MATHÉMATIQUES ET SCIENCES HUMAINES

JEAN-MARC BERNARD

Analyse descriptive des données planifiées

Mathématiques et sciences humaines, tome 126 (1994), p. 6-98

<http://www.numdam.org/item?id=MSH_1994__126__6_0>

AVERTISSEMENT

Ce numéro spécial s'adresse essentiellement à ceux dont l'analyse de la variance constitue le pain quotidien, et, en tout premier lieu, aux chercheurs et étudiants en Psychologie expérimentale. Il s'inscrit dans l'entreprise de Rouanet, Lépine (1977) avec le parti-pris de mettre les aspects formels de la théorie au second plan au profit de ses aspects opérationnels : Que nous dit l'analyse des comparaisons pour, concrètement, analyser des données planifiées ? D'où l'importance accordée aux exemples et au logiciel EyeLID, un des instruments que cette théorie a engendré. De ce point de vue, ce numéro spécial relève plus du courant de "Informatique et Sciences humaines" *que de celui de* "Mathématiques et Sciences humaines".

Ce numéro est composé de sept chapitres, les références bibliographiques étant regroupées en fin de numéro. Les chapitres I et II présentent les notions de base pour comprendre, à la fois, le langage LID et la notion de protocole dérivé, *objets qui sont tous deux au coeur du logiciel EyeLID. Les chapitres III et IV traitent de l'analyse descriptive de plans de complexité croissante, et constituent, en même temps, une introduction au langage LID par l'exemple. Le chapitre V établit la jonction entre le point de vue des protocoles dérivés et celui des* comparaisons. *Le chapitre VI fournit une introduction à l'analyse inductive. Le chapitre VII, enfin, est d'avantage tourné vers la pratique; après une brève description du logiciel EyeLID, suit l'analyse de plusieurs exemples de données réelles issues de la Psychologie expérimentale, selon la ligne méthodologique, qui je l'espère, se sera dégagée des chapitres antérieurs. Le lecteur pressé d'en arriver au but pourra commencer par ce dernier chapitre, jusqu'au moment, où, peut-être, sa curiosité le ramènera à chercher à mieux en comprendre les bases.*

t Les parties de texte introduites ainsi ne sont pas indispensables en première lecture. Elles constituent en général, soit des parenthèses concernant des points plus théoriques et/ou techniques, soit des anticipations par rapport aux chapitres ultérieurs qui s'adressent au lecteur plus averti.

Les chapitres III et VII sont abondamment illustrés de sorties du logiciel EyeLID qui, on le comprendra vite, si ce n'est déjà fait, est notre outil de prédilection pour l'analyse descriptive des données planifiées. Le lecteur conquis par les possibilités du logiciel pourra se le procurer auprès de la société INDIA S.A.. Pour l'autre lecteur, en voie d'être conquis, une disquette comprenant une version de démonstration du logiciel et utilisable sur tous les exemples donnés dans le texte, est disponible gratuitement auprès de INDIA S.A. :

EyeLID–2, *version 2.04 pour système MSDOS* ©
Logiciel d'analyse de données structurées multidimensionnelles
par J.–M. Bernard, H. Rouanet et R. Baldy

édité et distribué par

Interface & Dialogue (INDIA) S.A.
22 rue de Douai, 75009 Paris
Tel. : 44.53.26.70, Fax. : 46.62.02.32

Remerciements : *Je tiens à remercier chaleureusement les collègues et amis qui ont — avec attention, intérêt, et surtout courage — relu, commenté et critiqué constructivement les premières versions de ce texte : Denis Corroyer, Tiziana Keller et Pierre Vrignaud, sans oublier Henry Rouanet, ainsi que Charlotte Carcassonne pour sa relecture attentive du texte définitif.*

Jean–Marc BERNARD

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 7–18)

I. STRUCTURE DES DONNÉES, DONNÉES PLANIFIÉES*

RÉSUMÉ — *La structure des données, décrites formellement comme un "protocole de base", et les questions guident l'analyse statistique. Dans ce chapitre, on explicite les structures et les questions typiques des données planifiées. Ce cadre servira de base aux développements des autres chapitres de ce texte : une approche descriptive de l'analyse de la variance et de l'analyse des comparaisons, dans laquelle le concept central est celui de "protocole dérivé" et l'outil privilégié est le langage "*LID*" (implémenté dans le logiciel* EyeLID*).*

SUMMARY — Structured data, planned data.
*The structure of the data, formally described as a "basic protocol", and the questions guide the statistical analysis. In this chapter, we focus on planned data, and analyse the structures and questions that arise in this context. This framework will serve as a base for what is developed in the following chapters : a descriptive view of ANOVA and of the analysis of comparisons (*ANACOMP*), where the central concept is that of "derived protocols" and the privileged tool is the "*LID*" language (implemented in the* EyeLID *computer program).*

1. DES STRUCTURES ET DES QUESTIONS AUX MÉTHODES STATISTIQUES

Tout ensemble de données est structuré, qu'il s'agisse de données d'observation ou de données expérimentales recueillies selon un plan complexe. L'explicitation de ces structures, ainsi que celle des questions qu'on pose aux données fournit un guide sûr pour le choix des méthodes statistiques pertinentes.

L'importance d'une description formalisée est clairement apparue avec la construction de l'Analyse des Comparaisons dûe à H. Rouanet et D. Lépine (1977) : formalisation ensembliste pour l'expression des structures (Lépine, 1977a et 1977b), et formalisation linéaire pour l'expression des questions (Rouanet, Lépine, 1976). Cette formalisation est riche de conséquences : d'abord elle fournit un cadre unifié pour décrire des types de données ou des problèmes apparemment différents, ensuite elle permet l'expression des structures et des questions à l'aide d'un langage. Ce langage a été l'objet de mises en oeuvre informatiques : langage des plans et des demandes de comparaisons des logiciels VAR3 (Lebeaux, Lépine, Rouanet, 1976) et PAC (Lecoutre, Poitevineau, 1992), langage d'interrogation des données (LID) du logiciel EyeLID (Bernard, Baldy, Rouanet, 1988 ; Bernard, Rouanet, Baldy, 1993).

*Ce numéro spécial de la revue *Math. Inf. Sci. hum.* n° 126, intitulé "L'analyse descriptive des données planifiées" est composé de sept chapitres, dont celui-ci. Il a été rédigé par J.-M. Bernard, Groupe Mathématiques et Psychologie, CNRS et Université René Descartes, Sorbonne, 12 rue Cujas, 75005 - Paris.

Dans cette section, nous rappelons brièvement les grandes lignes de cette formalisation ; ce cadre général nous permettra de préciser l'objet de ce texte. Un certain nombre de notions, introduites "au passage", seront définies au §2..

1.1. Structure des données : protocole, support U, espace d'observables V

Considérons un questionnaire dans lequel un ensemble d'individus (IND) est interrogé sur son salaire (SAL) et son niveau d'études (NIV). Le terme "*protocole*" désigne la description formalisée d'un tel ensemble de données, comme une *application* "U->V", d'un ensemble U d'*unités statistiques*, appelé le *support* du protocole, ici U=IND, vers un *espace d'observables* V, ici le produit cartésien, V=SAL × NIV : à chaque individu de IND, on associe un couple de valeurs de SAL × NIV. Les deux ensembles U et V peuvent chacun être munis de structures particulières :

- Le support U peut être un ensemble d'individus, comme ici, mais aussi de villes, d'entreprises, *etc.*. Il peut également posséder une structure plus complexe. Si on s'intéressait aux différences de salaires selon le sexe et l'année, on considérerait des individus *emboîtés* dans leur sexe (SEX) et *croisés* avec un ensemble d'années (AN). Le support serait alors un ensemble de "Individus-Années", décrit par les *facteurs* IND, SEX et AN, ce qu'on noterait : "IND<SEX>*AN".
- L'espace d'observables V peut être *univarié* ou *multivarié* : ici, on a 2 variables, SAL et NIV. On distingue divers types de variables : binaires (ou dichotomiques, ou en $\{0,1\}$), catégorisées (ou nominales), ordinales, métriques, ou numériques (Rouanet, Le Roux, Bert, 1987). La spécification de l'*échelle de mesure* détermine les procédures statistiques applicables.

Bien entendu, on peut simplement baptiser "variable" tout ce qui est susceptible de varier. Tous les ensembles mentionnés sont de telles "variables au sens large". Mais si, à une étape de l'analyse, la *question* posée porte sur les variations du salaire selon le sexe et le niveau d'études, pour une année donnée, on introduira une dissymétrie entre des *variables explicatives* (IND, SEX et NIV), et des *variables à expliquer* (SAL), et on décrira le protocole par l'application : "IND<SEX&NIV> -> SAL". Ainsi, la structure des données n'est pas figée dès leur recueil, et peut dépendre des questions qu'on se pose.

1.2. Les questions et les méthodes

Schématiquement, on peut distinguer deux grandes catégories de questions, intimement liées, comme on vient de le voir, à la structure du protocole adoptée à un instant donné :

- Lorsqu'il y a beaucoup de variables explicatives (U complexe), celles-ci sont considérées comme des *facteurs structurants*, ou simplement des *facteurs*, et un type de question privilégié est l'étude de l'*effet de ces facteurs* sur les variables à expliquer. Ceci conduit aux méthodes d'*analyse de la variance* (ANOVA) et de l'*analyse des comparaisons* (ANACOMP).
- Lorsqu'il y a beaucoup de variables à expliquer (V complexe), un type de question privilégié est l'étude des *corrélations* entre variables. On recherche alors de nouvelles variables résumant les variables initiales, par des méthodes de *réduction dimensionnelle* : *e.g. analyse des correspondances* et *analyse en composantes principales*.

Ces deux catégories de questions — comparaison et corrélation — ne sont pas exclusives, U et V pouvant être tous deux complexes, et certaines méthodes participent des deux approches à la fois : en particulier l'*analyse des données multidimensionnelles* (Bernard *et al.*, 1989 ; Rouanet, Le Roux, 1993) et l'*analyse de la variance multivariée*.

1.3. Données d'observation et données expérimentales

Du point de vue structurel, les deux types de données — d'observation et expérimentales — peuvent être décrits avec la formalisation ensembliste esquissée précédemment.

Une des différences principales entre une *observation* et une *expérimentation* est, que dans cette dernière, on cherche à *contrôler* un certain nombre de variables explicatives *a priori* (*e.g.* en décidant de faire varier systématiquement leurs modalités, ou en les maintenant constantes) dans le cadre d'un *plan d'expérience* [1]. De ce fait, dans l'expérimentation, la structure des données est presque entièrement imposée par le plan de recueil des données. Au contraire, dans l'observation, certaines variables peuvent être considérées comme explicatives ou à expliquer, selon la question posée à une étape donnée de l'analyse.

Ceci a une conséquence importante sur l'interprétation des résultats. Le contrôle des facteurs dans l'expérimentation, notamment lorsque leurs modalités sont *affectées* aux sujets, permet l'interprétation des effets trouvés en termes de causalité. Par contre la notion d'"effet" pour des données d'observation n'a en général qu'un sens métaphorique.

1.4. De la description à l'induction

Une première étape pour répondre à une question donnée est l'*étape descriptive*, dans laquelle on cherche à résumer les données, à en fournir une description condensée. A l'issue de cette étape, on disposera de conclusions synthétiques mais qui portent seulement sur les données observées. Cependant, souvent, les données ne constituent qu'une partie d'une population plus vaste de données potentielles. Le but de l'*étape inductive* est alors de généraliser les propriétés trouvées au niveau descriptif, soit à des données futures, soit à l'ensemble de la population. La distinction des deux étapes, descriptive et inductive, est essentielle (Rouanet, Bernard, Le Roux, 1990 ; Rouanet *et al.*, 1991).

2. STRUCTURATION DES DONNÉES PLANIFIÉES

Le panorama général que nous venons d'esquisser permet de préciser l'objet de ce texte : *l'analyse descriptive des données planifiées*. Par *données planifiées*, on entend des données où la dissymétrie entre variables explicatives et variables à expliquer est posée au départ. Cette dissymétrie provient toujours des questions qu'on pose aux données, que ces questions président à leur recueil — comme dans les recherches expérimentales, ou dans les recherches d'observation à caractère confirmatoire —, ou qu'elles soient posées *a posteriori* — dans le cadre d'une recherche d'observation à caractère exploratoire. Ainsi, même si les exemples traités dans la suite de ce texte relèvent surtout de l'expérimentation, le champ d'application des méthodes exposées est plus vaste.

L'accent sera donc mis sur les données dans lesquelles le support U est complexe; mais on envisagera aussi bien des espaces d'observables univariés que multivariés. Les variables explicatives seront ici toujours catégorisées, et les variables à expliquer toujours numériques. Selon l'usage en expérimentation, on parlera de *facteurs* (ou quand il s'agit de variables manipulées, de *variables indépendantes*) pour les premières, et de *variables dépendantes* ou simplement de *variables* pour les secondes.

[1] Mais, comme l'argumente Reuchlin (1992), le contrôle parfait est chose rare, et la plupart des recherches se situent plutôt, en fait, sur un continuum dont les deux pôles extrêmes sont "observation" et "expérimentation".

La notion de *protocole* est simplement la formalisation d'un *ensemble de données à traiter*. Dans cette section nous rappelons les éléments qui interviennent dans cette description formelle, (pour plus de détails, voir Ehrlich, 1975 ; Hoc, 1983 ; Rouanet, Lépine, 1977). Deux concepts sont ici essentiels : la *structure ensembliste du protocole*, *i.e.* la structure du tableau dans lequel on peut ranger les données, qui sera exprimée par le *plan du protocole* ; et sa *structure statistique* qui ouvre la voie de l'induction.

L'ensemble de données à traiter est qualifié de *protocole de base* lorsqu'il est considéré comme primitif. Comme on le verra, une question spécifique amène souvent à en considérer seulement une "partie" ou un "résumé", *i.e.* un *protocole dérivé*. Dans cette première section, il n'y a pas lieu de distinguer et on parlera de "protocole" sans précision.

Nous illustrerons ce texte à l'aide de deux dossiers expérimentaux, le dossier "Négligence" présenté ci-après et le dossier "Horloge" présenté au §2.4.. Ils serviront d'abord à illustrer les chapitres plus théoriques (I à VI), mais seront aussi l'objet d'analyses plus poussées au chapitre VII.

2.1. Le dossier "Négligence"

Cette recherche porte sur la "pseudo-négligence" (Chokron, Imbert, 1993) qu'on observe chez des sujets normaux et qui présente des similarités avec l'"héminégligence" de sujets atteints d'une lésion cérébrale. La tâche des sujets consiste à déterminer le milieu subjectif d'une baguette de $24cm$ avec la seule aide d'informations kinesthésiques. La pseudo-négligence se traduit par une déviation systématique vers la droite (pour des droitiers) de ce milieu subjectif par rapport au milieu objectif de la baguette.

Les données portent sur 24 sujets (des femmes droitières) répartis selon 2 conditions (12 sujets pour chacune) : "active" (`c1`) où le sujet peut librement déplacer son doigt posé sur un curseur mobile le long de la baguette ; ou "passive" (`c2`) où le sujet commande un moteur déclenchant le mouvement de la baguette dans un sens ou dans l'autre, alors que son doigt ne bouge pas. Chaque sujet exécute cette tâche dans 6 situations expérimentales obtenues par le croisement de : la main utilisée, gauche (`m1`) ou droite (`m2`) ; et l'orientation du regard, 30° à gauche (`o1`), 0° (`o2`) ou 30° à droite (`o3`). Pour chaque sujet et chaque situation, on mesure la déviation en *cm* (notée `DEV`) entre le milieu subjectif et le milieu objectif de la baguette. Une déviation à droite est notée par une valeur positive, à gauche par une valeur négative. L'objectif principal de l'expérience est l'étude de l'effet de la condition sur la déviation, et des possibles variations de cet effet selon l'orientation. Le tableau des données se présente sous la forme suivante :

		m1			m2		
		o1	o2	o3	o1	o2	o3
c1	s1	1.95	0.95	0.55	0.15	-0.80	-0.65
c1	s2	3.00	3.10	**1.55**	-0.10	-0.30	-0.60
...	...	...	...	...	...	...	...
c1	s12	1.80	1.65	0.55	1.00	-1.00	1.30
c2	s13	-0.30	-0.10	-0.55	1.30	-1.90	0.75
c2	s14	1.40	-1.00	0.95	-0.20	0.00	-0.50
...	...	...	...	...	...	...	...
c2	s24	0.20	-0.45	-0.80	0.65	2.20	-0.10

2.2. Description d'un protocole, facteurs, variables

2.2.1. *Facteurs et variables*

Décrire la structure du protocole consiste d'abord à distinguer les *facteurs* et les *variables*.

- Les *facteurs* sont parmi les "variables au sens large" celles qui décrivent les conditions d'observation. Ici les facteurs sont au nombre de 4 :

  S, les sujets à 24 modalités, (s1,s2,..,s24);
  C, les 2 conditions (c1 active, c2 passive);
  M, les 2 mains (m1 gauche, m2 droite);
  O, les 3 orientations (o1, 30° à gauche; o2, 0°; ou o3, 30° à droite).

- Les *variables dépendantes*, ou simplement les *variables*, sont parmi les "variables au sens large" celles dont on observe les variations d'une condition à l'autre. Ici on a une seule variable numérique, notée DEV, la déviation entre milieu subjectif et objectif mesurée en *cm*, qui peut valoir entre $-12cm$ et $12cm$ (de gauche à droite).

Aux facteurs déjà indiqués, on peut toujours adjoindre le facteur U qui indexe les 144 unités (24 sujets × 6 situations), et est appelé le *support* du protocole. Une description complète des conditions d'observation doit également faire intervenir tous les facteurs maintenus constants dans l'expérience; ici, il y aurait par exemple le sexe des sujets (féminin), la longueur de la baguette ($24cm$), *etc.*. Il est commode de désigner l'ensemble de ces *facteurs constants*, par un facteur unique Z à une seule modalité z1, qu'on pourra lire "unité ayant participé à l'expérience". Les deux facteurs U et Z jouent un rôle important à la fois sur le plan formel et sur le plan méthodologique.

