHAPITRE II - UN "PARADIGME" EXPERIMENTAL : L'EXAMEN D'UNE INTERACTION, ET UN EXEMPLE DE DONNEES EXPERIMENTALES*

Afin de rendre tangible la problématique de l'analyse des données expérimentales, nous présenterons tout de suite un "paradigme" expérimental et des données concrètes.

Le paradigme (omniprésent dans l'expérimentation) sera l'examen d'une interaction entre deux facteurs expérimentaux. Pour notre propos, il nous suffira de considérer le cas le plus simple, où chacun des deux facteurs comporte seulement 2 modalités.

Comme illustration concrète de ce paradigme, nous prendrons les données d'une expérience de Holender et Bertelson (*) que nous décrirons ici schématiquement. Il s'agissait d'une expérience de temps de réaction dans laquelle les deux facteurs expérimentaux principaux étaient : la fréquence du stimulus présenté : stimulus "fréquent" (présenté dans 75 % des épreuves), stimulus "rare" (présenté dans 25 % des épreuves) ; et la durée de la période préparatoire (délai entre la présentation d'un signal avertisseur et celle du stimulus), laquelle était tantôt "courte" (0.5 seconde), tantôt "longue" (5 secondes). L'expérience comportait trois sessions. A l'intérieur de chaque session, on faisait alterner des blocs d'épreuves à période préparatoire courte et des blocs d'épreuves à période préparatoire longue ; dans les épreuves de chaque bloc apparaissaient les deux types de stimulus.

[En fait, parmi les trois sessions, la première était considérée comme une session préliminaire d'entraînement ; c'est pourquoi la plupart des analyses seront restreintes aux données relatives aux deux dernières sessions, au cours desquelles les performances sont apparues comme à peu près stables].

---

(*) Expérience 4 de D. Holender et P. Bertelson extraite de "Selective preparation and time uncertainty", *Acta Psychologica*, 1975, 39, 193-203 - Cf. également Réf. 1977b.

Les données de cet exemple, que nous désignerons par "données H & B", nous serviront d'illustration tout au long de ce texte. L'ensemble de toutes les données analysées peut être figuré sous forme d'un tableau ayant la structure du Tableau I (*) , dans lequel on a adopté les notations suivantes :

c0,c1,c2 désignent les 3 sessions successives ;
b1 désigne la période préparatoire courte, b2 la période longue ;
a1 désigne le stimulus fréquent, a2 le stimulus rare.

Les colonnes du tableau correspondent à ce que nous appellerons les conditions (expérimentales).

Pour chaque sujet, on inscrira, pour chaque condition, les valeurs des temps de réaction observés lors des différentes épreuves. Au lieu d'épreuves, on dira aussi "répétitions", les épreuves relatives à une condition et à un sujet donné pouvant être regardées comme autant de "répétitions" d'un même épreuve fondamentale de temps de réaction.

Dans le tableau I, on a simplement indiqué le nombre d'observations, c'est-à-dire ici d'épreuves, pour le sujet s1 et les 8 conditions relatives aux deux sessions c1 et c2.

Les nombres indiqués sont ceux des observations prises effectivement en compte dans l'analyse ; au cours de l'expérience, on avait, pour chaque sujet, procédé à 192 épreuves par session : 72 pour chacune des 2 conditions mettant en jeu le stimulus fréquent a1, et 24 pour chacune des 2 conditions mettant en jeu le stimulus rare a2 ; d'où $192 \times 3 = 576$ répétitions par sujet, et pour l'ensemble des sujets $12 \times 576 = 6912$ répétitions ; mais certaines épreuves (en petit nombre) ont donné lieu à des erreurs (lorsque le sujet ne donnait pas la réponse correspondant au stimulus présenté, mais à l'autre stimulus) et, pour la présente analyse, seules ont été retenues les épreuves sans erreur.

Quelques mots sur les objectifs de cette expérience et d'expériences analogues : des recherches antérieures avaient mis en évidence l'effet de chacun des deux facteurs expérimentaux (fréquence du stimulus et durée de la période préparatoire), pris séparément, sur le temps de réaction : celui-ci est nettement plus court, d'une part lorsque le stimulus est plus fréquent, d'autre part lorsque la période préparatoire est plus courte. L'expérience

---

(*) Les tableaux et figures se trouvent à la fin du texte.

envisagée ici avait pour objectif essentiel d'examiner l'effet conjoint des 2 facteurs expérimentaux : ceux-ci agissent-ils sur le temps de réaction d'une manière additive (d'où absence d'interaction) ou non-additive (d'où effet d'interaction).

Le problème théorique plus fondamental qu'on cherche à aborder est le suivant : si on se place dans un schéma selon lequel le temps de réaction est décomposable en stades successifs (modèle "sériel" du temps de réaction), les facteurs expérimentaux ont-ils une influence sur des stades distincts ou sur des stades communs ? Dans le premier cas, les effets des facteurs seront nécessairement additifs ; en d'autres termes, on aura absence d'interaction ; dans le deuxième cas, la prédiction la plus plausible sera un effet d'interaction. D'où l'importance du "paradigme" précédent dans les recherches expérimentales portant sur le mécanisme du temps de réaction.

## *CHAPITRE III - INTRODUCTION A LA FORMALISATION ENSEMBLISTE*

Parmi les notions de base de la formalisation ensemblista, on présentera seulement, dans ce chapitre, celles qui mettent en jeu les concepts mathématiques les plus élémentaires relatifs aux relations et applications. Le notions de base plus profondes telles que treillis de finesse, facteurs saturés, plan minimal, relation de confusion entre facteurs, et facteurs superflus, sont exposées dans la Réf. 1977a ; cf. également B. et M-P . LECOUTRE, 1977, pour une illustration détaillée.

Nous commencerons par illustrer la formalisation ensembliste sur les "données H & B" présentées au chapitre II, et nous ferons suivre cette illustration d'un aperçu théorique destiné à faire entrevoir la portée générale de la formalisation.

### *Illustration de la formalisation ensembliste*

[Les termes techniques soulignés dans ce paragraphe désignent les notions dont une présentation générale sera donnée au paragraphe suivant].

En vue de la formalisation ensembliste, on considèrera qu'une valeur observable est ici un temps, exprimé par une valeur numérique (l'unité de

mesure une fois choisie : ici, la milliseconde). Comme espace d'observation on prendra donc un ensemble numérique, choisi suffisamment large pour contenir toutes les valeurs observables, par exemple $\mathbb{R}$, ensemble des nombres réels. Les données d'ensemble de l'expérience pourront être représentées comme une famille d'observations, ou protocole, qu'on notera $(x_i)i \in I$ ou, de manière équivalente, comme une application $I \longrightarrow R$ qu'on appellera application-
$i \longmapsto x_i$
protocole (de support I à valeur dans $\mathbb{R}$). Si on désigne par n le nombre des observations du protocole, on pourra toujours numéroter les observations de 1 à n, et prendre comme support du protocole l'ensemble des n premiers nombres entiers ; mais pour l'analyse, ce support serait insuffisant car il ne fait pas intervenir l'organisation des données. Pour exprimer celle-ci, on procèdera d'abord à la description des observations, laquelle sera effectuée ici à partir des 5 facteurs élémentaires suivants : fréquence du stimulus, période préparatoire, sessions, sujets, épreuves.

Le facteur "Fréquence du stimulus" sera défini comme l'ensemble {a1,a2}, dont les éléments a1 et a2 constituent les modalités du facteur, le facteur lui-même étant désigné soit par A (écriture non-indiciée), soit par $A_2$ (écriture indiciée, l'indice 2 spécifiant le nombre de modalités du facteur A). De même, le facteur "Période préparatoire" sera défini comme l'ensemble B ou $B_2$ = {b1,b2}.

Les deux facteurs A et B constitueront les "facteurs expérimentaux", liés directement aux objectifs de l'expérience. On introduira ensuite :

- le facteur "Session" ou ("Sessions") C (la lettre C pourra désigner soit l'ensemble des deux dernières sessions $C_2$ = {c1,c2}, soit l'ensemble des 3 sessions $C_3$ = {c0,c1,c2}) ;

- le facteur "Sujets", que nous noterons S ou $S_{12}$, 12 sujets ayant passé l'expérience ;

- le facteur "Epreuves" ou "Répétitions" que nous noterons R, écriture non-indiciée (cf. ci-après l'alinéa sur les conventions d'écriture) ; rappelons que le nombre de modalités du facteur R est légèrement inférieur à 6912.

N.B. : Si l'on considère les modalités d'un facteur comme totalement ordonnées, on les appellera également des niveaux ; l'ordre pourra s'imposer natu-

rellement (exemple : les niveaux du facteur Session C), ou être purement conventionnel (par exemple, pour les facteurs A et B) mais commode ; lorsque nous ordonnerons les modalités de A et B, nous adopterons l'ordre du présent numérotage).