2.2.2. *Application-protocole et application-description*

Le protocole est décrit par deux applications : l'*application-description* va de U vers le produit-cartésien des facteurs, et associe à chaque unité sa description selon ces facteurs; l'*application-observation* va de U vers l'espace d'observables V (ici V=DEV), et associe à chaque unité une valeur observée (numérique ou multinumérique si V est multivarié) [2]. On intègre ces deux composantes dans l'*application-protocole* : "U<S&C&M&O> -> DEV". A chaque unité (*i.e.* u9), décrite par une certaine combinaison de modalités (s2c1m1o3), on associe une valeur observée (1.55) (cf. tableau des données, p.10). Par rapport au tableau des données, où chaque "case" correspond à une unité, ces deux applications ont une interprétation simple : l'application-description décrit la structure des marges du tableau, l'application-protocole décrit le contenu des cases.

2.3. Des facteurs élémentaires au plan

2.3.1. *Facteurs élémentaires*

Les facteurs précédents, donnés dans la description du protocole, sont dits *élémentaires*. Par facteur élémentaire, on entend, à la fois l'ensemble de ses *modalités*[3], par exemple C={c1,c2}, et aussi l'application qui va du support U vers cet ensemble de modalités. De façon équivalente, le facteur se définit par un ensemble de classes disjointes, une par modalité, d'éléments de U; c'est la *partition* de U associée au facteur; ainsi, pour le facteur C, les 144 unités se divisent en 2 classes, les 72 premières unités décrites par la modalité c1, et les 72 autres décrites par c2.

[2] La notion générale de protocole fait également intervenir des *poids* associés aux unités (cf. II.§2.3.).
[3] Quand les modalités sont ordonnées, on parle aussi de *niveaux*.

2.3.2. *Facteurs composés*

Un *facteur composé* de deux facteurs élémentaires est l'ensemble des *modalités composées* des deux facteurs ; par exemple le facteur composé de M et de O, noté à l'aide du symbole de la *composition* '&' (lire "et" ou "composé avec"), M&O, est l'ensemble : {m1o1,m1o2,m1o3,m2o1,m2o2,m2o3}. Un facteur composé définit aussi une partition de U, ici en 6 classes de 24 unités de base chacune. La notion de facteur composé s'étend à la composition d'un nombre quelconque de facteurs élémentaires, qu'on notera C&M&O, S&C&M&O, *etc.*. Formellement, il n'y a pas de différence entre facteur élémentaire et facteur composé : à chacun est associé un ensemble de modalités et une partition des unités de base. Ainsi, par la suite, le mot "facteur" sans précision désignera tout facteur élémentaire ou composé.

2.3.3. *Relations entre facteurs : croisement, emboîtement, confusion*

Le facteur composé M&O comprend *toutes* les combinaisons possibles de modalités de M et de modalités de O. Dans ce cas, les deux facteurs M et O sont dits *croisés*, et le facteur composé M&O est appelé un *croisement*, ce qu'on note M*O.

Le facteur composé S&C a autant de modalités que S seul (24), puisque chaque sujet est affecté à une seule condition. Ceci définit la relation non symétrique d'*emboîtement* : "S est emboîté dans C", ce qu'on note S<C>. Les éléments de S peuvent être rangés dans des "boîtes" étiquetées par chacune des modalités de C. Ici l'emboîtement est *équilibré* : il y a autant de sujets (12) dans chacune des 2 conditions.

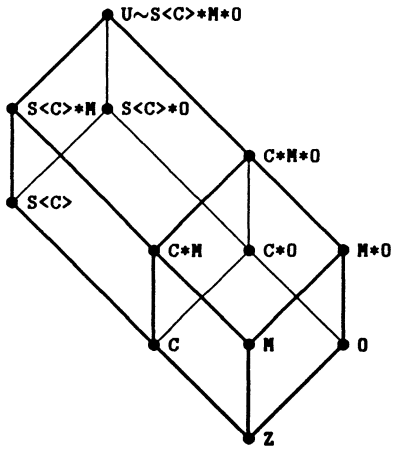
Deux facteurs sont en relation de *confusion*, notée '∼' (lire "confondu avec"), s'il existe une bijection entre eux. Dans le cas d'une relation de confusion A∼B, un des facteurs A ou B peut être considéré comme *superflu* : on peut l'"oublier" pour l'analyse (mais non pour l'interprétation des résultats !).

L'étude de la structure du protocole se fait en considérant les relations entre facteurs élémentaires pris 2 à 2, puis 3 à 3, *etc.*. Pour les relations binaires, on trouve ici : S<C>, S*M, S*O, C*M, C*O, M*O. Pour les relations d'ordre supérieur, on trouve : S<C>*M, S<C>*O, S*M*O, C*M*O, puis S<C>*M*O [4].

2.3.4. *Partitions, treillis de finesse des partitions*

Un facteur décrit le protocole avec une certaine *finesse* : le nombre de classes dans lesquelles il permet de ranger les unités. Le support U est, par définition, un facteur d'une finesse extrême : il divise les 144 unités en 144 classes comprenant une unité chacune. A l'opposé, le facteur Z est le plus grossier : il range les 144 unités en une seule classe de 144 unités. Chaque autre facteur est intermédiaire entre U et Z. En comparant entre eux les facteurs quant à leur finesse ou leur grossièreté réciproque, on définit le *treillis de finesse des partitions* (Duquenne, 1986 ; Lépine, 1977a ; Duquenne, Monjardet, 1982) présenté à la figure ci-après.

[4] Un certain nombre de règles formelles, découlant de propriétés ensemblistes, permettent de passer des relations binaires à certaines relations ternaires : A<B> et A*C ⇒ B*C et A<B>*C ; A<B> et B<C> ⇒ A<C> et A<B<C>> ; A<B>, A<C> et B*C ⇒ A<B*C>. Par contre, le croisement 2 à 2 de plusieurs facteurs (A*B, A*C et B*C) n'implique pas le croisement global (A*B*C) ; un contre-exemple en est fourni par la relation de *carré-latin*. Une situation typique de carré-latin est celle de la passation de deux conditions avec "contrebalancement des ordres" : les facteurs "condition", "ordre" et "essai" forment alors un carré-latin.



t Un treillis sur un ensemble (ici celui des partitions associées aux facteurs) est une relation d'ordre partiel (ici la finesse) qui posséde un suprémum (ici U), et un infimum (ici Z) : pour tout facteur A, on a U&A~U et A&Z~A. L'étude des relations entre facteurs est équivalente à celle du treillis : le facteur composé A&B est le premier (en montant) élément du treillis à la fois plus fin que A et plus fin que B ; A<B> ssi A est plus fin que B, *i.e.* ssi il y a un chemin toujours descendant de A vers B ; A~B ssi les partitions associées à A et à B sont identiques ; A*B ssi la partition associée à A&B possède A×B classes.

2.3.5. *Notation indiciée des facteurs*

La notation indiciée d'un facteur permet de retrouver son nombre de modalités :

S24 Indique que le facteur élémentaire S a 24 modalités ;

M2*O3 Indique que le facteur M a 2 modalités, et O, 3 modalités. Le produit des indices donne le nombre de modalités du facteur composé M*O : $2 \times 3 = 6$.

S12<C2> Indique que le facteur élémentaire S est emboîté dans le facteur C à 2 modalités, avec 12 modalités de S par modalité de C. En cas d'emboîtement déséquilibré, l'indice de S serait omis. A nouveau, le produit des indices donne le nombre de modalités du facteur composé S<C> : $12 \times 2 = 24$.

2.3.6. *Plan d'un protocole*

Un plan d'un protocole est un facteur associé à la partition la plus fine, *i.e.* qui comporte autant de modalités que U. Bien entendu U constitue toujours un plan, mais on cherchera plutôt un plan *riche*, *i.e.* qui fait intervenir le plus grand nombre de facteurs élémentaires, ici S<C>*M*O. Un plan correspond à une description des observations qui permet d'identifier chacune d'elles de façon unique. Avec la notation indiciée du plan, S12<C2>*M2*O3, on voit que le facteur composé S&C&M&O comporte $12 \times 2 \times 2 \times 3 = 144$ unités, comme U.

Un plan est un *plan quasi-complet (PQC)* s'il peut être décrit par une *formule unique* qui ne fait intervenir que des croisements et/ou des emboîtements. Pour les PQC, un plan correspond à une simple lecture de la structure des marges du tableau de données : le tableau p. 10 se présente comme le croisement de 24 lignes (facteur S<C>) et de 6 colonnes (facteur M*O), d'où le plan "S<C> * M*O".

2.3.7. *Structure ensembliste du protocole*

On résume la structure du protocole en décrivant le support U par un plan (voire plusieurs plans) aussi riche que possible et en indiquant, à droite de la flèche '->', la ou les variables dépendantes, soit ici :

S12<C2>*M2*O3 -> DEV

On parlera, pour cette structure, de la *structure ensembliste des données* — les données sont décrites à l'aide des ensembles que constituent les facteurs et les variables —, à distinguer de leur *structure statistique* (cf. §2.7.).

2.4. Le dossier "Horloge"

Le second dossier qui illustre ce texte porte sur l'exploration mentale d'un environnement imaginé (Amorim, Stucchi, 1994). Le sujet voit la lettre "F" sur un écran (tridimensionnelle, projetée en 2 dimensions) et doit imaginer que la lettre se trouve au centre d'une horloge. Dans la condition `c1` "sujet-centrée", on indique au sujet l'heure à laquelle il se trouve sur l'horloge, et la tâche consiste à donner l'heure à laquelle pointe la lettre ; dans la condition `c2` "objet-centrée", on indique l'heure à laquelle pointe la lettre et le sujet doit donner l'heure à laquelle lui-même se trouve. Les 24 sujets de l'expérience passent chacun les deux conditions, avec 12 essais par condition qui correspondent au croisement de : 6 angles que peut faire la lettre avec le sujet (15°, 45°, 75°, 105°, 135°, 165°) ; et 2 côtés possibles, gauche ou droit [5]. Enfin les sujets sont répartis en 4 groupes, de 6 sujets chacun, obtenus par le croisement de : 2 dimensions suggérées de l'horloge, $3m$ ou $30m$; et 2 ordres de passation des 2 conditions. A chaque essai, on mesure : l'erreur (`ERR`) en degrés, une erreur positive indiquant une surestimation ; et le temps de réponse (`TR`). Les objectifs de l'expérience portent avant tout sur l'étude du facteur "angle", et ensuite sur celle des facteurs "condition" et "dimension".

Les facteurs du protocole sont les suivants :

`S`, les 24 sujets ;
`C`, les 2 conditions ;
`E`, les 12 essais ;
`A`, les 6 angles ;
`L`, les 2 côtés (latéralité) : gauche (`l1`) et droit (`l2`) ;
`G`, les 4 groupes de sujets ;
`D`, les 2 dimensions : $3m$ (`d1`) et $30m$ (`d2`) ;
`O`, les 2 ordres : `c1-c2` (`o1`) et `c2-c1` (`o2`).

Un premier plan découle de ce que les Sujets, les Conditions et les Essais sont croisés dans leur ensemble, `S*C*E`, ce qui représente $24 \times 2 \times 12 = 576$ unités. On obtient les autres plans en tenant compte des emboîtements et des confusions entre facteurs : `S<G>`, "chaque sujet n'appartient qu'à un seul des 4 groupes" ; `G~D*O`, "les groupes correspondent au croisement de la dimension et de l'ordre" ; et `E~A*L`, "les essais correspondent au croisement de l'angle et du côté". En omettant les facteurs `G` et `E`, superflus et introduits uniquement pour plus de clarté, le plan le plus riche est quasi-complet : `S<D*O>*C*A*L`. La structure ensembliste du protocole est ainsi :

```
S6<D2*O2>*C2*A6*L2 -> ERR,RT
```

2.5. Aspects méthodologiques liés à la structuration des données

2.5.1. *Statut des facteurs ; structure statistique des données*

Au niveau ensembliste, seul aspect considéré jusqu'ici, tous les facteurs jouent des rôles analogues (à part les facteurs `U` et `Z`). Des considérations d'ordre méthodologique amènent à introduire des distinctions entre facteurs élémentaires qui seront essentielles lors de l'analyse des données (pour plus de détails, voir *e.g.* Ehrlich 1975) :

La première distinction renvoie aux objectifs de l'expérience : les *facteurs principaux* sont ceux impliqués dans les hypothèses ou les questions de l'expérience ; les *facteurs secondaires*, les autres :

- Dans le dossier "Négligence", les facteurs principaux sont d'abord la "condition" `C`, l'objectif essentiel étant de comparer les conditions "active" et "passive" ; puis

[5] Les données initiales comprennent en fait 144 essais, car on a ici moyenné sur un facteur secondaire "heure" à 12 modalités, l'heure indiquée (position du sujet en `c1`, ou celle de la lettre en `c2`).

l'"orientation" O, un second objectif étant d'étudier d'éventuelles variations de l'effet de C selon O. Les facteurs M et S sont secondaires.
- Dans le dossier "Horloge", les objectifs de la recherche portent d'abord sur le facteur A, puis sur les facteurs C et D ; ces trois facteurs sont donc principaux. Les facteurs S et O sont clairement secondaires. Enfin, le statut du facteur L est intermédiaire; il n'y a pas d'hypothèse précise le concernant, mais plutôt une question ouverte.

La seconde distinction tient au mode de recueil des modalités d'un facteur : les modalités d'un *facteur systématique* sont choisies pour elles-mêmes et non échangeables ; les modalités d'un *facteur de groupe* sont échangeables, et constituent en général une partie d'une population plus vaste. Dans les deux dossiers "Négligence" et "Horloge", seul S (les "sujets") est facteur de groupe, comme souvent en Psychologie Expérimentale. Les modalités d'un facteur de groupe sont traitées de façon symétrique. Lors de l'étape descriptive, cela se traduira par le calcul de *statistiques de groupe*, *e.g.* moyenne, variance, *etc.* (Rouanet, Lépine, 1973). Lors de l'étape inductive, on cherchera à généraliser les conclusions à la population.

L'adjonction, à la structure ensembliste du protocole, de la distinction entre facteurs de groupe et facteurs systématiques, définit sa *structure statistique*.

2.5.2. *Facteurs constants et confusion de facteurs*

Les facteurs constants n'interviennent pas dans l'analyse mais leur prise en compte est essentielle pour permettre de dégager la portée des conclusions descriptives ou inductives de l'expérience. Par exemple, dans le dossier "Horloge", les conclusions ne pourront, en toute rigueur, être émises que pour la lettre "F", et ne sauront être étendues à d'autres lettres, voire à d'autres situations, qu'en recourant à des hypothèses psychologiques provenant par exemple de résultats expérimentaux antérieurs.

Dans le cas où certains facteurs sont confondus entre eux, l'interprétation se heurtera à la difficulté, voire l'impossibilité, d'attribuer un effet à un facteur plutôt qu'à un autre. Un problème similaire intervient si certains facteurs sont corrélés entre eux (la confusion en est un cas extrême) ; ceci se produit notamment pour tous les facteurs dont les modalités sont seulement *observées* et non *affectées* aux sujets, et qui sont, par nature, confondus ou corrélés avec de nombreux autres (voir *e.g.* Abdi, 1987, pp. 26–35 ; Reuchlin, 1992).

2.5.3. *Objectifs de la recherche et planification*

Lors de la planification d'une expérience se posent certains choix cruciaux, qui sont éclairés par les considérations précédentes. Nous ne faisons ici que les évoquer :

- Choix des modalités des facteurs systématiques : d'un choix trop "étriqué" parmi toutes les valeurs possibles d'une variable peut résulter une sous-estimation de son effet, un choix trop "large" risque au contraire de masquer l'effet d'autres facteurs.
- Choix du plan : l'idéal de tout expérimentaliste est le plan complet, qui permet d'étudier l'effet de chaque facteur pris isolément, et toutes leurs interactions. Mais cet idéal est rarement réalisable. Il peut s'agir d'une impossibilité *logique* : les sujets sont nécessairement emboîtés dans leur sexe; d'une impossibilité *méthodologique* : l'effet d'apprentissage d'une condition à l'autre, oblige à préférer l'emboîtement des sujets dans les conditions, ou bien à introduire un facteur "ordre de passation" dans lequel les sujets sont nécessairement emboîtés ; ou enfin, d'une impossibilité *pratique* : l'idéal conduirait à un trop grand nombre d'observations.

2.6. Des données brutes aux données analysées : plan d'expérience et plan d'analyse

A partir des données brutes, caractérisées par un *plan d'expérience* aussi détaillé que possible, on se ramènera souvent, à une certaine étape de l'analyse, à un *plan d'analyse* plus simple. Cette simplification pourra consister en l'élimination de facteurs superflus ou secondaires. De premières analyses pourront conduire à considérer l'expérience comme un ensemble de plusieurs expériences qu'on traitera séparément, et on restreindra alors l'analyse à un sous-ensemble de modalités d'un facteur, voire à un sous-ensemble de sujets, à certains facteurs ou à certaines variables. On pourra également être amené à recoder certaines variables initiales en des variables plus pertinentes.

2.7. Structures ensemblistes et statistiques envisagées dans ce texte

Dans ce texte, nous envisagerons principalement des protocoles dont la structure ensembliste est un plan quasi-complet, avec une ou plusieurs variables dépendantes numériques. On considérera des protocoles comprenant un seul facteur de groupe (noté génériquement S, pour "Sujets"), avec, en particulier, les structures statistiques remarquables suivantes (les facteurs G et T peuvent être élémentaires ou composés ; au lieu de "Traitement" pour le facteur T, on pourra lire selon les cas "Occasions", "Essais", *etc.*) :

`S<G>`	des "Sujets" emboîtés dans des "Groupes" ;
`S*T`	des "Sujets" croisés avec des "Traitements" ;
`S<G>*T`	des "Sujets" emboîtés dans des "Groupes" et croisés avec des "Traitements".

2.7.1. *Déclarer les structures : le langage des plans*

Le langage qui vient d'être introduit constitue un *langage de déclaration de la structure du protocole de base*, ou *langage des plans*. Les logiciels dont il va être question dans la section suivante connaissent (à des degrés divers) ce langage :

- Dans le logiciel VAR3 par exemple, ce langage des plans fait intervenir : facteurs indiciés, relations '*' et '<>', et désignation par S d'un unique facteur de groupe.
- Dans le logiciel EyeLID, la structure n'est pas "déclarée", et chaque unité statistique est décrite en terme de modalités de chacun des facteurs ; ainsi la structure ensembliste peut être quelconque (les facteurs peuvent être en relation quelconque '&'). Il y est néanmoins possible de vérifier l'existence de relations particulières entre facteurs. Par contre, le logiciel EyeLID ne comporte pas, dans sa version actuelle (EyeLID–2), de connaissance sur la structure statistique.

3. EXPRIMER DES QUESTIONS AVEC LE LANGAGE LID

Dans le cadre de l'analyse des données planifiées, les questions seront toujours du type : *"Quel est l'effet d'un facteur, ou d'une source de variation liée à un ou plusieurs facteurs (i.e. une interaction entre deux facteurs) sur la (ou les) variable(s) ?"* De telles questions pourront résulter d'une hypothèse de recherche précise, d'un modèle théorique qu'on cherche à mettre à l'épreuve, ou bien d'interrogations plus ouvertes.

Pour ce type de questions, la méthode privilégiée, surtout dans le contexte des données expérimentales, est l'analyse de la variance (ANOVA), présentée habituellement en mettant l'accent sur un modèle (statistique) général, et sur l'analyse inductive. L'*analyse des comparaisons* (ANACOMP) a été développée par Rouanet et Lépine (1977), puis ses extensions bayésiennes par Lecoutre (1984) comme une nouvelle perspective sur l'ANOVA

traditionnelle (voir aussi, Hoc, 1983). Dans cette approche, le modèle général est relégué au second plan, les données et les *questions spécifiques* passant au premier plan [6].

3.1. Un outil conceptuel : le langage d'interrogation des données (LID)

La formalisation entreprise par Rouanet et Lépine est intimement liée au développement parallèle du logiciel VAR3 (Lebeaux, Lépine, Rouanet, 1976) avec l'introduction de langages assurant l'interface entre logiciel et utilisateur : *langage des plans*, pour la description des structures, qui vient d'être décrit ; et *langage des demandes d'analyse*, pour la spécification des questions. Ce dernier a été le premier langage LID en date, mais l'appellation "LID" n'a été introduite qu'avec le logiciel EyeLID en 1988. Le langage LID (sous ses divers dialectes) constitue l'armature commune à l'ensemble des logiciels créés au "Groupe Mathématiques et Psychologie" : VAR3 et EyeLID déjà cités, et le logiciel PAC (Lecoutre, Poitevineau, 1992) [7] [8].

Le langage LID constitue une interface entre les logiciels et l'utilisateur, certes, mais c'est aussi, plus fondamentalement, un moyen de communication entre le statisticien et l'usager des statistiques. Le langage LID, parce qu'il s'appuie sur une formalisation, véhicule un certain nombre de concepts essentiels de la Statistique [9].

3.1.1. *Questions et langage* LID

Le langage LID permet d'exprimer chaque question spécifique, dans le cadre de la structure des données, et conduit à deux objets-clés : les comparaisons et les protocoles dérivés. Le langage LID peut être considéré, soit comme un *langage de demandes de comparaison* — c'est l'approche adoptée de façon privilégiée dans les logiciels VAR3 et PAC —, soit comme un *langage de construction et d'interrogation de protocoles dérivés* — c'est l'approche privilégiée dans EyeLID et dans ce texte.

3.2. L'entreprise "EyeLID"

Le concept de *protocole dérivé*, bien que toujours présent dans les textes sur l'ANACOMP, a été comparativement moins mis en avant que le concept de comparaison, à l'exception de (Rouanet *et al.* 1975–1976).

L'entreprise "EyeLID" a débuté en Grande-Bretagne en 1985[10], et a eu comme premier objectif d'accorder une place privilégiée aux protocoles dérivés, dans le cadre de données multivariées (Bernard, Baldy, Rouanet, 1988 ; Bernard, Rouanet, Baldy, 1993 ; Bernard

[6] Dans ce texte, l'expression ANACOMP désignera donc un ensemble de méthodes plus large que l'ANOVA usuelle, qui comprend en plus : l'approche descriptive, l'intégration de données non expérimentales, et l'approche spécifique.

[7] Mentionnons aussi d'autres logiciels qui ont jalonné cette "route linguistique" : STEL (Duquenne, 1976), PRELIM de J.-C. Bergès, et VARUNIG de M.-O. Lebeaux et H. Rouanet.

[8] Ce texte est largement consacré au logiciel EyeLID (sous sa version actuelle EyeLID–2, version 2.04), mais on s'efforcera de "parler" le langage LID général, quitte à indiquer, le cas échéant, que tel ou tel "mot" n'existe pas dans EyeLID.

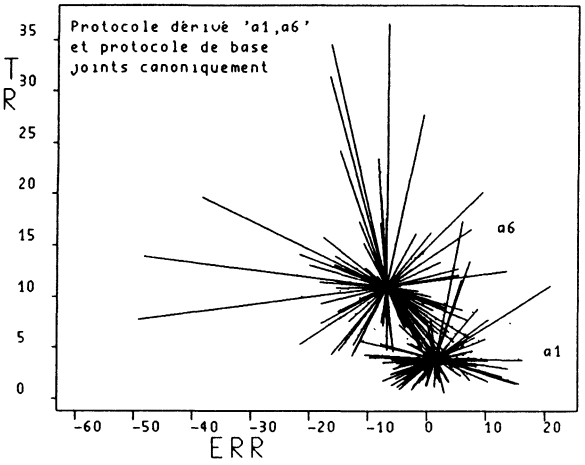
[9] C'est pourquoi il est un instrument privilégié dans beaucoup d'enseignements de notre groupe. Ce texte est d'ailleurs, en grande partie, le fruit d'un cours donné par l'auteur entre 1990 et 1993, dans le cadre du C3-C4 de Maîtrise de Psychologie Expérimentale de l'Université Paris V, cours qui s'est principalement appuyé sur le livre de J.-M. Hoc (1983).

[10] A l'occasion d'un exposé de H. Rouanet à la "Royal Statistical Society" (Rouanet, 1985), puis d'une ATP franco-britannique dirigée par B.S. Everitt et H. Rouanet (voir ATP, 1988).

et al. 1989). Mettre l'accent sur les protocoles dérivés, c'est surtout redonner une place importante à l'analyse *descriptive* des comparaisons, *i.e.* à tout ce qui est logiquement en amont de l'inférence, mais malheureusement trop souvent omis ou écourté. Si une comparaison représente un "angle" sous lequel on regarde les données, un protocole dérivé représente ce qu'on voit des données sous cet angle : *un protocole dérivé se voit.* D'où le nom "EyeLID" : on y *regarde* les données ("EYE"), en particulier graphiquement, sous l'angle de questions exprimées en langage "LID" [11].

Un exemple parlera plus que de longues phrases. Pour le dossier "Horloge", à la question "Quel est l'effet de l'Angle sur les deux variables **ERR** et **TR**?", et en se limitant aux deux angles extrêmes (**a1** et **a6**), on trouvera un premier élément de réponse dans le graphique 1 : on a superposé le protocole dérivé moyen "**a1,a6**" au protocole de base, en reliant chaque point moyen aux unités de base qui lui correspondent. On visualise ainsi, à la fois la différence moyenne entre **a1** et **a6**, mais aussi la dispersion des données pertinentes autour des points moyens. La *décomposition "inter-intra"*, au coeur de l'ANACOMP n'est pas loin !

Graph a1,a6->DEV,TR + $probase (gr.1)



3.3. Analyse descriptive et inductive

Dans ce texte, on insistera sur les aspects descriptifs de l'analyse des données planifiées (chapitres II. à V.). Une raison majeure en est que beaucoup des problèmes méthodologiques qui se posent dans l'analyse de données planifiées ne sont pas liés à l'induction ; ils existent dès le niveau descriptif : interprétation d'une interaction, plans non-équilibrés. Avec l'approche adoptée ici, on s'apercevra en fin de compte que l'étape inductive est conceptuellement plus simple qu'il n'y paraît dans sa présentation usuelle.

Ensuite, nous défendons l'idée que l'analyse inductive doit prendre l'analyse descriptive comme point de départ : en réponse à une question, l'analyse descriptive permet de dégager certaines *propriétés des données* : "*tel effet observé est important, tel autre est faible*". Le but de l'analyse inductive est alors simple à formuler : chercher à *généraliser* de telles propriétés, *e.g.* à une population plus vaste. A cet égard les *méthodes d'inférence bayésienne* constituent des méthodes privilégiées.

Dans ce texte, les analyses inductives seront abordées brièvement, et ce uniquement dans le cadre de données expérimentales univariées, avec une présentation des méthodes d'ANOVA (juger de l'existence des effets) et des méthodes bayésiennes (juger de leur importance) (cf. chapitre VI.).

[11] L'utilisation de la *couleur* dans les représentations graphiques d'EyeLID est essentielle par l'accessibilité qu'elle confère à des données nombreuses et/ou complexes. Mais le lecteur sera malheureusement condamné à seulement imaginer la couleur dans ce texte !

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 19–26)

II. PROTOCOLES PONDÉRÉS, PROTOCOLES DÉRIVÉS

RÉSUMÉ — *Un ensemble de données structurées a été décrit comme un "protocole de base" (cf. I.§2.). Une question posée aux données s'exprime par un "protocole dérivé". On introduit ici la notion de "protocole pondéré" qui recouvre ces deux notions, puis les statistiques associées qui interviennent dans l'ANOVA et l'*ANACOMP. *Enfin, on décrit les principales dérivations qui constituent les "atomes" opérationnels du langage* LID.

SUMMARY — Weighted protocols, Derived protocols
A structured data-set has been described as a "basic protocol" (cf. I.§2.). A question asked to the data can be expressed by a "derived protocol". We introduce the notion of "weighted protocol" which covers both previous ones, and define the associated statistics that are involved in ANOVA and ANACOMP. *Finaly, we describe the main derivations that constitute the operational "atoms" of the* LID *language.*

1. INTRODUCTION

Dans le chapitre I., nous avons décrit un ensemble de données structuré comme un *protocole de base*. Analyser les données, c'est "poser des questions" au protocole de base, en se référant, naturellement, à cette structure. Ces questions sont exprimées à l'aide du langage d'interrogation des données (LID). La réponse à ces questions consiste en des "protocoles dérivés" du protocole de base — c'est une généralisation de la notion de résumés statistiques —, auxquels on applique certaines procédures.

Toute la "philosophie" du logiciel EyeLID est contenue dans ce qui précède : EyeLID *manipule des protocoles*, de base ou dérivés, qui sont tous des représentants particuliers d'un unique type d'*objet*, des *protocoles pondérés*. On va ici décrire comment on passe du protocole de base aux protocoles dérivés, à la fois du point de vue du langage ("Comment ça se dit ?") et de celui des opérations concrètes ("Comment ça le fait ?").

2. PROTOCOLE DE BASE ET PROTOCOLES DÉRIVÉS

2.1. Protocole de base

Considérons à nouveau l'exemple "Négligence" (cf. I.§2.1. p. 10). Le protocole de base a pour support l'ensemble `U` de 144 unités (on parle aussi d'*unités de base*). Il est caractérisé, par l'*application-description* — le support `U` est décrit par le plan `S12<C2>*M2*03` —, et par l'*application-observation* — à chaque unité u, on associe une valeur observée de l'espace d'observables, notée x^u (cf. I.§2.2.2.). De ce dernier point de vue, le protocole de base est ainsi décrit comme un *ensemble d'observations* (u, x^u).

2.2. Protocoles dérivés

Considérons la question : "Quel est l'effet de l'orientation (facteur `O`) sur la déviation (variable `DEV`) pour la condition active (`c1`) et la main gauche (`m1`)?" Cette question amène, entre autres, à considérer le *protocole dérivé* des moyennes de la variable `DEV` selon `O` en se restreignant à `c1m1`, qu'on désigne en langage LID par : "`O/c1m1->DEV`". Il comporte 3 *unités dérivées*, une pour chacune des 3 modalités du facteur `O`. A chaque unité dérivée, est associée la moyenne de la variable `DEV` calculée sur les unités de base qui lui correspondent. Ce protocole dérivé est obtenu par plusieurs *dérivations* : restriction à `c1m1`, puis moyennage sur le facteur `S`. Une dérivation quelconque peut être ainsi décrite par une série de dérivations élémentaires que nous détaillons au §5..

2.3. Protocoles pondérés

Une notion plus générale de protocole fait intervenir, outre l'application-description et l'application-observation, une *pondération*, w_U, *i.e.* un ensemble de *poids*, w_u associés à chaque unité u. Une observation est alors un triplet : (u, x^u, w_u) [1]. La notion générale de *protocole pondéré* qui vient d'être définie recouvre à la fois celle de protocole de base et celle de protocole dérivé :

- Pour le protocole de base, la pondération est en général *élémentaire* : $\forall u, w_u = 1$.
- Pour un protocole dérivé la pondération est en général non-élémentaire. Par exemple, pour le protocole "`O/c1m1->DEV`", les valeurs sont des moyennes calculées sur 12 unités de base chacune, d'où un poids $w_u = 12$ par unité dérivée (ce mode de dérivation des poids est spécifique du moyennage, cf. §5.3.).

On notera f_u les poids relatifs : $f_u = w_u / \sum w_u$. Si les poids w_u sont tous égaux, le protocole est dit *équipondéré* ; s'ils valent tous 1, il est dit *élémentaire*.

Soulignons que le protocole de base et tout protocole dérivé sont des objets structurellement identiques : un ensemble pondéré d'unités, décrites par les modalités de facteurs, et auxquelles sont associées des valeurs de variables.

3. PROTOCOLES DÉRIVÉS ET DEMANDES DU LANGAGE LID

Le langage LID du logiciel EyeLID permet d'énoncer des *demandes d'analyse* sur un protocole de base donné. Une demande d'analyse définit deux choses : un protocole dérivé (sa structure et les opérations pour y parvenir), et une procédure à lui appliquer. Le protocole dérivé est spécifié par une *formule*, par exemple "`O/c1m1->DEV`". La procédure est spécifiée par un *mot-clé* situé *à gauche* de la formule : elle peut consister, soit à représenter le protocole (`Table` pour un tableau, `Graph` pour un graphique, *etc.*), soit à en calculer certaines statistiques (`Mean` pour sa moyenne, *etc.*.) [2].

La formule de protocole dérivé contient une liste de *variables dérivées* introduite par le symbole '`->`'. Ici, on considérera qu'il y a une seule variable dérivée, la variable initiale `DEV`, et cette liste sera souvent omise. Pour un protocole multivarié, les dérivations qui vont être introduites se feraient séparément pour chaque variable.

---

[1] On adopte ici les notations duales des *variables* et des *mesures* : quand on regroupe deux unités, les variables x^u (indice en haut), se dérivent par *moyennage*, et les *mesures* w_u (indice en bas) par *sommation* (cf. Rouanet, Le Roux, 1993, chp. II).

[2] Le logiciel EyeLID a été développé initialement en Grande-Bretagne, d'où ses mots-clés en anglais.

4. STATISTIQUES UNIVARIÉES SUR UN PROTOCOLE PONDÉRÉ[3]

Voici la représentation "brute" (mot-clé **Raw**) du protocole pondéré "**O/c1m1**" qui fournit valeur (x^u) et poids (w_u) pour chaque unité :

Raw O/c1m1 -> DEV

	x^u	w_u
o1	1.408	**12**
o2	1.238	**12**
o3	0.342	**12**

On sera constamment amené à résumer de tels protocoles en en calculant diverses statistiques numériques. Chaque statistique s'exprime par un *mot-clé* du langage LID. Le tableau 1 donne quelques-unes des statistiques univariées usuelles en précisant : mot-clé en langage LID, signification, formule de calcul, et résultat sur le protocole "**O/c1m1**". Les statistiques du tableau 1 ne constituent qu'un petit échantillon des statistiques qu'offre le langage LID[4]. D'autres statistiques seront introduites aux §5.4. (statistiques brutes) et au §6. (statistiques corrigées par les degrés de liberté). Dans ce texte, on fera principalement appel à des indices *métriques* qui sont privilégiés en ANACOMP. Parmi les statistiques du tableau 1, il faut distinguer :

- Les *indices structurels* ne dépendent pas des valeurs, mais seulement des poids ou de la structure du protocole dérivé : **Tweight** ("Total weight", noté W dans les formules), **Nbunit** ("Number of units", noté U).
- Les *indices métriques de centralité* incluent la moyenne pondérée, on dira simplement la moyenne, (**Mean**), et la moyenne équipondérée (**Emean**) ; ces deux indices coïncident ici car **O/c1m1** est équipondéré.
- Les *indices métriques de dispersion* sont : la "somme des carrés centrés" aussi appelée "inertie centrée" ou "inertie" (**Ss**, pour "Sum of squares") ; la "variance" (**Var**) ; l'"écart-type" (**Sd**, pour "Standard deviation") ; et la "différence" (**Diff**), pour un protocole n'ayant que 2 unités. Ces statistiques mesurent toutes la dispersion des valeurs du protocole autour de sa moyenne; pour un protocole constant (*i.e.* $\forall u, u',\ x^u = x^{u'}$), elles valent toutes *0.* Comme **Diff**, la statistique **Sd** s'exprime dans la même unité de mesure que la variable, des *cm*, et a une interprétation simple : elle représente un "écart moyen des valeurs à la moyenne générale" [5].

Mot-clé		Signification	Formule	O/c1m1
Tweight		Poids total	$W = \sum_u w_u$	36
Nbunit		Nombre d'unités	U	3
Mean	*	Moyenne pondérée	$M = \sum_u (w_u x^u)/W = \sum_u f_u x^u$	0.996
Emean	*	Moyenne équipondérée	$\sum_u x^u / U$	0.996
Diff	*	Différence (pour $U = 2$)	$x^1 - x^2$	——
Ss		Somme des carrés centrés	$\sum_u w_u (x^u - M)^2$	7.878
Var	*	Variance	$\mathbf{Ss}/W = \sum_u f_u (x^u - M)^2$	0.219
Sd	*	Ecart-type	$\sqrt{\mathbf{Var}} = \sqrt{\sum_u f_u (x^u - M)^2}$	0.468

Tableau 1: *Statistiques numériques univariées sur un protocole pondéré* $(x^u, w_u)_{u \in U}$.