Quant aux modalités du facteur S (les sujets), leur numérotage est a priori arbitraire ; dans la suite nous adopterons un numérotage induit par les observations effectuées (correspondant à l'ordre croissant des effets d'interaction en valeur absolue ; cf. Tableau IV).

On vérifiera que chacun des facteurs précédents satisfait à la double propriété : (1) à chaque observation on peut associer exactement une modalité du facteur (qui constitue la description de l'observation selon ce facteur) ; (2) réciproquement, à chaque modalité du facteur correspond au moins une observation. On prendra cette double propriété comme caractérisant la notion générale de facteur, ce qui permettra en particulier de définir, à partir des facteurs élémentaires envisagés précédemment, des facteurs composés.

Ainsi considérons le produit (cartésien) des deux ensembles $A_2$ et $B_2$ : $A_2 \times B_2$ = {a1b1,a2b1,a1b2,a2b2} (nous écrivons chacun des 4 couples du produit cartésien par simple juxtaposition) ; à chacun des 4 couples correspond au moins une observation, d'où il s'ensuit que le produit cartésien $A_2 \times B_2$ est lui-même un facteur, ce qu'on exprimera en disant que les deux facteurs A et B sont croisés.

[La structure de croisement des deux facteurs expérimentaux est ici nécessitée par l'objectif de la recherche, qui est l'examen de l'interaction entre ces deux facteurs].

Le facteur composé des facteurs A et B sera appelé le croisement des facteurs A et B, et noté soit A*B, soit $A_2$*$B_2$, au moyen du symbole étoile "*" désignant le croisement.

Plus généralement, à chacune des combinaisons a1b1c1, a2b1c1, ... du produit cartésien A×B×C, correspond au moins une observation ; les facteurs A,B,C seront dits croisés dans leur ensemble, et leur croisement sera le facteur composé J = A*B*C. Le nombre des modalités de J est le produit du nombre de modalités des facteurs A,B,C, d'où les écritures indiciées $J_{12} = A_2*B_2*C_3$ (en prenant C = $C_3$) et $J_8 = A_2*B_2*C_2$ (avec C = $C_2$).

Le facteur composé J que nous venons de définir sera appelé ici le facteur Condition (ou Conditions). Chaque sujet passe dans toutes les conditions, donc le facteur Sujet S est lui-même croisé avec le facteur Condition C, d'où les croisements S*J et (en remplaçant J par A*B*C) S*A*B*C.

Enfin à chaque épreuve, ou répétition, correspond exactement un sujet et une condition ; par exemple, si la répétition r399 correspond au sujet s1 et à la condition a1b1c2, on écrira r399<s1a1b1c2> ; en ce qui concerne la relation entre les facteurs R et S*J, on dira que le facteur R est emboîté dans le croisement S*J ; le facteur composé de R, S et J sera appelé emboîtement de R dans S*T, et sera noté R<S*J> au moyen du symbole chevrons "<>" désignant l'emboîtement.

On pourra, bien entendu, dans les formules, remplacer J par A*B*C, d'où les formules S*A*B*C et R<S*A*B*C>.

*Emboîtement équilibré* : lorsque les divers nombres de modalités du facteur emboîté pour chacune des modalités du facteur emboîtant sont tous égaux, on dira que l'emboîtement est équilibré ; l'emboîtement précédent R<J> est non équilibré.

*Plan du protocole* : à chaque modalité r du facteur R correspond une seule observation : nous dirons que le facteur R est un plan du protocole. Tout facteur composé dont R est un facteur composant sera également un plan du protocole ; ainsi, à la modalité r<sabc> (répétition r correspondant au sujet s passant dans la condition abc) du facteur composé R<S*A*B*C> est associée une seule observation (laquelle n'est autre, bien entendu, que celle associée à la modalité r). Tout plan du protocole est en bijection avec un support du protocole. Mais parmi tous les plans qu'on peut constituer à partir des facteurs élémentaires, le plan R<S*A*B*C>, composé de tous les facteurs élémentaires, est le plus riche, donc sera le plus "signifiant" pour l'interprétation des données.

De plus, ce plan est représentable par une formule ne faisant intervenir que des croisements "*" et des emboîtements "<>" ; nous exprimerons cette propriété en disant que ce plan est quasi-complet.