[3]Les valeurs sont données ici avec 3 décimales, mais les calculs ultérieurs ont toujours été effectués avec une plus grande précision, et le calcul à la main n'en donnera, en général, qu'une valeur approchée (ceci sera également vrai pour les autres chapitres de ce texte).

[4]Mentionnons, pour mémoire, quelques autres statistiques descriptives du langage LID : le minimum **Min**, le maximum **Max**, l'étendue **Range**, la médiane (interpolée) **Median**, les quartiles **Qu1** et **Qu3**.

[5]Pour un protocole de 2 unités, on a : **Sd**= $\sqrt{f_1 f_2} \times$**Diff**.

5. DÉRIVATIONS ÉLÉMENTAIRES

5.1. Dérivation par restriction, réunion de protocoles

Restreindre un protocole de base, c'est construire un protocole dérivé qui ne comporte qu'une partie des unités du protocole de base, et qu'on appelle un *sous-protocole* du protocole de base. La restriction peut s'exprimer de plusieurs façons dans le langage LID ; pour désigner le protocole `O/c1m1`, on a utilisé le "slash" ('/', lire "conditionnellement à" ou "restreint à"), mais on aurait pu également adopter une écriture qui rappelle le plan du protocole : `c1*m1*O`. La restriction peut également s'exprimer à l'aide de facteurs partiels : "`S<C>*M*o1,o3`" est un sous-protocole du protocole de base, dans lequel on se restreint aux modalités `o1` et `o3`.

Lorsqu'on considère plusieurs protocoles obtenus par restriction à des modalités distinctes, on parle de *protocoles disjoints* : par exemple les sous-protocoles "`S<C>*M*o1`" et "`S<C>*M*o2`". Cette notion vaut, plus généralement, pour des protocoles dont les supports sont disjoints. On sera souvent amené à faire la *réunion* de protocoles disjoints.

5.2. Regroupement de modalités

A une certaine étape de l'analyse, on veut opposer, non plus les 3 orientations entre elles (facteur `O`), mais `o1` et `o2` d'une part à `o3` d'autre part, en se restreignant à `c1m1`. On considérera pour cela le protocole dérivé "`o1_o2,o3/c1m1`" qui comporte 2 unités dérivées d'étiquettes respectives "`o1_o2`" et "`o3`" [6]. L'opérateur '`_`' (lire "ou") conduit à un *regroupement* de modalités. Par défaut, ce regroupement se fait par moyennage.

5.3. Dérivation par concentration ou par "mot-clé à droite"

L'opération de *concentration* correspond d'abord à l'idée de résumer plusieurs valeurs par une seule valeur. Dans la construction du protocole `O/c1m1`, on moyenne sur les facteurs absents de la formule, ici `S` ; par exemple la valeur de l'unité `o1/c1m1` s'obtient par moyennage du sous-protocole `S<c1>*m1*o1`. Dans le langage LID, la dérivation par moyennage (`Mean`) est la "dérivation par défaut". Mais on peut également concentrer les valeurs par d'autres statistiques ; par exemple, la formule "`O Sd /c1m1`" indiquerait qu'on concentre en calculant des écarts-types. De même, `O/c1m1` peut aussi se dire "`O Mean /c1m1`", pour rappeler qu'on dérive par moyennage, voire encore "`O Mean S /c1m1`", pour préciser qu'on moyenne sur le facteur `S` [7]. Comme on voit, la *statistique concentratrice* est spécifiée par un *mot-clé*, par exemple "`Sd`" ; on parle de *mot-clé à droite*, parce qu'il apparaît à droite dans la formule de protocole dérivé. Dans une formule LID, il faut clairement distinguer "mot-clé à gauche", le "Quoi ?", de "mot-clé à droite", le "Comment ?" : "`Sd O`" consiste à demander le *calcul de l'écart-type* du protocole `O` (*dérivé par moyennage*), alors que "`O Sd`" désigne un protocole de support `O` *dérivé par écart-type* à partir du protocole de base.

Tous les mots-clés du tableau 1 p. 21 (ainsi que ceux indiqués en note p. 21) peuvent s'utiliser comme mots-clés à droite. La colonne "Formule" de ce tableau indique comment

[6] Considérer une telle opposition, est équivalent à considérer implicitement un nouveau facteur `O'` construit par regroupement de modalités de `O`. EyeLID fait cette opération explicitement et traduit la formule "`o1_o2,o3/c1m1`" en "`o'1,o'2/c1m1`".

[7] Les trois formules précédentes peuvent aussi s'écrire respectivement : "`O/c1m1 Sd`", "`O/c1m1 Mean`" et "`O/c1m1 Mean S`". On préfèrera en général les premières écritures parce que plus proches du fonctionnement effectif de EyeLID : le protocole de base est filtré, au préalable, par "`/c1m1`", les autres dérivations n'intervenant que sur ce protocole filtré.

calculer les valeurs dérivées $x^{u'}$ à partir des x^u. Une idée supplémentaire, dans l'opération de concentration est qu'on concentre aussi les poids w_u en un poids dérivé $w_{u'}$. A chaque statistique concentratrice, est associé un mode particulier de dérivation des poids. Le tableau 2 indique comment se dérivent les poids pour les *statistiques canoniques* (statistiques pour lesquelles le choix du poids dérivé s'impose pour des raisons théoriques abordées au chapitre V.) : **Mean**, **Emean**, **Diff**. On a déjà évoqué que, pour le moyennage (**Mean**), le poids dérivé est la somme des poids des unités moyennées; si le moyennage porte sur un protocole élémentaire, le poids est donc simplement un effectif. Pour **Emean** et **Diff**, par contre, on notera que les poids dérivés peuvent ne pas être entiers.

Statistique	Formule de dérivation des poids
Mean	$w_{u'} = \sum_u w_u$
Emean	$w_{u'} = \frac{U^2}{\sum_u 1/w_u}$
Diff	$w_{u'} = \frac{w_1 w_2}{w_1 + w_2}$

Tableau 2: *Dérivation des poids pour les statistiques canoniques* **Mean**, **Emean**, **Diff**.

Deux cas extrêmes d'utilisation de l'opération de concentration doivent être soulignés (dans les deux cas, on considère un protocole de base de support **U**) :

- La formule "**U**" désigne un protocole dérivé de **U** unités dérivées par moyennage. Mais chaque moyennage porte ici sur une seule unité de base (puisque **U** est un plan) ; et la moyenne d'une seule valeur est cette valeur elle-même. La formule "**U**" réalise ainsi une concentration triviale et redonne simplement le protocole de base [8].
- La formule "**Z**" désigne un protocole dérivé d'une seule unité **z1** (puisque **Z** est constant) dont la valeur est la moyenne générale du protocole de base. La formule "**Z**" réalise une concentration maximum, de toutes les unités de base en une unité dérivée [9].

5.3.1. *Dérivation par moyennage pondéré et équipondéré*

La dérivation par moyennage pondéré (**Mean**) est privilégiée dans le langage LID ; comme on l'a dit, c'est la dérivation par défaut pour un regroupement. Une notion plus générale de moyenne est celle d'une moyenne pondérée par des coefficients, les poids. Se pose alors un choix crucial quand on regroupe plusieurs unités dont les poids sont différents : doit-on calculer la moyenne en en tenant compte (**Mean**) ou non (**Emean**) ? Il est impossible de donner une réponse générale à cette question. Dans le contexte de l'analyse des données planifiées, les poids proviennent toujours d'effectifs différents, par exemple d'un emboîtement non-équilibré. Si la disparité de ces effectifs provient de choix (ou de contraintes) de l'expérimentateur, on préfèrera souvent ne pas en tenir compte, et on privilégiera le moyennage équipondéré [10].

[8] La formule "**U**" fournit les données de base, mais celles-ci sont uniquement décrites par le facteur **U**. Si la formule est "**S<C>*M*O**", *i.e.* le plan le plus riche, alors le protocole dérivé *est* le protocole de base (décrit par tous ses facteurs).

[9] On distingue "**Mean U**" qui désigne seulement la *valeur* de la moyenne générale, de "**Z**" ou "**Z Mean U**" qui désigne un *protocole* : une unité **z1**, à laquelle sont associés cette *valeur* et son *poids*.

[10] Pour avoir un aperçu de la difficulté du problème, encore accrue pour les données d'observation, on pensera à l'indice des prix, calculé à partir d'un "panier de la ménagère" (un ensemble d'unités), pondéré par des coefficients sélectionnés "avec soin". Mais souvent le taux d'inflation ainsi calculé varie entre le gouvernement (choix n°1 des coefficients) et tel syndicat (choix n°2). Qui a "raison" ?

5.3.2. *Dérivation par différence, par contraste*

Pour un protocole de support A comportant 2 unités, l'effet du facteur A s'évalue par une différence : "Diff a1,a2" ou "Diff a2,a1", selon ce qui est le plus signifiant. Pour un facteur A à plus de 2 modalités, on pourra de même évaluer des effets partiels : "Diff a1,a3" (a1 contre a3) ou "Diff a1,a2_a3" (a1 contre a2 et a3 regroupés).

Lorsqu'un facteur B est croisé avec un facteur (ou facteur partiel) A2, l'étude de l'effet de A pour chaque modalité de B fera intervenir le protocole *dérivé par différence*, "B Diff a1,a2", de support B, dont les valeurs sont les "Diff a1,a2/b" et les poids dérivés conformément au tableau 2.

t La notion de différence se généralise en celle de *contraste*, puis en celle de *comparaison* (cf. chapitre IV.). C'est pour assurer la cohérence entre le point de vue des protocoles dérivés et celui des comparaisons que la dérivation par différence s'accompagne de la dérivation des poids donnée au tableau 2 p. 23.

5.4. Dérivation par centrage ; statistiques brutes

Centrer un protocole, c'est construire un nouveau protocole de même structure, où chaque valeur x^u est remplacée par l'écart de cette valeur à la *moyenne pondérée* du protocole : $x^{u'} = x^u - M$. Le poids des unités reste inchangé. Si le protocole a pour support U, le protocole centré associé se désigne par "U(Z)".

Certaines dérivations du langage LID font intervenir plusieurs centrages. Pour un protocole de support U décrit par le facteur A, le protocole *dérivé intra-A* (ou *centré selon* A), noté U(A), se construit en centrant séparément chaque sous-protocole U/a, soit par A centrages indépendants. De même, une *dérivation d'interaction* entre deux facteurs croisés A et B peut être décrite par un double-centrage, à la fois selon A et selon B. On reviendra sur cela au chapitre III.

Les statistiques de dispersion du tableau 1 p. 21 (Ss, Var, Sd) peuvent toutes être qualifiées de "centrées", dans la mesure où chacune ne se calcule en fait que sur le protocole centré associé au protocole considéré : leurs formules ne font intervenir que les écarts $(x^u - M)$. On peut définir des analogues "brutes", *i.e.* non-centrées, de ces statistiques en remplaçant les écarts $(x^u - M)$ par la valeur brute x^u : d'où les statistiques Rss, Rvar, Rsd ('R' pour "Raw") données dans le tableau 3 [11].

Mot-clé	Signification	Formule	O/c1m1
Rss	Somme des carrés bruts	$\sum_u w_u(x^u)^2$	43.579
Rvar *	Variance brute	$\texttt{Rss}/W = \sum_u f_u(x^u)^2$	1.211
Rsd *	Ecart-type brut	$\sqrt{\texttt{Rvar}}$	1.100

Tableau 3: *Statistiques numériques univariées brutes, sur un protocole pondéré* $(x^u, w_u)_{u \in U}$.

Avec l'introduction de ces nouveaux mots-clés, on peut considérer que l'opération de centrage a lieu, ou bien au niveau du calcul d'une statistique, ou bien au niveau du protocole lui-même. Dans la suite de ce texte, on adoptera systématiquement le second point de vue : on préfèrera utiliser "Rss O(Z)/c1m1" — somme des carrés bruts du protocole centré O(Z)/c1m1 associé à O/c1m1 —, plutôt que "Ss O/c1m1" — somme des carrés centrée du protocole non-centré O/c1m1.

[11] On notera que, du fait de sa définition, la Rss de la réunion de deux protocoles disjoints, s'obtient par addition de leurs Rss respectives ; cette propriété sera constamment utilisée au chapitre III.

Graphiquement, centrer un protocole, c'est simplement le déplacer de façon à amener son point moyen à *0*, opération qui n'affecte ni sa variance, ni les écarts entre valeurs. Or, dans l'ANACOMP, on ne considère souvent d'un protocole que sa variance ou de tels écarts, *i.e.*, en fin de compte, que le protocole centré associé. D'où le fait que *l'explicitation des opérations de centrage est une clé essentielle pour comprendre l'*ANACOMP. En particulier le calcul des degrés de liberté — qui apparaît souvent un peu "magique" — devient alors limpide (cf. §6.). Ainsi, cette explicitation, qui n'est pas usuelle et peut sembler introduire une lourdeur inutile dans les notations, est, au contraire, porteuse de simplification.

6. NOMBRE DE DEGRÉS DE LIBERTÉ ASSOCIÉ A UN PROTOCOLE

Considérons un protocole de support `U`, comprenant U unités [12], protocole de base ou dérivé. Sur le plan de la structure, chaque unité représente une "case" dans laquelle on doit ranger la valeur observée correspondante. Admettons que ne sont connus que cette structure et ces dérivations, et que les valeurs observées ne sont pas disponibles (comme par exemple avant le recueil des données), et que, de plus, on se donne la "liberté" de remplir les cases à notre gré. Le *nombre de degrés de liberté brut* (*ddl brut*, ou simplement *ddl* en abrégé, et `Rdf` en LID) associé au protocole est égal au nombre de cases qu'on peut ainsi remplir librement. Soulignons que le nombre de ddl ne dépend ni des valeurs ni des poids : il s'agit d'une propriété structurelle d'un protocole.

- Si le protocole n'est pas centré, le nombre de ddl est égal au nombre de cases : `Rdf`$= U$. En conséquence les `Rdf` s'ajoutent par réunion de protocoles disjoints.
- Si le protocole est centré (dérivation `(Z)`), sa moyenne vaut 0. On peut remplir toutes les cases librement sauf la dernière qui doit être calculée de façon à ce que la moyenne soit effectivement 0. Le centrage a fait perdre 1 ddl d'où : `Rdf`$= U - 1$.
- Plus généralement, toute opération de centrage introduit une contrainte entre les valeurs — une certaine moyenne doit valoir *0* — d'où la perte d'un ddl. On perd ainsi autant de ddl que d'opérations de centrage indépendantes. Si un protocole de support `U<A>` est centré selon `A`, alors son nombre de ddl vaut `Rdf`$= U - A$ [13].

Le nombre de ddl centré (`Df`) est le nombre de ddl brut du protocole centré associé : si le protocole est déjà centré, on a `Df=Rdf`, s'il est non-centré, on a `Df=Rdf`-1.

Dans le contexte de l'ANOVA, plusieurs indices de dispersion, déduits de ceux considérés au §4., et faisant intervenir les ddl, sont couramment utilisés : le "carré-moyen" (`Ms` pour "Mean square"), la "variance corrigée" (`Varcor`) et l'"écart-type corrigé" (`Sdcor`). Ces trois indices sont centrés, et on peut, à nouveau, en définir une version "brute" correspondante. Ces divers indices sont donnés dans le tableau 4.

Dans le "tableau d'analyse de la variance" usuel, on trouve systématiquement les statistiques `Ss`, `Df` et `Ms`. Mais à nouveau, dans ce texte, nous préférerons recourir à leurs équivalents bruts — `Rss`, `Rdf` et `Rms` — en explicitant les centrages au niveau des formules de protocoles dérivés.

La formule de la variance corrigée du tableau 4 n'est pas classique et nous a été suggérée par H. Rouanet. Elle redonne, pour le cas d'un protocole élémentaire non-centré, la formule usuelle, `Varcor` $=$ `Ss`$/(U-1)$, mais est plus générale car applicable à un

[12] On notera U, A, B, *etc.*, le nombre de modalités d'un facteur `U`, `A`, `B`, *etc.*.

[13] De même, le décompte des centrages indépendants pour un protocole d'interaction, "`A.B`", entre deux facteurs croisés `A` et `B`, conduit à la formule classique : `Rdf A.B`$= (A-1)(B-1)$ (cf. III.§4.3.3.).

Mot-clé		Signification	Formule	O/c1m1
Rdf		Degrés de liberté brut		3
Df		Degrés de liberté centré		2
Ms		Carré moyen centré	Ss/Df	3.939
Varcor	*	Variance corrigée	$(\mathtt{Ss} \times U)/(\mathtt{Df} \times W)$	0.328
Sdcor	*	Ecart-type corrigé	$\sqrt{\mathtt{Varcor}}$	0.573
Rms		Carré moyen brut	Rss/Rdf	14.526
Rvarcor	*	Variance corrigée brute	$(\mathtt{Rss} \times U)/(\mathtt{Rdf} \times W)$	1.211
Rsdcor	*	Ecart-type corrigé brut	$\sqrt{\mathtt{Rvarcor}}$	1.100

Tableau 4: *Statistiques numériques univariées, faisant intervenir les ddl.*

protocole quelconque. L'indice Rsdcor sera utilisé comme indice privilégié de la grandeur des effets (cf. chapitre III.) [14].

7. STATISTIQUES DESCRIPTIVES ET AUTRES STATISTIQUES

Parmi les statistiques définies dans ce chapitre, il faut distinguer les *statistiques purement descriptives*, en bref *statistiques descriptives* (Rouanet, Le Roux, Bert, 1987 ; Rouanet, Bernard, Le Roux, 1990, chp. I). La notion de statistique descriptive est intuitivement simple : une statistique est descriptive si elle ne dépend pas du nombre de "sujets", *i.e.* du nombre de modalités du facteur de groupe S (on suppose ici qu'il y en a un seul). La notion de statistique descriptive est ainsi relative à la structure statistique du protocole de base. En conséquence, "être ou ne pas être descriptive" n'est pas une propriété intrinsèque d'une statistique, et peut dépendre du protocole dérivé considéré (Son support comporte-t-il S ou non ?) [15].