*Conventions d'écriture*

En règle générale, un facteur élémentaire sera représenté par une lettre majuscule. La lettre minuscule correspondante représentera une modalité quelconque du facteur ; une modalité particulière sera représentée par la lettre minuscule suivie du numéro de cette modalité composé de chiffres de hauteur normale. Pour un facteur composé, on écrira une modalité en juxtaposant les écritures des modalités des facteurs élémentaires (cependant, cf. les conventions relatives à l'emboîtement). La notation indiciée des facteurs élémentaires dans une formule sera toujours, dans ce texte, considérée comme facultative (*). Si dans une formule le facteur est non-emboîté, la notation indiciée consistera à faire figurer en indice le nombre de modalités de ce facteur. Si le facteur est emboîté, et que l'emboîtement est équilibré, on pourra faire figurer en indice, dans la formule de l'emboîtement, le nombre (commun) de modalités de ce facteur emboîté par modalité du facteur emboîtant. Par exemple, on pourra écrire R<S*A*B*C> (formule non-indiciée), ou $R<S_{12}*A_2*B_2*C_2>$ (formule indiciée pour tous les facteurs non-emboîtés) ou $R<S*A_2*B_2*B_2>$ (formule indiciée pour A,B,C mais non pour S) ; mais l'emboîtement étant non-équilibré, on ne pourra pas, dans la formule de l'emboîtement, indicier le facteur R.

*Autres exemples*

*1)* $E<F_3>*C_3$ : le facteur E n'est pas indicié, ce qui, selon le contexte, pourra soit signifier que l'emboîtement n'est pas équilibré, soit qu'il est équilibré mais qu'on a jugé inutile de spécifier le nombre de modalités de E par modalité de F.

*2)* $E_2<F_3>*C_3$ : cette fois l'écriture implique que l'emboîtement est équilibré (avec 2 modalités de E pour chacune des 3 modalités de F) ; le facteur E a donc en tout 2 x 3 = 6 modalités, d'où s'ensuivra, par exemple, l'écriture indiciée du croisement E*C : $E_6*C_3$, etc.

*Remarques*

1) La notion de plan du protocole (qui traduit l'organisation des données en vue de leur analyse) est distincte de celle de "plan d'expérience" (ou plus généralement de "plan de recueil des données"), mais les deux notions

---

(*) Une formule de plan (quasi-complet), moyennant l'adoption d'une règle d'énumération des observations, permettra de se donner implicitement une indexation du protocole (cette propriété est exploitée dans la conception des programmes-machines ; cf. V. Duquenne, 1977).

sont apparentées étroitement, car lorsque les données ont été recueillies selon un plan, l'organisation des données en vue de l'analyse (donc, ce que nous appelons, au sens technique, le plan du protocole), sera conditionnée, sinon entièrement déterminée, par ce plan de recueil.

2) Mais de son côté, la notion de facteur du protocole est essentiellement différente de celle des facteurs dont il est question en "analyse factorielle" ; en effet, les facteurs du protocole sont posés au départ des analyses statistiques, alors que les facteurs d'une analyse factorielle sont issus de cette analyse (pour un exemple où interviennent les deux espèces de facteurs, cf. chapitre 8).

*Autres notions de base*

1) Un facteur à une seule modalité sera appelé facteur constant ; les facteurs constants seront superflus pour l'analyse statistique, mais ils pourront être crucieux lors de l'interprétation des résultats.

2) Une partie stricte d'un facteur sera appelée facteur-partiel. Par exemple, le protocole d'ensemble des données, avec les 3 sessions, admettra $C_3$ = {c0,c1,c2} comme facteur et $C_2$ = {c1,c2} comme facteur-partiel.

*APERCU THEORIQUE SUR LA FORMALISATION ENSEMBLISTE (Réf. 1977a)*

*(1) Espace d'observation, protocole.*

La notion d'espace d'observation constitue la première notion primitive de la formalisation ; les données à traiter seront toujours représentées comme une famille d'observations $(x_i)i \in I$, c'est-à-dire par définition, une famille finie de termes $x_i$ à valeurs dans un espace d'observation $\mathcal{U}$ ($x_i \in \mathcal{U}$). Au lieu de famille d'observations, on dira aussi protocole.

Un ensemble (fini) arbitraire I indexant la famille des observations sera appelé un support du protocole. Souvent, on caractérisera le protocole par l'application-protocole $x : I \longrightarrow \mathcal{U}$, $i \longmapsto x_i$ ayant le support I pour ensemble de départ et l'espace d'observation $\mathcal{U}$ pour ensemble d'arrivée.