Dans l'ANACOMP, la pondération d'un protocole résulte de regroupements d'unités, et provient donc des effectifs, *i.e.*, entres autres, du nombre de "sujets". De ceci, il découle qu'une statistique qui ne dépend que des x^u et/ou des f_u (et non de W) est nécessairement descriptive *pour tout protocole dérivé*. Une telle statistique, *intrinsèquement descriptive*, est dite "descriptive". Les statistiques descriptives sont indiquées par un '*' dans les tableaux (1 p. 21, 3 p. 24 et 4 p. 26) [16]. A l'opposé, les statistiques Tweight, Rss et Ss ne sont jamais descriptives. Enfin, le caractère descriptif des autres statistiques du tableau 4 p. 26 n'est pas intrinsèque, car les degrés de liberté peuvent provenir ou non du nombre de "sujets".

L'étape descriptive, on l'imagine, consiste surtout à calculer des statistiques descriptives. L'étape inductive fait intervenir, de surcroît, des éléments inductifs (des nombres de sujets) qui expriment le potentiel inductif des données. La dissociation de ces deux composantes est essentielle pour distinguer la *teneur des résultats* — "Que disent les données ?" —, de leur *portée* — "Avec quel degré de généralisabilité le disent-elles ?".

---

[14] Les statistiques Varcor, Rvarcor, Sdcor et Rsdcor sont, de prime abord, moins intuitives que leurs homologues non-corrigés, mais on peut entrevoir leur motivation ainsi : l'"écart-type brut" mesure un *écart à 0 moyen* des U valeurs d'un protocole ; si celui-ci a Rdf ddl, les valeurs sont liées entre elles par (U − Rdf) contraintes, qui diminuent leur "capacité" globale à s'écarter de *0* ; d'où l'idée de "corriger" les indices de dispersion par (U/Rdf) pour "dépénaliser" les protocoles contraints structurellement.

[15] Pour vérifier si telle statistique *sur tel protocole dérivé* est ou non descriptive, il suffit de la calculer, d'une part à partir du protocole de base initial, et d'autre part à partir d'un protocole de base obtenu en dupliquant l'ensemble des "sujets" : si les deux résultats coïncident, la statistique *pour ce protocole dérivé particulier* est descriptive.

[16] Les statistiques Varcor, Rvarcor, Sdcor, Rsdcor sont aussi considérées comme descriptives, mais ne sont indépendantes de W que de façon approchée ; on les qualifie parfois de "quasi-descriptives".

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 27-54)

III. PROTOCOLES DÉRIVÉS POUR CERTAINES STRUCTURES REMARQUABLES

RÉSUMÉ — *On présente ici l'analyse descriptive des données planifiées en considérant successivement des structures ensemblistes de complexité croissante :* `A`, `A<B>`, `A*B` *et* `A<B>*C`, *avec comme cas particuliers les structures statistiques remarquables :* `S`, `S<G>`, `S*T` *et* `S<G>*T`. *On aborde à la fois la décomposition standard, vue en termes de protocoles dérivés, et l'analyse de questions spécifiques.*

SUMMARY — Derived protocols for some typical structures
We present here the descriptive analysis of planned data, by successively considering several set-theoretic structures of growing complexity : `A`, `A<B>`, `A*B` *and* `A<B>*C`, *with, as particular cases, the statistical structures :* `S`, `S<G>`, `S*T` *and* `S<G>*T`. *Both the standard decomposition, in terms of derived protocols, and the analysis of specific questions are envisaged.*

1. INTRODUCTION

1.1. De l'analyse de la variance traditionnelle (ANOVA) ...

En mettant, d'abord, de côté l'aspect inductif, on peut résumer ce que fait l'analyse de la variance traditionnelle (ANOVA) en quelques mots : décomposer la *variance des données* — qui traduit l'existence d'un effet conjoint des facteurs du plan —, en un certain nombre de *sources de variation* additives, traduisant chacune l'effet d'un facteur, ou un effet lié à plusieurs facteurs (*i.e.* une interaction). La structure ensembliste des données, exprimée par le plan du protocole, détermine une *décomposition standard* des sources de variation. La spécification de leur structure statistique sépare, ensuite, ces sources en *sources systématiques*, sur lesquelles on peut faire de l'inférence, et *sources adjointes*, qui servent de termes de référence aux premières. L'ensemble des analyses est résumé dans le *tableau d'analyse de la variance* qui donne, pour chaque source de variation, diverses statistiques dont : somme des carrés, degrés de liberté, carré-moyen, et test F.

1.2. ... à l'analyse des comparaisons (ANACOMP) et des protocoles dérivés

Les travaux de H. Rouanet & D. Lépine, amorcés au début des années 70, constituent une reconstruction de l'ANOVA, et ont débouché sur une nouvelle théorie autonome par rapport à l'ANOVA traditionnelle : l'"*analyse des comparaisons* (ANACOMP)" (voir *e.g.* Rouanet, Lépine, 1976 et 1977 ; et l'avant-propos de Rouanet dans Lecoutre, 1984). Rappelons certaines idées-forces de cette nouvelle approche, en mettant ici de côté ses aspects inductifs (abordés au chapitre VI.), pour expliciter en quoi ce texte s'y rattache et en prolonge certains aspects.

Dans l'ANACOMP, les sources de variation correspondent à des *questions* posées aux données. Mettre l'accent sur les questions, c'est redonner au chercheur la liberté d'interroger ses données d'une façon adaptée à ses objectifs. Pour les structures ensemblistes les plus simples, les questions possibles sont dictées par la structure et on retrouve la décomposition standard. Par contre, plus la structure est complexe, plus la panoplie des questions s'élargit ; la décomposition standard n'est plus alors qu'un "cadre de référence" et ne contient pas nécessairement les questions liées aux objectifs de la recherche. On y ajoutera — voire on y substituera — un ensemble de *questions spécifiques planifiées*. C'est la première idée-force de l'ANACOMP : compléter l'analyse "automatique" de la décomposition standard, par *une analyse guidée par les objectifs*. Le langage LID permet d'exprimer ces questions spécifiques.

Poser une question, c'est "regarder les données sous un certain angle". Il y a ainsi deux aspects complémentaires dans la notion de question : l'angle sous lequel on regarde, et ce qu'on voit sous cet angle. Le premier aspect se formalise par la notion de *comparaison* ; le second, par celle de *protocole dérivé*. Les deux aspects sont indissociables pour avoir une compréhension globale de l'ANACOMP, mais, dans ce chapitre, on choisit de faire jouer aux protocoles dérivés un rôle primordial ; le point de vue des comparaisons sera abordé au chapitre V. Prendre les protocoles dérivés comme objet d'étude privilégié, c'est surtout — deuxième idée-force — *redonner une place importante aux aspects descriptifs*, souvent négligés dans l'ANOVA traditionnelle. Le logiciel EyeLID a été conçu avec cette vision descriptive : pour lui, tout est protocole dérivé ; et tout protocole dérivé se regarde [1].

1.3. Structure ensembliste et structure statistique des données

A un premier niveau, l'analyse descriptive ne fait intervenir que la structure ensembliste des données. On considérera ici plusieurs structures ensemblistes remarquables de complexité croissante : un facteur `A` (§2.), un emboîtement `A<B>` (§3.), un croisement `A*B` (§4.), une structure-mixte `A<B>*C` (§5.), et enfin le cas de protocoles plus complexes (§6.). Ceci donnera l'occasion d'introduire progressivement les "briques" élémentaires à l'aide desquelles on peut décomposer toute structure complexe. Les plans considérés ici seront tous équilibrés ; le cas déséquilibré et ses répercussions sont étudiées au chapitre IV.

Pour chacune de ces structures ensemblistes, on envisagera la situation où un des facteurs est un facteur de groupe (qu'on désigne par "`S`"), d'où l'étude des structures statistiques remarquables : `S`, `S<G>`, `S*T`, et `S<G>*T`. Au niveau descriptif, la répercussion principale en sera l'introduction d'*effets-calibrés* comme indicateurs privilégiés de l'*importance des effets*. Mais cela jettera également les bases de l'analyse inductive envisagée au chapitre VI.

1.4. L'analyse spécifique ; protocole de base et protocole initial

L'idée d'une analyse guidée par les objectifs conduit à adopter la démarche de l'*analyse spécifique* : pour une question d'intérêt, ne considérer du protocole de base qu'un protocole dérivé pertinent sur lequel s'effectue l'analyse : ce protocole dérivé pertinent a ainsi le statut d'un nouveau protocole de base, mais on l'appellera "*protocole initial*" pour le distinguer du protocole de base primitif. L'explicitation de cette démarche, utilisée de façon intuitive dans les premières sections, sera abordée au §6.2..

[1] Ce chapitre constitue en partie une introduction par l'exemple au langage LID du logiciel EyeLID. On se reportera au chapitre VII. pour une présentation plus systématique du langage LID, et des commandes graphiques de EyeLID.

2. ANALYSE DE L'EFFET D'UN FACTEUR : STRUCTURE "A"

On considère ici la structure la plus simple d'un protocole décrit par un seul facteur A et les protocoles qu'on peut en dériver. Cette section est également l'occasion d'introduire des considérations générales sur la notion d'effet et celles qui y sont liées : *représentant*, *grandeur* et *importance* d'un effet.

2.1. Facteur à 2 modalités

Reprenons le dossier "Négligence" (cf. I.§2.1.), de structure "S12<C2>*M2*O3->DEV", et considérons le protocole dérivé "C/m1o1" des moyennes des conditions (C) pour la main gauche (m1) et l'orientation à gauche (o1). On le considère ici comme un protocole initial, en ne retenant que sa structure C2->DEV. Ce protocole est représenté ci-après de façon "brute", *i.e.* un tableau de valeurs (x^u) pondérées (w_u) (mot-clé "Raw" en LID) :

Raw C -> DEV

	x^u	w_u
c1	1.408	12
c2	0.292	12

2.1.1. *Protocoles dérivés*

De ce protocole initial, on peut dériver plusieurs protocoles : les deux sous-protocoles c1 et c2 qui correspondent chacun à une ligne du tableau du protocole initial ; le protocole moyen Z ; le protocole centré C(Z) des écarts à la moyenne ; et le protocole de la différence "Z Diff C" puisque le facteur C a 2 modalités :

Raw Z -> DEV

	x^u	w_u
z1	0.850	24

Raw C(Z) -> DEV

	x^u	w_u
c1	0.558	12
c2	-0.558	12

Raw Z Diff C -> DEV

	x^u	w_u
z1	1.117	6

Le protocole Z représente le niveau moyen de performance ; pour être plus explicite, on pourrait aussi le désigner par "Z Mean C". Le protocole "Z Diff C" est dérivé du protocole initial par *différence* sur C, d'où son poids canonique : $6 = (12 \times 12)/(12 + 12)$ (cf. II.§2) [2]. Il représente l'effet du facteur C : sa valeur, $1.117cm$, est l'écart entre c1 et c2. Mais le protocole C(Z) traduit lui aussi l'effet du facteur C : il indique que c1 est situé à $0.558cm$ au dessus de la moyenne, et que c2 est situé à $0.558cm$ en dessous.

Le tableau ci-après donne les statistiques Rss et Rdf associées à chacun de ces protocoles (cf. II.§4.). On trouve par exemple : Rss C $= 12 \times 1.408^2 + 12 \times 0.292^2 = 24.822$. La statistique Rdf est simplement le nombre d'unités pour les protocoles non-centrés (C, c1, c2, Z et "Z Diff C") ; mais vaut seulement 1 pour C(Z) : 2 unités moins 1 centrage (cf. II.§6.). Ces résultats nécessitent plusieurs remarques :

- Le protocole C est la réunion des deux protocoles disjoints c1 et c2, d'où la décomposition additive de C en "c1+c2" pour les Rss et les Rdf.
- On constate également que le protocole C se décompose additivement, en termes de Rss et de Rdf, en "Z+C(Z)".
- Les deux protocoles C(Z) et "Z Diff C" ont mêmes statistiques Rss et Rdf.

[2] On rappelle que, dans une formule du type "Z Mean C" ou "Z Diff C", 'Z' désigne le support du protocole dérivé (ici d'une seule unité z1), et ce qui est à droite, *e.g.* "Mean C", désigne le mode de dérivation des valeurs et des poids pour chaque unité dérivée (cf. II.§5.3.).

Protocole	Rss	Rdf
C	24.822	2
c1	23.801	1
c2	1.021	1

Protocole	Rss	Rdf
Z	17.340	1
C(Z)	7.482	1
Z Diff C	7.482	1

2.1.2. *Effet d'un facteur à 2 modalités*

Lorsqu'un facteur a deux modalités, il est naturel de représenter son *effet* par, au choix, une des différences "Diff c1,c2" ou "Diff c2,c1", plutôt que par C(Z) : la différence choisie constitue alors un *représentant* privilégié de l'effet Dans cet effet, il faut distinguer soigneusement : la *valeur* de l'effet "Diff c1,c2", dont le signe est le *sens* de l'effet ; et la *grandeur* de l'effet : cette dernière pourra être évaluée à l'aide de divers indices, comme, par exemple ici, la valeur absolue de la différence. Cette distinction deviendra essentielle pour un facteur à plus de 2 modalités.

Mais comme on va voir, la notion d'effet ne se réduit pas à calculer une simple différence. Ce chapitre fera intervenir de nombreux autres types d'"effets" : un effet sera toujours représenté par un (ou plusieurs) protocole(s) dérivé(s). Parmi les indices qu'on peut associer à un "effet" (dans ce sens large), le Rdf et la Rss sont des indices privilégiés : le premier en tant que propriété structurelle de l'effet (sa dimensionnalité), la seconde en tant qu'indicateur non-descriptif de la grandeur de l'effet. Voici les deux propriétés fondamentales de ces indices (elles renvoient à la formalisation de la notion de *comparaison* qui sera abordée au chapitre V.) :

- Si deux effets correspondent à des facettes des données mutuellement "indépendantes", alors les Rss et les Rdf s'ajoutent : *e.g.* ci-dessus Z et C(Z) ; cette indépendance des effets découle de l'*orthogonalité* des comparaisons associées ;
- Si deux protocoles dérivés traduisent le même effet, alors leurs Rss et leurs Rdf sont respectivement égaux : *e.g.* ci-dessus C(Z) et "Z Diff c1,c2". Dans ce dernier cas, on parlera de "*protocoles équivalents (pour l'effet)*".

2.2. Facteur à plus de 2 modalités

Reprenons l'exemple du protocole "O/c1m1->DEV", dérivé du dossier "Négligence" (déjà utilisé en II.§2.2. p. 20), considéré comme un protocole initial de structure "O3->DEV". La majeure partie de ce que nous venons de voir se généralise ici : à partir du protocole initial O, on définit les sous-protocoles o1, o2 et o3, le protocole moyen Z, et le protocole centré O(Z). Ces protocoles sont donnés ci-après, ainsi que les Rss et Rdf associés. On retrouve les deux décompositions (pour les Rss et les Rdf) : O=o1+o2+o3, et O=Z+O(Z).

Raw O -> DEV

	x^u	w_u
o1	1.408	12
o2	1.238	12
o3	0.342	12

Raw Z -> DEV

	x^u	w_u
z1	0.996	36

Raw O(Z) -> DEV

	x^u	w_u
o1	0.413	12
o2	0.242	12
o3	-0.654	12

Protocole	Rss	Rdf
O	43.579	3
o1	23.801	1
o2	18.377	1
o3	1.401	1
Z	35.701	1
O(Z)	7.878	2

Le protocole centré O(Z) traduit encore l'effet du facteur O, mais puisque O a 3 modalités, cet effet ne peut plus être représenté par une seule différence, il est maintenant

"*multidimensionnel*" [3]. Caractériser l'effet d'un facteur comporte trois aspects que nous détaillons ci-après : *représenter l'effet*, *calculer un indicateur global de grandeur de l'effet*, et *mesurer l'importance de l'effet*.

2.2.1. *Notion d'effet, représentant de l'effet*

Dire que le facteur `O` a un effet, exprime que les valeurs pour chaque modalité de `O` ne sont pas toutes identiques. Ainsi, l'existence d'un effet se traduit par le fait que certaines valeurs s'écartent de la moyenne générale, donc que le protocole centré `O(Z)` n'est pas constamment nul. Mais, de façon équivalente, elle se traduit aussi par l'existence de différences entre certaines valeurs. A l'opposé, il y aurait absence d'effet si le protocole centré `O(Z)` était constamment nul ; ou, encore, si les différences entre modalités prises deux-à-deux étaient toutes nulles.

Ainsi pour représenter l'effet de `O`, on peut choisir entre deux visions : soit considérer le protocole centré `O(Z)`, soit considérer un ensemble de différences entre les valeurs. La première vision fait jouer des rôles symétriques aux modalités de `O` ; ce n'est pas forcément le cas pour la seconde, tout dépend du choix des oppositions entre modalités. Nous détaillons cette seconde approche ci-après.

On peut d'abord choisir de représenter l'effet de `O` par l'ensemble de *toutes* les différences entre couples de modalités : `Diff o1,o2`= 0.171, `Diff o2,o3`= 0.896, et `Diff o1,o3`= 1.067. Ces différences décrivent complètement l'effet de `O`, mais avec une certaine redondance, puisqu'on a : $0.171 + 0.896 = 1.067$.

Un autre possibilité consiste à chercher à décomposer l'effet de `O` en plusieurs effets partiels unidimensionnels *non redondants*. L'effet global de `O`, représenté par `O(Z)`, a $(n-1) = 2$ ddl, et peut, en conséquence, être décomposé additivement à l'aide de seulement 2 effets unidimensionnels, *i.e.* par 2 différences *convenablement choisies*. Ce type de décomposition n'est pas unique ; un exemple d'une telle décomposition est : "`Z Diff o1,o2`" et "`Z Diff o1_o2,o3`". L'effet global de `O` est décrit à l'aide de 2 oppositions : `o1` contre `o2` d'une part, et le regroupement "`o1_o2`" contre `o3` d'autre part. Le tableau ci-après donne les deux protocoles dérivés correspondant (d'une unité chacun) et les statistiques **Rss** et **Rdf** associées, sur lesquelles on vérifie l'additivité de la décomposition [4] :

Protocole	x^u	w_u	**Rss**	**Rdf**
`O(Z)`			7.878	2
`Z Diff o1,o2`	0.171	**6**	0.175	1
`Z Diff o1_o2,o3`	0.981	**8**	7.703	1

t Ce qui vient d'être fait consiste, on le verra au chapitre V, à décomposer la *comparaison globale* sur le facteur `O` à $(n-1)$ ddl en $(n-1)$ *contrastes orthogonaux*. L'additivité découle de l'orthogonalité des oppositions choisies. Pour un facteur à n modalités, l'ensemble des $(n-1)$ oppositions ("`o1,o2`", "`o1_o2,o3`", "`o1_o2_o3,o4`", *etc.*) fournit toujours une décomposition orthogonale.