On appellera sous-protocole par restriction, ou simplement (dans le texte présent) sous-protocole d'un protocole, une partie de ce protocole, laquelle sera caractérisée par une restriction de l'application-protocole.

. On désignera, sous le terme général de dérivation, toute procédure permettant d'engendrer, à partir d'un protocole d'une classe donnée, un nouveau protocole, qu'on appellera protocole dérivé. La dérivation par restriction apparaît comme la plus simple des dérivations.

*(2) L'organisation des données : espace de description, descripteur, facteur, plan d'un protocole.*

La notion d'espace de description constitue la deuxième notion primitive de la formalisation ; un espace de description sera généralement un produit (cartésien) de plusieurs ensembles appelés descripteurs. Formellement, un descripteur d'un protocole sera un ensemble tel qu'à chaque observation $x_i$ du protocole on puisse associer un élément de cet ensemble (qu'on appellera la description de l'observation $x_i$ selon ce descripteur. Par exemple, si l'espace de description comporte deux descripteurs, on associera à $x_i$ les deux descriptions $f_i$ et $g_i$ ; ou, ce qui revient au même, on caractérisera la description du protocole par deux applications $f : i \longmapsto f_i$ et $g : i \longmapsto g_i$, ou encore par l'application composée $i \longmapsto f_i g_i$ ($f_i g_i$ désignant le couple des descriptions). A chaque descripteur on associera un facteur du protocole, ou brièvement facteur, défini comme l'ensemble-image de l'application correspondante ; ainsi, à l'application f on associera le facteur $F = f(I)$. En conséquence, on pourra caractériser un facteur F par une application surjective $f : i \longmapsto f_i$, dont l'ensemble d'arrivée est le facteur F. En d'autres termes, un facteur du protocole peut être défini comme un descripteur dont chacun des éléments (qu'on appellera les modalités du facteur) est la description d'au moins une observation ; ou encore : un facteur est un descripteur qui indexe une partition des observations.

Un facteur constant est un facteur qui indexe la partition grossière des observations ; un plan du protocole est un facteur qui indexe la partition la plus fine des observations (ou de manière équivalente tel que chacune de ses modalités est la description d'exactement une observation) ; un plan d'un protocole pourra être mis en bijection avec un support quelconque de ce protocole.

Si F est un facteur à plusieurs modalités et si F' est une partie stricte de F, on dira que F' est un facteur-partiel du plan. (Un facteur-partiel n'est donc pas un facteur ; mais beaucoup de propriétés des facteurs se généraliseront aux facteurs-partiels).

Etant donné un protocole, un facteur qui n'est pas un plan de ce protocole, ou un facteur-partiel, pourra être un plan d'un protocole dérivé de ce protocole et donc être mis en bijection avec un support de ce protocole dérivé (ce qu'on appellera un support dérivé).

*(3) Facteurs composés et structures ensemblistes élémentaires remarquables : emboîtement, croisement.*

La donnée de plusieurs facteurs permet de définir leur facteur composé, qu'on définit de la façon suivante : si $f : i \longmapsto f_i$ et $g : i \longmapsto g_i$ sont des applications associées aux facteurs F et G, le

facteur composé de F et G est par définition l'ensemble-image de l'application $i \longmapsto f_i g_i$.

Un facteur composant d'un facteur composé sera appelé un sous-facteur du facteur composé.

Dans ce qui suit, F,G,H désigneront des facteurs (qui pourront être des facteurs composés). Nous introduirons maintenant deux structures ensemblistes élémentaires "remarquables" : celle d'emboîtement et celle de croisement.

*Structures d'emboîtement* (*) : le facteur F sera dit emboîté dans le facteur G si chaque modalité de F correspond à exactement une modalité de G ; le facteur composé de F et G sera appelé l'emboîtement de F dans G et noté F<G> ; on appellera les symboles "<" et ">" : chevrons (commençant et finissant) (**). On remarque la dissymétrie de la structure d'emboîtement : F est le facteur emboîté, G le facteur emboîtant. Si f est une modalité de F, et g la modalité de G associée à f dans l'emboîtement, le couple fg sera noté f<g> lorsqu'on voudra expliciter la structure d'emboîtement. Les notions précédentes s'étendent aux facteurs-partiels. Si G' est une partie de G, l'ensemble des couples f<g> tels que $g \in G'$ sera appelé emboîtement de F dans G' et noté F<G'>. Si G est un facteur à plusieurs modalités et si l'inclusion de G' dans G est stricte, l'emboîtement F<G'> sera un facteur-partiel ; et si en particulier G' ne comporte qu'une seule modalité g, ce facteur-partiel sera appelé emboîtement (partiel) de F dans g et noté F<g> (écriture qui pourra être considérée comme la simplification de F<{g}> : le symbole ensembliste de l'accolade est manifestement ici inutile).

On dira que l'emboîtement F<G> est équilibré si les emboîtements-partiels F<g> ont le même nombre de modalités lorsque g parcourt G.

Enfin, on remarquera qu'avec les définitions adoptées ci-dessus, un plan sera un facteur emboîté dans n'importe quel facteur, et un facteur constant un facteur emboîtant de n'importe quel facteur.

*Structure de croisement* : les facteurs F et G seront dits croisés si pour chaque couple fg ($f \leftarrow F$, $g \leftarrow G$), on a au moins une observation ; le facteur composé de F et G sera alors appelé le croisement de F et G et noté (au moyen

---

(*) La définition donnée, dans ce texte introductif de l'emboîtement constitue une simplification de la notion plus restrictive qu'on trouvera exposée dans la Réf. 1977a.

(**) Nous avons utilisé antérieurement (par exemple dans la Réf. 1975-76), pour désigner l'emboîtement, le symbole des crochets "[ ]" (mais nous n'avons jamais utilisé dans ce sens le symbole des parenthèses "()" auquel nous réservons une signification toute différente : cf. ci-dessous, n° (5) et chapitre 4). Le remplacement des crochets par les chevrons nous permet d'utiliser désormais le même symbole dans les exposés théoriques et pour les entrées et sorties des machines (alors que les crochets n'appartiennent apparemment pas aux symboles connus de la plupart des imprimantes d'ordinateurs !). Ce remplacement nous permettra également d'utiliser les crochets comme métasymbole à valeur parenthétique dans une formule.

du symbole étoile : "*") : F*G (ou G*F, la structure de croisement étant manifestement symétrique). Cette définition s'étend immédiatement aux facteurs-partiels. On remarquera qu'un facteur constant est croisé avec n'importe quel facteur.

*Remarque* : pour exposer la notion de croisement, on pourrait très certainement se contenter du symbole du produit cartésien (classiquement la croix "x") ce que d'ailleurs nous avons fait dans certains textes antérieurs cf. Réf. 1975-76) ; le motif de l'introduction d'un nouveau symbole (l'étoile *), est de permettre, au moyen d'une seule écriture (F*G), d'une part d'exprimer que des facteurs sont croisés, d'autre part de désigner leur facteur composé (croisement).

### *(4) Généralisation : facteur composé complet, quasi-complet.*

La relation de croisement s'étend à plus de deux facteurs. Ainsi, si 2 facteurs F et G sont croisés et si le croisement F*G est lui-même croisé avec H, les 3 facteurs F,G,H seront dits croisés dans leur ensemble et leur facteur composé sera noté F*G*H. (Mais à ce propos, on signalera que trois facteurs peuvent être croisés 2 à 2 sans être pour autant croisés dans leur ensemble : v. par exemple le plan des "données Cochran & Cox" analysées au chapitre 8. Par ailleurs, si F est emboîté dans G et croisé avec H, l'emboîtement F<G> est croisé avec H ; le facteur composé de F,G et H sera noté F<G>*H.

Etant donné plusieurs facteurs, leur facteur composé sera dit :

- complet si tous les facteurs sont croisés (dans leur ensemble) ;

- quasi-complet si : d'une part, tous les facteurs composés binaires sont, soit des emboîtements, soit des croisements ; et si, d'autre part, tous les facteurs qui sont 2 à 2 croisés sont croisés dans leur ensemble. (Intuitivement, un facteur quasi-complet est un facteur "complet aux emboîtements près").

Les notions de facteur complet ou quasi-complet seront dans la suite surtout utilisées dans le cas particulier où le facteur sera un plan : (notamment un plan de protocole dérivé) : d'où les termes de plan complet, plan quasi-complet.

### *(5) Usage des parenthèses dans la formalisation ensembliste.*

Lorsque, le facteur F étant emboîté dans G, la modalité g correspond à la modalité F, on considérera l'écriture f(g) (qu'on lira : "f dans g") comme une autre écriture de la modalité f, qui explicite la structure d'emboîtement (intuitivement, la modalité f se trouve "à l'intérieur de la modalité g") ; en outre, on notera F(g) l'ensemble des modalités de F emboîtées dans la modalité g.