[3] La multidimensionnalité tient ici à ce qu'on *compare* globalement plus de 2 unités, et donc que l'effet a plus d'un ddl. Lorsque le protocole est multivarié, l'effet est également "multidimensionnel", mais dans un autre sens : on parlera alors plutôt d'un effet "multivarié".

[4] Au lieu des deux oppositions précédentes, on pourrait considérer "`o1,o3`" et "`o1_o3,o2`" qui redonneraient, par addition, la même **Rss** globale. Lorsque l'effet est multidimensionnel, plusieurs décompositions additives sont possibles, dont certaines seulement seront privilégiées du fait des questions qu'on se pose. D'une représentation de l'effet à une autre, les statistiques **Rss** et **Rdf** sont invariantes.

2.2.2. *Indicateurs de la grandeur d'un effet*

Lorsque l'effet est multidimensionnel, on ne peut plus le caractériser par *une valeur* et un sens. Par contre, on peut encore définir un indice numérique global de *la grandeur de l'effet* : on choisira pour cela un indice descriptif, qui s'exprime dans la même unité de mesure que celle de la variable (ici des *cm*). Divers indices peuvent être envisagés, découlant chacun d'une choix particulier du représentant de l'effet.

En considèrant `O(Z)` comme représentant de l'effet, on est conduit au calcul d'un indice de dispersion entre les valeurs de `O(Z)`. On pourrait ainsi envisager n'importe laquelle des statistiques descriptives `Rvar`, `Rsd`, `Rvarcor` et `Rsdcor`, chacune valant *0* pour un protocole constamment nul. Dans ce texte, on utilisera systématiquement l'indice `Rsdcor` comme indicateur de la grandeur de l'effet ; pour un protocole centré, il s'interprète comme un "écart moyen des valeurs à la moyenne" (corrigé par les ddl) ; ici il vaut : `Rsdcor O(Z)`$= 0.573cm$.

Choisir de représenter l'effet par un ensemble de différences conduit à définir un indice de grandeur comme une "moyenne" de différences. Un indice de ce type fréquemment utilisé est le *diamètre* (mot-clé `Diam`) [5] ; il est défini comme la moyenne quadratique des différences $(x^u - x^{u'})_{u<u'}$ pondérée par les coefficients $(w_u w_{u'})$, mais s'exprime aussi en fonction de la variance de `O` :

$$\texttt{Diam}^2 \;=\; \frac{\sum_{u<u'} w_u w_{u'} (x^u - x^{u'})^2}{\sum_{u<u'} w_u w_{u'}} = \texttt{Var} \times \frac{2W^2}{W^2 - \sum_u w_u^2}, \quad \textit{avec} \;\; W = \sum_u w_u. \qquad (1)$$

Les deux façons de représenter l'effet conduisent toutes deux à définir la grandeur de l'effet de `O` à l'aide d'indices qui sont liés à la variance de `O`. C'est une des idées clés à la base de l'analyse de la variance : *un effet se traduit par de la variance.*

Pour le cas de l'effet d'un facteur équipondéré, les divers indicateurs envisageables sont liés entre eux par la relation : `Rsdcor` $= \texttt{Rsd} \times \sqrt{U/(U-1)} = \texttt{Diam}/\sqrt{2}$ (chacun étant évalué sur "`O(Z)`)". L'indice `Diam`, qui généralise `|Diff|` pour un facteur à deux modalités, est plus naturel. Mais on lui préfèrera l'indice `Rsdcor` parce que plus général : d'une part, il est défini pour n'importe quel protocole (même ceux ne comportant qu'une unité), d'autre part le calcul d'effets-calibrés se fera toujours par un rapport de deux `Rsdcor` (cf. §2.5.) [6].

2.2.3. *Importance d'un effet*

Les notions précédentes (représentant, valeur pour un effet à un ddl, grandeur de l'effet) ne font intervenir que le protocole dérivé considéré. Une autre notion, l'*importance de l'effet* fait intervenir, de surcroît, une valeur de référence extérieure à celui-ci. On jugera de l'importance de l'effet en *rapportant* la grandeur de l'effet à cette valeur de référence, d'où un indice *sans unité* (Corroyer, Rouanet, 1994). Il y a plusieurs approches, non exclusives, pour caractériser l'importance de l'effet :

- La valeur de référence peut être externe au protocole de base ; elle sera choisie *a priori* compte-tenu du contexte expérimental, de la "sémantique du domaine" (Reuchlin, 1992) ; d'où une certaine part de subjectivité dans son choix.

[5] Il est souvent noté 'ℓ', *e.g.* dans Lecoutre (1984 et 1991).

[6] Par définition, deux protocoles dérivés équivalents (pour l'effet), ont même `Rss` et `Rdf`. Par contre, ils n'ont pas forcément toujours même `Rsdcor`. Ce qui est essentiel est que les effets-calibrés, définis comme rapports de deux `Rsdcor`, seront eux invariants.

- Il peut s'agir également d'une référence interne au protocole de base : l'étendue des valeurs, une moyenne particulière qui représente un "niveau *0*" de performance, une différence "étalon", *etc.*.
- A partir d'indices d'importance des deux types précédents il est impossible, cependant, de donner des critères généraux pour pouvoir qualifier un effet d'"important" ou de "faible". En revanche, lorsque la structure statistique du protocole est spécifiée, certaines valeurs de référence internes conduisent à de tels critères généraux. Cette approche, privilégiée dans la suite de ce texte, nous conduira à qualifier l'importance des effets à l'aide d'*effets-calibrés* (cf. §2.5.).

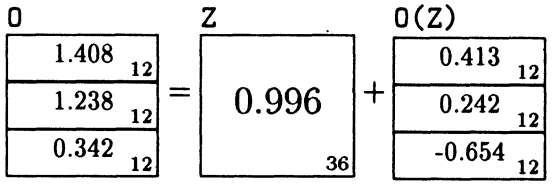
2.3. Décompositions liées à la structure "A"

Dans les exemples précédents, on a vu que le protocole initial se décompose additivement en protocoles dérivés, et ce de plusieurs façons. De façon générale, un protocole de support A admet toujours les deux décompositions additives suivantes :

$$\mathtt{A} = \mathtt{a1} + \mathtt{a2} + \cdots \quad (2)$$

$$\mathtt{A} = \mathtt{Z} + \mathtt{A(Z)} \quad (3)$$

La décomposition (3) est appelée la *décomposition canonique d'un facteur*. Avant tout, de telles équations signifient qu'il est possible de reconstituer de façon additive le protocole de gauche, à l'aide des protocoles de droite. Par exemple, l'équation (3), appliquée au protocole O/c1m1 considéré au §2.2. peut s'écrire :



Mais surtout, ces équations signifient qu'il y a décomposition additive des sommes des carrés bruts (Rss) et des ddl bruts (Rdf), et donc indépendance des différentes composantes. L'équation (2) découle immédiatemment de ce que a1, a2, *etc.* sont des protocoles disjoints (cf. II.§6.). Les décompositions (2) et (3) sont fondamentales : beaucoup des décompositions plus complexes qu'on va aborder s'en déduisent.

Enfin, on a également vu (§2.2.1.), que l'effet global d'un facteur, représenté par A(Z) à $(n-1)$ ddl pouvait être décomposé additivement en $(n-1)$ effets unidimensionnels. Un exemple d'une telle décomposition est :

$$\mathtt{A(Z)} = \mathtt{Z\ Diff\ a1,a2} + \mathtt{Z\ Diff\ a1\_a2,a3} + \mathtt{Z\ Diff\ a1\_a2\_a3,a4} \cdots \quad (4)$$

Pour le cas d'un facteur A2 à 2 modalités, cette équation redonne l'équivalence entre A(Z) et "Z Diff A", et à l'aide de ce résultat, l'équation (3) s'écrit alors :

$$\mathtt{A2} = \mathtt{Z\ Mean\ A} + \mathtt{Z\ Diff\ A} \quad (5)$$

2.4. Notion de "source de variation"

L'effet du facteur O, *i.e.* O(Z), se traduit par une certaine variance. On parlera ainsi de la *source de variation* "O(Z)". La notion de source de variation est plus générale que celle d'effet d'un facteur : il peut s'agir d'effets partiels, d'effets conjoints de plusieurs facteurs, d'effets d'interaction entre plusieurs facteurs, d'effets résiduels, *etc.*.

De ce point de vue, la composante Z de la décomposition (3) ne peut être qualifiée de source de variation, sauf quand elle est vue comme un "écart à une valeur de référence". Dans le dossier "Négligence" la variable (DEV) représente une déviation entre milieu subjectif et milieu objectif d'une baguette. On peut ainsi se poser des questions du type : "Telle déviation moyenne est-elle différente de *0cm* ?" Cette question revient à s'intéresser à l'écart "Z-*0*" et l'équation (3) peut se lire : "A=*0*+(Z-*0*)+A(Z)". Le terme Z peut alors être assimilé à une source de variation [7].

t Quand l'origine de la variable (le *0*) est arbitraire, le terme Z l'est aussi, et la "source de variation" Z peut être écartée de l'analyse. L'analyse d'un protocole de support A revient alors à celle du protocole centré A(Z), puisqu'on ne considère que des écarts à la moyenne générale. Mais, même dans ce cas, on retombe sur un *0* non-arbitraire dès qu'on considère un protocole dérivé par différence.

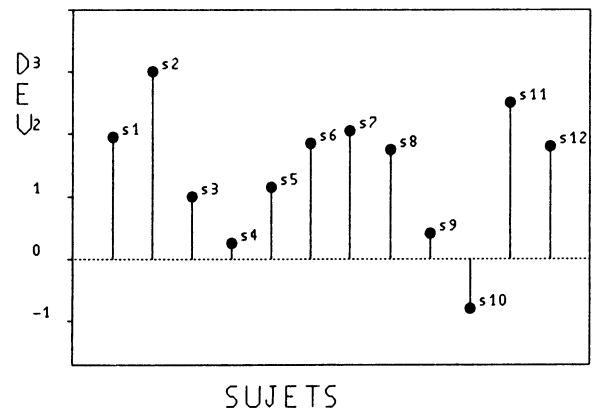
2.5. Structure statistique "S" ; notion d'effet-calibré

Ce qui précède ne fait appel qu'à la structure ensembliste du support. Des considérations supplémentaires doivent être introduites lorsqu'il est muni d'une *structure statistique*, c'est-à-dire ici lorsque le support représente un facteur de groupe. Prenons, par exemple, comme protocole initial, le protocole dérivé "S/c1m1o1->DEV" du dossier "Négligence". Ce protocole initial est donné ci-après sous forme d'un tableau et de deux graphiques (1 et 2) équivalents (pour les valeurs de DEV) [8]. Considérer S comme un facteur de groupe, revient à dire qu'on ne s'intéresse pas individuellement aux sujets, qu'on les traite de façon symétrique. Dans la représentation graphique du protocole, ceci amène en particulier à ignorer les identificateurs des sujets, comme dans le graphique 2.

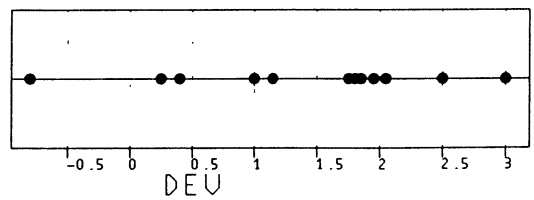
Table S/c1m1o1->DEV

s1	1.95
s2	3.00
s3	1.00
s4	0.25
s5	1.15
s6	1.85
s7	2.05
s8	1.75
s9	0.40
s10	-0.80
s11	2.50
s12	1.80

Graph S/c1m1o1 -> S,DEV (gr.1)



Graph S/c1m1o1 -> DEV,1 (gr.2)



[7] Si la question était : "La déviation moyenne est-elle supérieure à x *cm* ?", au lieu du protocole A->V1, on considérerait le protocole A->V1-x (à chaque valeur initiale, on enlève la valeur de référence x).

[8] Les graphiques de EyeLID (mot-clé "Graph") sont bivariés : ils sont obtenus par une demande du type "Graph ...->V1,V2". Quand le protocole est univarié, sa variable est affectée aux abscisses (V1) ou aux ordonnées (V2) ; la seconde "variable" pour le graphique peut être soit un facteur du protocole ("->S,DEV" pour le graphique 1), soit une constante ("->DEV,1" pour le graphique 2).

Du point de vue des résumés statistiques, cela amène à ne calculer que des *statistiques de groupe* : dénombrement, moyenne, écart-type, *etc.*. On remarquera, par exemple, que, pour 11 des 12 sujets, la valeur de DEV est supérieure à $0cm$, *i.e.* une déviation à droite (comme attendu d'après les hypothèses de l'expérience) et que, pour 9 d'entre eux, elle dépasse $1cm$. Pour l'analyse de la déviation moyenne, il est seulement nécessaire de prendre en compte le protocole Z, et les variations inter-individuelles prises dans leur ensemble, *i.e.* le protocole S(Z), sans chercher à décomposer ce dernier plus avant comme à l'équation (4). Ainsi, la décomposition canonique S=Z+S(Z) est particulièrement privilégiée.

Dans cette décomposition, la structure statistique conduit à distinguer : *a)* "Z" qui ne fait pas intervenir S, et qui constitue une *source systématique* ; *b)* "S(Z)" qui le fait intervenir, et qui constitue une *source adjointe*. Les décompositions analogues que nous introduirons par la suite, pourront comporter plusieurs sources systématiques et/ou plusieurs sources adjointes. Dans tous les cas, à chaque source systématique d'intérêt (dont on cherche à qualifier l'effet) sera associée une source adjointe particulière qui lui sert de *terme de référence* ; ici à Z est associée S(Z).

La distinction de ces deux types de sources de variation a des répercussions importantes à plusieurs niveaux : au niveau descriptif, elle offre la possibilité de définir un indice d'importance de l'effet à caractère "objectif" ; au niveau inductif, elle jette le pont vers les méthodes d'inférence statistique.

Au niveau descriptif, en effet, un indicateur privilégié d'importance de l'effet, l'*effet-calibré*, s'obtient en *rapportant la grandeur de l'effet systématique d'intérêt à la grandeur de l'effet adjoint associé*. Intuitivement, cela revient à "calibrer" l'effet d'intérêt, par un effet qui traduit les variations inter-individuelles ; cette procédure généralise la procédure psychométrique usuelle de calcul d'un "écart-réduit". Pour la structure statistique "S", l'effet calibré de Z, qu'on appellera plutôt l'*écart-calibré* de Z à *0* est ainsi défini par : Ec Z = Rsdcor Z/Rsdcor S(Z) [9]. Tous les effets-calibrés envisagés dans ce texte s'expriment sous cette forme générale : étant donné une source d'intérêt, notée "eff", et la source adjointe associée, notée "adj", l'écart-calibré est défini comme [10] :

$$\mathrm{Ec}_{eff} = \frac{\mathrm{Rsdcor}_{eff}}{\mathrm{Rsdcor}_{adj}}. \qquad (6)$$

Défini comme un rapport de deux indices de grandeur (tous deux descriptifs), l'indice d'importance de l'effet "Ec" est lui-même descriptif. De plus, il peut être qualifié d'"objectif" puisqu'il est déterminé uniquement par les données et la structure statistique. Enfin, cet indice a un caractère "absolu", dans la mesure où les valeurs qu'il prend pour divers effets, voire pour diverses expériences, sont comparables entre elles [11]. Ainsi, il est possible de fixer des *critères généraux d'importance de l'effet* à l'aide de "*valeurs-repères*" pour l'effet-calibré (Corroyer, Rouanet, 1994 ; Rouanet, 1994). Dans ce texte on adoptera les valeurs-repères de (Rouanet, 1994) : Ec > 0.6 pour un effet important, et Ec < 0.4

[9] Cette écriture n'est pas classique, bien qu'en fait équivalente à la définition usuelle : Ec Z = Mean Z/Sdcor S.

[10] Dans sa version actuelle (2.04), le logiciel EyeLID connaît seulement la structure ensembliste des données, et non leur structure statistique : la distinction entre source systématique et source adjointe lui est inconnue ; le mot-clé "Ec" n'y est donc pas implémenté.

[11] L'effet-calibré "Ec" défini ici est le même que dans (Rouanet, 1994). Il diffère, pour certaines questions, de l'indice du logiciel PAC (Lecoutre, Poitevineau, 1992). L'indice "Ec" présente la propriété désirable suivante. Considérons un vaste ensemble de sujets "S" sur lesquels on peut mesurer deux valeurs, correspondant à un facteur à deux modalités C2, et que ces valeurs sont non-corrélées ; l'étude de l'effet du facteur C2 peut se faire en recueillant des données, soit selon le plan S*C2, soit selon le plan équilibré S<C2> ; l'indice Ec a la même valeur (aux fluctuations d'échantillonnage près) quel que soit le plan adopté.

pour un effet faible, et $0.4 < \texttt{Ec} < 0.6$ pour un effet intermédiaire [12]. Bien entendu, ces valeurs-repères sont conventionnelles et ont donc seulement une valeur indicative.

Pour l'exemple `S/c1m1o1`, les indices de grandeur des deux termes valent, `Rsdcor Z`= 1.408*cm* et "`Rsdcor S(Z)`"= 1.060*cm*, d'où un écart-calibré : `Ec`= 1.328. On peut ainsi qualifier l'écart de la moyenne `Z` à *0cm* d'important.

[t] Au niveau inductif, comme on verra au chapitre V., la source systématique et la source adjointe associée servent respectivement de numérateur et de dénominateur pour le calcul des statistiques de test usuelles (t ou F). Ici, on obtient : $F = \frac{\texttt{Rms Z}}{\texttt{Rms S(Z)}} = 21.175$ à [`Rdf Z=1, Rdf S(Z)=11`] ddl ; ou de façon équivalente $t = \sqrt{F} = 4.602$ à [`Rdf S(Z)=11`] ddl. Mais il est plus parlant de voir l'indice inductif t comme `Ec` $\times \sqrt{12}$, écriture dans laquelle il apparaît comme fonction de deux choses : d'une part, un indice descriptif d'importance de l'effet (`Ec`), et d'autre part, le potentiel inductif des données (12, le nombre de sujets).