On distinguera donc, d'un point de vue formel, f(g) (modalité de F) et f<g> (modalité de F<G> facteur composé), et de même F(g) (facteur-partiel de F) et F<g> (facteur-partiel de F<G>).

Par ailleurs, nous ne donnerons pas ici de signification ensembliste dans une formule, à l'écriture F(G) (écriture d'un facteur à l'intérieur d'une parenthèse), à moins bien entendu, que G ne se réduise à un facteur constant {g}, auquel cas on identifierait F(G) à F(g).

*De la formalisation ensembliste aux formules de facteurs quasi-complets.*

Un facteur composé quasi-complet peut s'exprimer à partir des facteurs composants au moyen d'une formule, dont l'écriture est "linéaire" (au sens de concaténation de lettres et de symboles) et fait intervenir uniquement les symboles "<>" et "*". Exemples de formules :

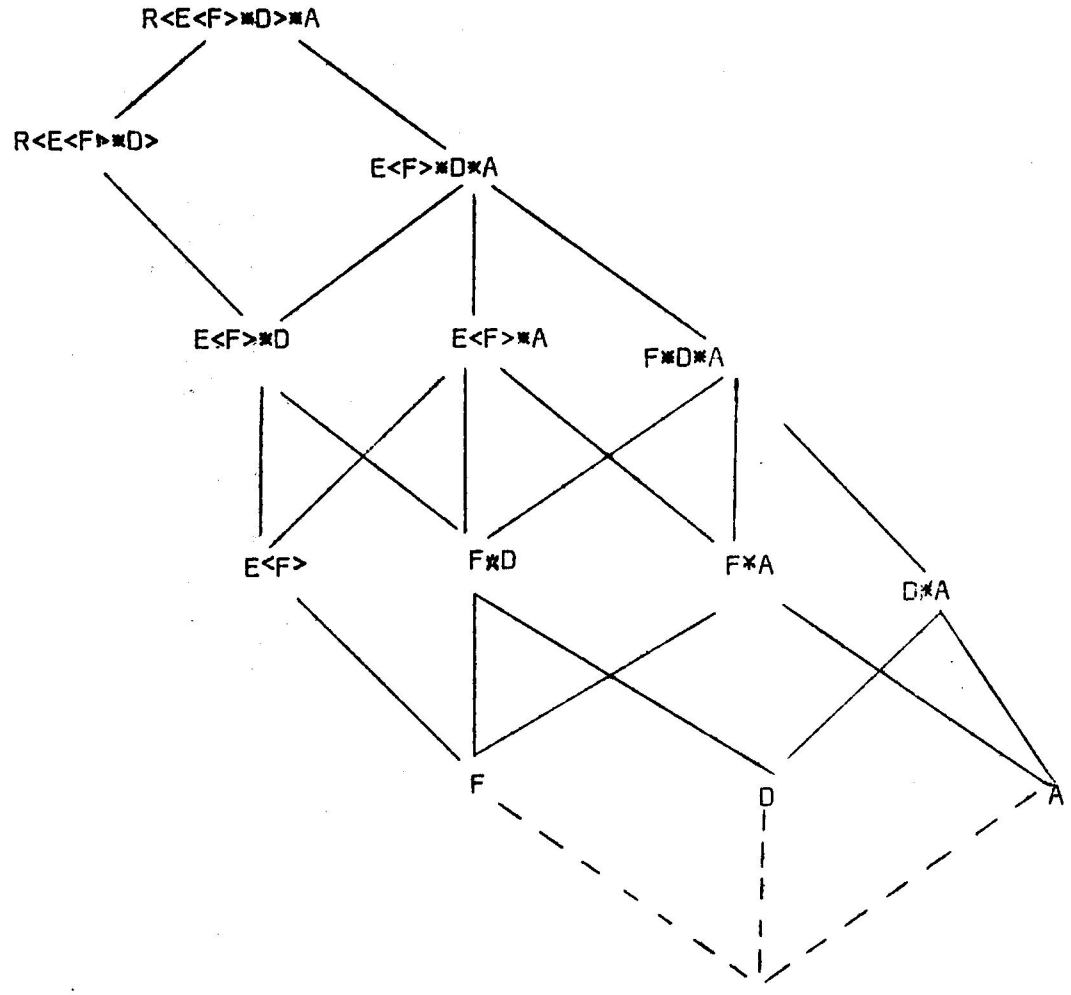
E<F> ;
R<E<F>> ;
A*B ;
E<F>*D ;
R<E<F>*D>*A , etc.