3. ANALYSE DES EFFETS LIÉS À UN EMBOÎTEMENT : "A<B>"

Sur le dossier "Horloge", de structure "`S<D*O>*C*A*L->ERR,TR`" (cf. I.2.4.), on se pose la question spécifique : "Quel est l'effet de la dimension suggérée (facteur `D`) sur le temps de réponse (variable `TR`), pour l'ordre `o1` ?". Pour y répondre, on considère le protocole dérivé pertinent "`S<D>/o1->TR`", obtenu par restriction à `o1`, moyennage sur les facteurs condition `C`, angle `A`, et côté `L`, et en ne considérant que la variable `TR` [13]. On choisit ici d'analyser la question à un niveau individuel, et on conserve donc le facteur `S`. Le protocole initial a la structure `S6<D2>->TR`. On le considérera par la suite comme un protocole élémentaire (*i.e.* de poids tous égaux à 1).

3.1. Protocole initial, sous-protocoles et protocoles dérivés par moyennage

Le protocole initial est donné ci-après sous forme de tableau des valeurs, qui fait clairement apparaître la relation d'emboîtement de `S` dans `D`. On donne également, pour chaque sous-protocole, `S<d1>` et `S<d2>` diverses statistiques descriptives résumées [14].

[12] Pouvoir qualifier un effet de "faible" revient à conclure qu'un certain *modèle* (celui d'un effet nul) est approximativement vérifié (ici au seul niveau descriptif) ; le sens de l'effet n'est alors pas pertinent. Par contre, la conclusion descriptive d'effet important, est toujours une conclusion orientée.

[13] Ce protocole peut se désigner d'autres façons en langage LID : "`S<D*o1> Mean C*A*L -> TR`" ou "`S<D> Mean C*A*L /o1 -> TR`". La première écriture épouse et rappelle la structure du protocole de base : le facteur `O` est remplacé par `o1` pour exprimer la restriction, et on explicite qu'on a moyenné sur `C`, `A` et `L` ; dans la seconde, on exprime la restriction directement à l'aide du symbole '`/`'. Dans l'écriture "`S<D> /o1 -> TR`", les moyennages sont implicites.

[14] Les statistiques de ces tableaux, définies en II.§4., s'obtiennent conjointement à l'aide des mot-clés `Desc` et `Odesc`. Le symbole '+' ne fait pas partie du langage LID et sert seulement à indiquer que les résultats de plusieurs demandes ont été juxtaposés, ou superposés (pour les graphiques).

Table S<D> -> TR

	d1	d2
s1	5.646	
s2	9.605	
s3	7.666	
s4	7.464	
s5	9.085	
s6	4.286	
s13		12.795
s14		12.104
s15		6.067
s16		9.065
s17		5.506
s18		6.458

Desc + Odesc S<d1> -> TR

Mean (W)	7.292
Minimum	4.286
Maximum	9.605
Range	5.319
Variance	3.414
Std. Dev.	1.848
Median	7.565
Quartile 1	5.646
Quartile 3	9.085

Desc + Odesc S<d2> -> TR

Mean (W)	8.666
Minimum	5.506
Maximum	12.795
Range	7.290
Variance	8.441
Std. Dev.	2.905
Median	7.761
Quartile 1	6.067
Quartile 3	12.104

A partir du protocole initial S<D>, on peut construire plusieurs protocoles dérivés par moyennage : D (ou encore "D Mean" et "D Mean S"), obtenu par moyennage sur S ; et Z par moyennage sur toutes les unités de S<D>.

Raw D -> TR

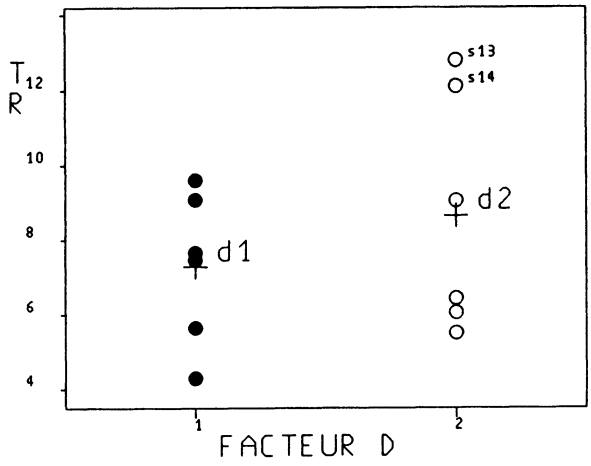
	x^u	w_u
d1	7.292	6
d2	8.666	6

Raw Z -> TR

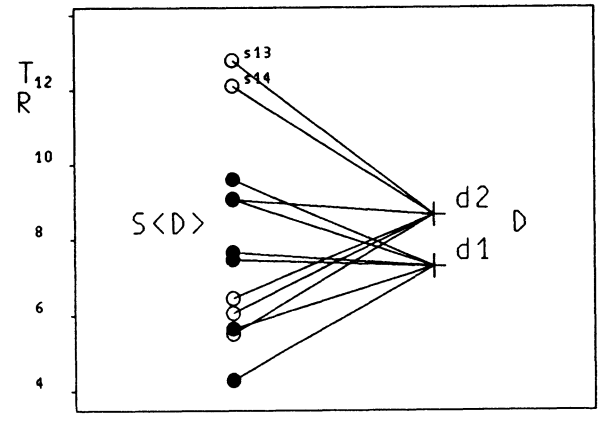
	x^u	w_u
z1	7.979	12

On a représenté simultanément les protocoles S<D> et D dans chacun des graphiques 3 et 4 [15]. Ces graphiques sont équivalents (du point de vue des valeurs de TR) mais chacun ne véhicule pas tous les aspects des données avec la même lisibilité.

Graph S<D>->D,TR + D->D,TR (gr.3)



Graph S<D>->1,TR + D->2,TR (gr.4)



Dans le graphique 3, les deux sous-protocoles S<d1> et S<d2> sont séparés. Les valeurs de TR sont proches dans les deux groupes, sauf pour deux sujets de d2 qui présentent des temps de réponse plus élevés (s13,s14) ; d'autre part, la dispersion des valeurs semble légèrement supérieure pour le groupe d2. Cette comparaison à vue est confirmée par les statistiques descriptives données précédemment.

Le graphique 4 donne un avant-goût de la notion essentielle de décomposition *inter-intra* (cf. §3.5.) : on visualise simultanément l'effet moyen du facteur D, et les effets individuels des sujets à l'intérieur de leur groupe respectif. On remarque que l'effet de

[15] Dans EyeLID on peut superposer graphiquement plusieurs protocoles, et les "joindre canoniquement" comme dans le graphique 4. Pour le graphique 3, les protocoles superposés sont "S<D>->D,TR" et "D->D,TR" ; pour le graphique 4, ce sont "S<D>->1,TR" et "D->2,TR".

D semble peu important si on le rapporte à ces différences inter-individuelles (comme on va voir, au §3.6., l'effet-calibré de D est peu important).

3.2. Protocoles dérivés caractérisant l'effet de D

Le facteur D a 2 modalités, et la caractérisation de son effet se fait comme au §2.1., à partir du protocole dérivé D : protocole dérivé par différence "Z Diff d2,d1", ou bien protocole centré D(Z). Ici on considère la différence "d2,d1", pour que l'effet soit positif.

Raw Z Diff d2,d1 -> TR

	x^u	w_u
z1	1.374	3

Raw D(Z) -> TR

	x^u	w_u
d1	-0.687	6
d2	0.687	6

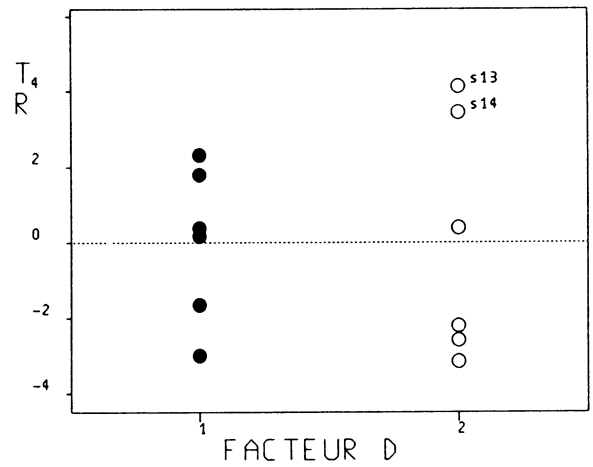
3.3. Protocoles dérivés "intras"

Si on centre séparément un sous-protocole, par exemple S<d1>, on obtient le *protocole dérivé intra*-d1, qui se désigne : S(d1) [16]. De même, on définit le protocole dérivé intra-d2 : S(d2). Enfin, le protocole dérivé intra-D, noté S(D), correspond à la réunion des protocoles intra-d pour toutes les modalités d de D. Le protocole S(D) est donné ci-après sous forme de tableau et de graphique (gr. 5).

Table S(D) -> TR

	d1	d2
s1	-1.646	
s2	2.313	
s3	0.374	
s4	0.172	
s5	1.793	
s6	-3.006	
s13		4.130
s14		3.438
s15		-2.599
s16		0.399
s17		-3.160
s18		-2.208

Graph S(D) -> D,TR (gr.5)



Dans le protocole intra-D, on centre séparément chaque groupe d, c'est-à-dire qu'on enlève à chaque sujet la moyenne de son groupe. Graphiquement, cette opération est simple à comprendre : on part du graphique 3 de S<D>, et on fait glisser chaque sous-protocole (S<d1> et S<d2>) de façon à amener leurs points moyens respectifs à *0*. Ainsi, on passe de S<D> à S(D), en "gommant" l'effet de D (et le niveau moyen de performance). Le protocole dérivé S(D) représente ainsi les variations inter-individuelles, une fois enlevées celles qui sont dues au facteur D. On parle d'un *protocole résiduel*.

3.4. Statistiques associées aux protocoles dérivés introduits

Le calcul des ddl se fait toujours à partir du nombre d'unités et du nombre de centrages indépendants (cf. II.§6.). Pour les protocoles non-centrés, le ddl est égal au nombre d'unités. Pour les protocoles centrés, ce dernier doit être réduit du nombre de centrages, d'où $(2-1)$ ddl pour D(Z), $(6-1)$ pour S(d1) et S(d2) et $(12-2)$ pour S(D).

[16] Il y a d'autres façons de désigner ce protocole dérivé, par exemple : "S(Z)/d1" ou "[S<d1>](Z)".

Le tableau ci-après donne les statistiques (`Rss`, `Rdf`) pour ces divers protocoles dérivés. Les décompositions `S<D>=S<d1>+S<d2>` et `S(D)=S(d1)+S(d2)` sont additives parce qu'il s'agit chaque fois d'une réunion de protocoles disjoints. Enfin, on retrouve l'équivalence entre `D(Z)` et "`Z Diff d2,d1`" déjà vue au §2.1. pour un facteur à 2 modalités.

Protocoles non-centrés	Rss	Rdf
`S<D>`	840.737	12
`S<d1>`	339.526	6
`S<d2>`	501.211	6
`D`	769.609	2
`Z`	763.948	1
`Z Diff d2,d1`	5.661	1

Protocoles centrés	Rss	Rdf
`S(D)`	71.128	10
`S(d1)`	20.482	5
`S(d2)`	50.647	5
`D(Z)`	5.661	1

3.5. Décompositions d'un emboîtement : `A<B>`

Pour un emboîtement quelconque `A<B>`, les sous-protocoles `A<b>` sont disjoints deux-à-deux. Il en est de même pour les protocoles intra-b `A(b)`. On a donc les deux décompositions :

$$\texttt{A<B>} = \texttt{A<b1>} + \texttt{A<b2>} + \cdots \quad (7)$$

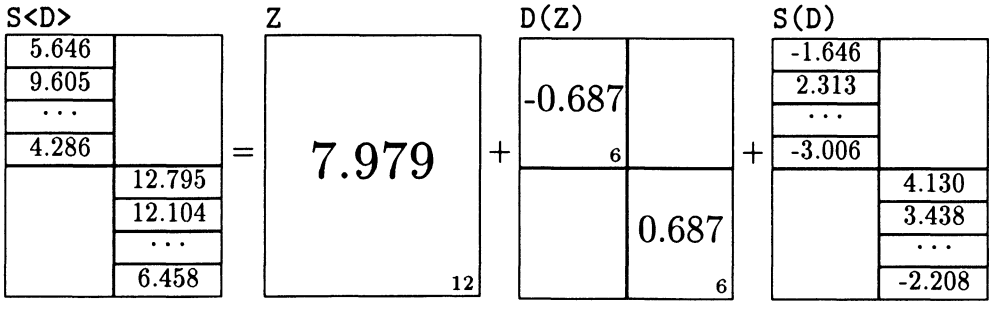
$$\texttt{A(B)} = \texttt{A(b1)} + \texttt{A(b2)} + \cdots \quad (8)$$

L'équation (3) appliquée à `A<b1>`, donne : `A<b1>=b1+A(b1)`; en faisant de même pour `A<b2>`, *etc.*, l'équation (7) devient : `A<B>=b1+A(b1)+b2+A(b2)+...`, ou encore `A<B>=b1+b2+...+A(b1)+A(b2)+...`. En appliquant (2) et (8), on obtient : `A<B>=B+A(B)`, dans laquelle on peut par (3) encore décomposer `B` en `Z+B(Z)`. En résumé, l'emboîtement `A<B>` admet les deux décompositions additives suivantes :

$$\texttt{A<B>} = \texttt{B} + \texttt{A(B)} \quad (9)$$

$$\texttt{A<B>} = \texttt{Z} + \texttt{B(Z)} + \texttt{A(B)} \quad (10)$$

La décomposition (10) est la *décomposition canonique d'un emboîtement*. Sur notre exemple, elle s'écrit :



Pour le sujet `s1` cette décomposition se lit : *a)* il a participé à l'expérience, d'où la composante 7.979 ; *b)* de plus il appartient au groupe `d1`, d'où −0.687 ; *c)* enfin il s'écarte de la moyenne de son groupe de −1.646 ; d'où son score 5.646 = 7.979 − 0.687 − 1.646.

La décomposition (10) peut aussi s'écrire en termes de protocoles centrés uniquement, en soustrayant, de chaque côté, le terme `Z`. Le terme de gauche devient alors "`A<B> - Z`", *i.e.* le protocole centré associé à `A<B>`, *i.e.* `[A<B>](Z)`. C'est souvent sous cette forme qu'est présentée la décomposition canonique d'un emboîtement; on parle alors de la

décomposition inter-intra ("inter-B", *i.e.* entre les modalités de B ; et "A-intra-B", *i.e.* A à l'intérieur de chaque modalité de B) [17] :

$$\texttt{[A<B>](Z)} \;=\; \texttt{B(Z)} + \texttt{A(B)}. \qquad (11)$$

Enfin, les protocoles centrés B(Z), A(b1) et A(b2) peuvent toujours se décomposer en plusieurs différences ou contrastes, comme on l'a fait au §2.2.1. (équation (4)).

3.6. Considérations liées à la structure statistique S<G>

Plaçons nous maintenant dans le cadre de la structure statistique S<G> — des "sujets" emboîtés dans des "groupes" —, où S est un facteur de groupe. C'est le cas dans l'exemple pris, avec G=D. Une question privilégiée est la comparaison des groupes, *i.e.* la source de variation G(Z). La source adjointe associée est S(G), qui représente les variations interindividuelles des sujets *à l'intérieur des groupes* [18]. La décomposition (11) est alors privilégiée. L'écart-calibré associé à G(Z) est donné par : Ec = Rsdcor G(Z)/Rsdcor S(G).

L'ensemble des statistiques pertinentes pour l'analyse de l'effet du facteur D est condensé dans le *tableau d'analyse de la variance* ci-après : chaque ligne correspond à une "source de variation" (*i.e.* un protocole dérivé), pour laquelle on donne diverses statistiques résumées. Traditionnellement, on y figure seulement les statistiques Rss, Rdf et Rms, suffisantes pour procéder à un test statistique (t ou F) concernant l'effet de D. On y a ajouté les statistiques descriptives caractérisant la grandeur et l'importance de l'effet. Ici la différence trouvée est de 1.374s en faveur de d2. L'écart-calibré vaut 0.364, et permet de qualifier l'effet de "faible", en prenant les valeurs-repères du §2.5..

Source de variation	Rss	Rdf	Rms	Diff	Rsdcor	Ec
D(Z)	5.661	1	5.661	-1.374	0.971	0.364
S(D)	71.128	10	7.113		2.667	

4. ANALYSE DES EFFETS LIÉS À UN CROISEMENT : A*B

On étudie ici les divers protocoles dérivés (et effets) associés au croisement de deux facteurs A et B. On se placera ici dans le cas d'un *croisement orthogonal* (le croisement est en fait équipondéré dans l'exemple qui suit), ce qu'on notera A $\perp$ B [19]. Certains résultats qui suivent ne sont valides que sous cette condition d'orthogonalité. Les cas "non-équipondéré" et "non-orthogonal" sont envisagés au chapitre IV.

4.1. Protocole initial

Sur le dossier "Horloge" de structure "S6<D2*O2>*C2*A6*L2->ERR,TR" (cf. I.§2.4. p. 14), on veut étudier l'effet conjoint des facteurs "dimension" (D) et "condition" (C) sur le temps de réponse (TR). Pour cela, on considère comme protocole initial le protocole dérivé par moyennage : D*C->TR (chaque moyenne porte sur 12 sujets × 6 angles × 2 côtés, d'où le

[17] On parle souvent de cette décomposition comme de la "décomposition inter-intra de la variance". Mais c'est d'abord une décomposition en terme de protocoles dérivés. On retrouve la décomposition usuelle des ddl, $(A-1) = (B-1) + (A-B)$, complètement explicitée par les opérations de centrage.

[18] Si la source systématique d'intérêt est Z, la source adjointe associée est également S(G).