A une formule de facteur quasi-complet on associera des sous-formules (la formule du facteur constituant elle-même l'une de ces sous-formules), les quelles désigneront aussi des facteurs quasi-complets. On dira qu'une sous-formule est saturée si chacun des facteurs emboîtés de cette sous-formule y figure avec tous ses facteurs emboîtants.

L'ensemble des sous-formules saturées d'une formule sera appelé décomposition canonique de cette formule ; voici les décompositions canoniques des formules précédentes :

| Formule | Sous-formules de la décomposition canonique |
|---|---|
| E<F> | F ; E<F>. |
| R<E<F>> | F ; E<F> ; R<E<F>>. |
| A*B | A ; B ; A*B. |
| E<F>*D | F ; D ; F*D ; E<F>; E<F>*D. |
| R<E<F>*D> | F ; D ; F*D ; E<F> ; E<F>*D ; R<E<F>*D> |
| R<E<F>*D>*A | F ; D ; F*D, E<F>; E<F>*D ; R<E<F>*D>, A ; F*A ; D*A ; F*D*A ; E<F>*A ; E<F>*D*A ; R<E<F>*D>*A. |

L'ensemble des sous-formules de la décomposition canonique d'un plan quasi-complet pourra être représenté sous forme d'un treillis, qui sera le "treillis de finesse" de ce plan (mais la notion générale de treillis de finesse s'applique à un plan quelconque et pas seulement à un plan quasi-complet). Ci-dessous, le treillis de finesse correspondant à la formule R<E<F>*D>*A :




N.B. - Dans les analyses, la notion de décomposition canonique d'une formule de plan quasi-complet servira essentiellement à obtenir la décomposition canonique des effets liés à un plan quasi-complet, que nous présenterons au chapitre IV.

*Remarque*

Il peut être utile de présenter quelques exemples d'écritures qui ne désignent pas des plans quasi-complets

(1) Premier exemple : [A<B>]<C> (où les crochets sont utilisés comme symbole séparateur) ; cette écriture exprime que les 3 facteurs A,B,C vérifient les deux propriétés suivantes :

. A est emboîté dans B ;

. l'emboîtement A<B>est lui-même emboîté dans C.

Mais ces propriétés n'entraînent pas que le facteur composé de A,B,C soit quasi-complet.

Contre-exemple : considérons 2 facteurs B et C, tels que le facteur composé, qu'on notera B⊗C, ne soit ni un emboîtement ni un croisement, et supposons que le facteur A soit emboîté dans B⊗C. Les deux propriétés ci-dessus seront vérifiées, ce qui justifie l'écriture [A<B>]<C> ; et pourtant, le facteur composé de A,B,C ne sera pas quasi-complet puisque l'un des facteurs binaires B⊗C, n'est ni un emboîtement, ni un croisement.

(2) Deuxième exemple : [A*B]<C> ; cette écriture exprime que les 3 facteurs A,B,C vérifient les deux propriétés :

{ . A et B sont croisés ;
{ . le croisement A*B est emboîté dans C.

Mais ces propriétés n'entraînent pas que le facteur composé de A,B,C soit quasi-complet.

Contre-exemple : supposons que chacun des facteurs A et B soit à deux modalités ; le croisement A*B sera à quatre modalités, et de même l'emboîtement [A*B]<C>; le facteur binaire A⊗C ne pourra donc pas être à plus de 4 modalités ; si donc c est à plus de 2 modalités, A⊗C ne pourra pas être un croisement, et le facteur composé de A,B,C ne sera donc pas quasi-complet.

On peut d'ailleurs construire des contre-exemples plus "sophistiqués" ; en effet 3 facteurs A,B,C (non-constants, avec tous les trois le

même nombre de modalités, par exemple 2) peuvent être tels que :

$$[A*B]<C> ;$$
$$[B*C]<A> ;$$
$$[C*A]<B> ,$$

donc être tels que tous les facteurs composés binaires sont cette fois des croisements ; mais pour que le facteur composé de A,B,C soit quasi-complet il faudrait en outre que les trois facteurs A,B,C soient croisés ce qui est impossible car on ne peut avoir à la fois (C n'étant pas constant) : $[A*B]<C>$ et $[A*B]*C$.

(N.B. : la structure précédente n'est autre que la structure, très courante dans l'expérimentation, de carré latin construit sur 3 facteurs ; nous rencontrerons une illustration de cette structure, au chapitre VIII, avec les "données de Cochran et Cox", avec un carré latin construit sur les trois facteurs Machines, Essais et Ordres).