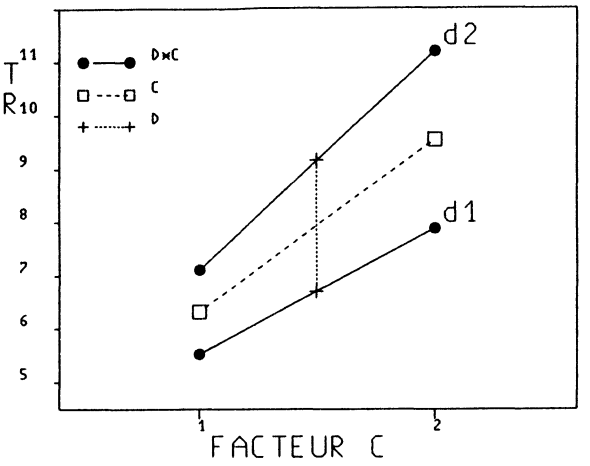
[19] Un croisement A*B, muni des poids w_{ab} est orthogonal si $w_{ab} = (w_a w_b)/W$, avec $w_a = \sum_b w_{ab}$, $w_b = \sum_a w_{ab}$ et $W = \sum_{ab} w_{ab}$. Le croisement est équipondéré si w_{ab} est constant. L'équipondération entraine l'orthogonalité.

poids de 144 par unité). Ce protocole est représenté de façon brute et de façon graphique ci-après.

Raw D*C -> TR

	c1		c2	
	x^u	w_u	x^u	w_u
d1	5.524	144	7.876	144
d2	7.107	144	11.193	144

Graph D*C + D + C ->C,TR (gr.6)



Quelle que soit la dimension d1 ou d2, la condition c2 donne un TR plus élevé que c1. Dans le graphique 6, cela se traduit par une pente positive des deux segments d1 et d2 (en traits pleins) : ces pentes représentent les *effets partiels* C(d). Mais la pente du segment d2 est plus forte que celle de d1. L'existence de différences entre les effets partiels C(d) traduit l'existence d'un *effet d'interaction* entre les facteurs C et D. On peut également lire le graphique du point de vue de D ; les effets partiels D(c) sont représentés par les écarts verticaux entre les points c/d1 et c/d2 ; l'existence d'une interaction se lit comme la différence entre ces écarts [20]. La notion d'interaction est, par essence, symétrique, même si, du fait du statut relatif des deux facteurs, on est souvent amené à la commenter dans un sens plutôt que dans l'autre.

Enfin, en construisant un autre segment passant au milieu des segments d1 et d2, on représente l'effet moyen, on dit aussi l'*effet principal*, du facteur C, de même que l'écart entre les "milieux" des segments d1 et d2 représente l'effet principal du facteur D. Ces deux effets sont également figurés dans le graphique 6 par des traits pointillés [21].

Nous allons étudier dans le détail les divers types d'effets que nous venons d'évoquer : effets partiels, effets principaux, et effet d'interaction.

4.2. Décomposition "inter-intra" d'un croisement

Envisageons d'abord une approche dissymétrique en nous interrogeant sur l'effet de C pour chaque dimension. Pour cela on considère d'abord le sous-protocole C/d1 de 2 unités à partir duquel on peut dériver les protocoles : d1 (moyenne), C(d1) (écarts à la moyenne), et "d1 Diff c2,c1" (différence) [22]. On procède de même pour le sous-protocole C/d2 et on obtient d2, C(d2) et "d2 Diff c2,c1". En réunissant ces protocoles (2 à 2 disjoints), on obtient les protocoles D, C(D) et "D Diff c2,c1" :

[20]Cette seconde lecture est un peu moins directe que la première, mais, si on avait construit le graphique "Graph D*C->D,DEV", tous les commentaires précédents s'inverseraient : les pentes deviendraient des écarts et vice-versa. Les deux graphiques en question sont connus comme les deux *diagrammes d'interaction* associés au croisement de 2 facteurs.

[21]Nous avons écrit "milieu" et "milieux" entre guillemets, parce que les choses ne sont pas si simples quand le protocole n'est pas équipondéré (cf. chapitre IV.).

[22]Pour le sous-protocole C/d1 (ou C*d1 selon les goûts), D est un facteur constant réduit à la seule modalité d1, d'où le fait qu'on peut dire simplement "d1", "C(d1)" et "d1 Diff c2,c1" au lieu de, respectivement, "Z/d1", "C(Z)/d1" et "Z Diff c2,c1/d1".

Raw D -> TR

	x^u	w_u
d1	6.700	288
d2	9.150	288

Raw C(D) -> TR

	c1		c2	
	x^u	w_u	x^u	w_u
d1	-1.176	144	1.176	144
d2	-2.043	144	2.043	144

Raw D Diff c2,c1 -> TR

	x^u	w_u
d1	2.352	72
d2	4.086	72

La décomposition "C/d=d+C(d)" est additive (cf. équation (3)), ainsi que "D*C=C/d1+C/d2" (réunion de protocoles disjoints). On en déduit l'additivité de la décomposition : D*C=D+C(D). A nouveau, par (3), le protocole D peut lui-même être décomposé en Z+D(Z), d'où finalement : D*C=Z+D(Z)+C(D). Du fait de la symétrie des facteurs C et D, on peut intervertir les rôles de C et D, et on obtient ainsi les deux *décompositions inter-intra* d'un croisement quelconque A*B [23].

$$A * B = Z + A(Z) + B(A) \quad (12)$$

$$A * B = Z + B(Z) + A(B) \quad (13)$$

Dans ces décompositions les termes A(Z) et B(Z) représentent les effets principaux des facteurs A et B. Sur notre exemple ces décompositions s'écrivent (D est en ligne, C en colonne) :

D*C

5.524 144	7.876 144	
7.107 144	11.193 144	

= Z

7.925 576

+ { D(Z)

-1.225 288
1.225 288

+ C(D)

-1.176 144	1.176 144
-2.043 144	2.043 144

} ou { C(Z)

-1.610 288	1.610 288

+ D(C)

-0.792 144	-1.659 144
0.792 144	1.659 144

}

Comme le facteur C a 2 modalités, les protocoles C(d) et "d Diff c2,c1" sont équivalents (pour les Rss et les Rdf), et donc le protocole C(D) est équivalent à "D Diff c2,c1". Symétriquement D(C) est équivalent à "C Diff d2,d1". On vérifie aisément ces propriétés, ainsi que l'additivité des décompositions (13) et (12) dans les tableaux de Rss et Rdf suivants :

Protocole	Rss	Rdf
D*C	38641.169	4
Z	36175.802	1
D(Z)	864.662	1
C(D)	1600.704	2
D Diff c2,c1	1600.704	2

Protocole	Rss	Rdf
D*C	38641.169	4
Z	36175.802	1
C(Z)	1492.443	1
D(C)	972.923	2
C Diff d2,d1	972.923	2

4.3. Interaction, protocole dérivé d'interaction

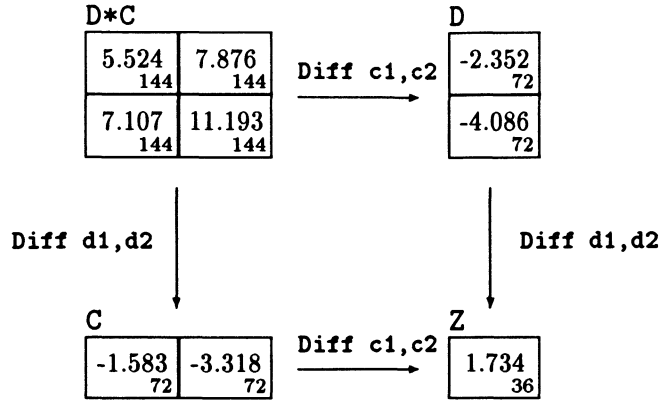
4.3.1. *Les deux visions dissymétriques de l'interaction*

L'existence d'une interaction s'exprime dans le protocole C(D) par le fait que les protocoles C(d) (pour d∈D) ne sont pas identiques, ou de façon équivalente par le fait que "D Diff c1,c2" n'est pas constant [24]. Mais cette dernière proposition veut dire encore que "D(Z) Diff c1,c2" n'est pas constamment nul, ou de façon équivalente

[23] Les décompositions (13) et (12) ne sont pas spécifiques de A*B et sont valables pour un facteur composé A&B quelconque. Pour le cas d'un emboîtement A<B> par exemple, l'équation (13) redonne la décomposition (8 p. 39), et l'équation (12) redonne la décomposition (3 p. 33) car le terme B(A) désigne alors un protocole constamment nul.

[24] On écrit plutôt "c1,c2" que "C" pour signaler qu'on utilise le fait que le facteur a 2 modalités.

que "Z Diff d1,d2 Diff c1,c2" n'est pas nul [25]. Ainsi, l'interaction apparaît comme une *différence de différences* ; l'unique valeur du dernier protocole constitue l'*effet d'interaction* qui vaut : 1.734. On l'obtiendrait également en permutant les rôles de C et D [26]. Ce qui vient d'être dit se résume par le diagramme suivant :



Dans la construction qui vient d'être faite, nous avons rencontré 4 protocoles dérivés qui représentent chacun l'effet d'interaction de façon équivalente, comme on pourra le constater dans le tableau de Rss et de Rdf suivant :

Protocole	Rss	Rdf
D(Z) Diff c1,c2	108.261	1
Z Diff d1,d2 Diff c1,c2	108.261	1
C(Z) Diff d1,d2	108.261	1
Z Diff c1,c2 Diff d1,d2	108.261	1

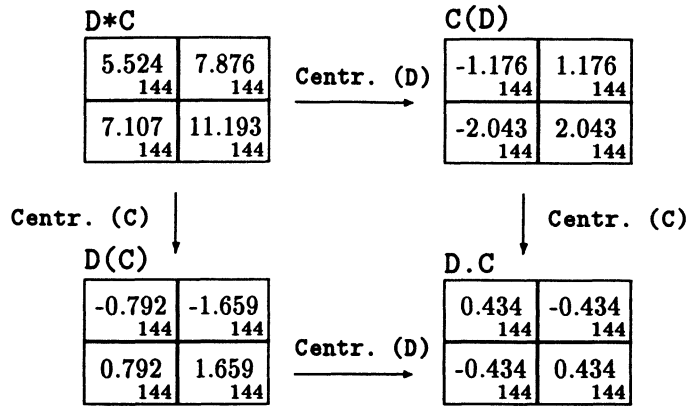
Chacun de ces protocoles équivalents a 1 ddl, d'où le fait qu'on peut caractériser l'effet d'interaction par une seule valeur numérique 1.734. Si, plus généralement, on était dans le cas d'un croisement A*B2, l'effet d'interaction serait multidimensionnel (à $(A-1)$ ddl), avec comme représentant le protocole "A(Z) Diff b1,b2".

4.3.2. *La vision symétrique de l'interaction*

Les protocoles construits précédemment représentent l'interaction mais en en faisant une lecture dissymétrique : "l'effet de C n'est pas le même selon les niveaux de D", ou bien "l'effet de D n'est pas le même selon les niveaux de C". Il est possible de représenter l'interaction de façon symétrique par le *protocole dérivé d'interaction*, qu'on désigne par "D.C" (ou "C.D"). On obtient ce protocole par *double centrage* à partir de D*C : un premier centrage selon C conduit à D(C), un second centrage selon D à partir de D(C) conduit à "D.C". On l'obtient également en centrant d'abord selon D puis selon C. Autrement dit, le diagramme suivant commute (parce qu'ici D*C est orthogonal) :

[25] On recourra fréquemment à de telles dérivations "en cascade", qui sont possibles du fait de la *récursivité* du langage LID. Par exemple, la demande "Z Diff d1,d2 Diff c1,c2" désigne un protocole dérivé d'une unité ("Z") dérivé par différence entre d1 et d2 ("Diff d2,d1") lui-même obtenu par différence entre c1 et c2 ("Diff c1,c2"). En termes d'ordre des dérivations, la lecture des demandes doit donc se faire de *la droite vers la gauche*.

[26] Les deux dérivations "Diff c1,c2" et "Diff d1,d2" *commutent*, *i.e.* l'ordre dans lequel elles sont effectuées n'importe pas ; nous verrons au chapitre IV. certaines dérivations qui ne commutent pas.



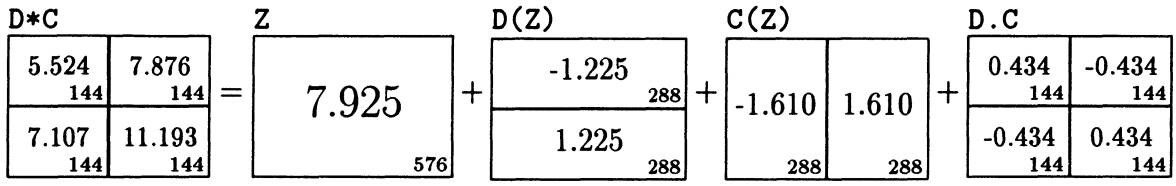
Le protocole "D.C" ainsi obtenu est centré à la fois selon C, et selon D : la propriété d'être *doublement centré* est caractéristique du protocole d'interaction entre deux facteurs quelconques. Ici, du fait que le croisement est orthogonal, la procédure de "double-centrage" suffit pour parvenir à un "protocole doublement-centré".

Une autre façon de présenter ce résultat est de dire que D(C) se décompose en "D(Z)+D.C" et C(D) en "C(Z)+D.C". Si on réintroduit l'un de ces résultats dans (12) ou (13), pour un croisement A*B, on obtient la *décomposition canonique d'un croisement en effets principaux et effet d'interaction (15)* :

$$A(B) = A(Z) + A.B \quad \text{ssi} \quad A \perp B \quad \text{d'où} \tag{14}$$

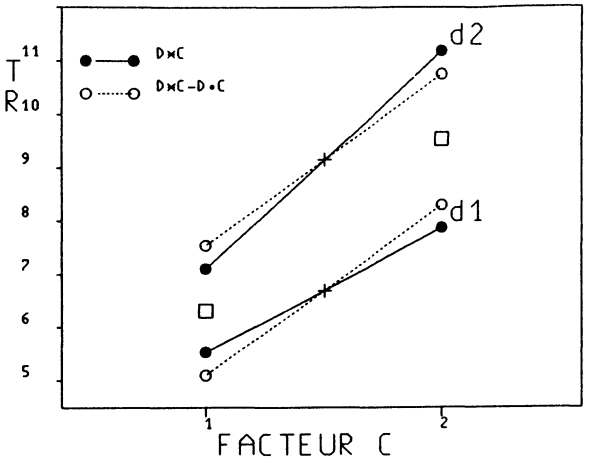
$$A * B = Z + A(Z) + B(Z) + A.B \quad \text{ssi} \quad A \perp B \tag{15}$$

La décomposition (15), pour notre exemple, est figurée ci-après. On vérifiera que le calcul direct de "Rss D.C" redonne la valeur 108.261 déjà trouvée au §4.3.1.. Comme les autres protocoles représentant l'interaction construits au §4.3.1., le nombre de ddl de "D.C" vaut 1 ; la justification en est donnée au §4.3.3..



Le graphique 7 permet de comprendre la signification des valeurs du protocole D.C. On y a représenté simultanément le protocole initial D*C, les protocoles moyens D et C, et le *protocole sans interaction* (obtenu en soustrayant terme-à-terme D.C à D*C). Dans le protocole initial, D*C, l'effet d'interaction se voit par le non-parallèlisme des segments d1 et d2. Le protocole sans interaction s'obtient à partir de D*C en faisant pivoter chaque segment autour de son point moyen (on ne change pas D) jusqu'à l'obtention de segments parallèles (on enlève l'interaction) et en conservant les mêmes points moyens pour les extrémités gauches et droites des segments (on ne change pas C). Le protocole d'interaction, quant à lui, correspond aux 4 écarts entre un '•' et un 'o'.

`Graph D*C->C,TR + D + C + D*C-D.C` (gr.7)



4.3.3. *Nombre de ddl de "A.B"*

Le nombre de ddl du protocole d'interaction "`A.B`" découle de sa caractérisation comme protocole doublement centré. Prenons l'exemple d'une interaction "`A.B`" entre deux facteurs `A3` et `B4`. Le tableau ci-après montre comment on peut utiliser notre *liberté* pour le remplir : chaque chiffre indique l'ordre de remplissage, les chiffres en *italiques* indiquent les cases choisies librement, les chiffres en **gras** correspondent aux cases déjà contraintes par les choix antérieurs. La formule du ddl de "`A.B`" est dès lors évidente [27] :

`A.B`

	b1	b2	b3	b4
a1	*1*	*4*	*7*	**10**
a2	*2*	*5*	*8*	**11**
a3	**3**	**6**	**9**	**12**

$$\texttt{Rdf A.B} = (A-1)\times(B-1) \qquad (16)$$

4.4. Structure statistique "S*T"

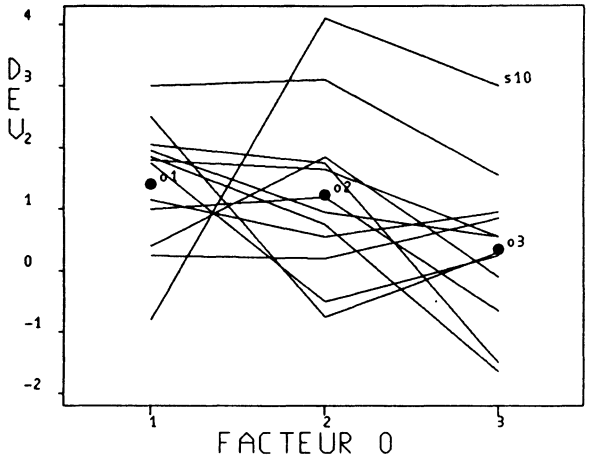
Considérons la question "`O/c1m1->DEV`" sur le dossier "Négligence" : "Quel est l'effet de l'orientation (`O`) sur la déviation (`->DEV`) en condition active (`c1`) pour la main gauche (`m1`) ?". On se demande de plus si la déviation moyenne en condition `c1m1` diffère de la valeur *0cm* : "`Z/c1m1->DEV`". On cherche à répondre à ces deux questions au niveau individuel. Pour cela on considère comme protocole initial le protocole dérivé `S*O/c1m1->DEV` où `S`, les "sujets", est un facteur de groupe. Ce protocole élémentaire a la structure `S12*O3`, et nos deux questions s'expriment par les effets `Z` et `O(Z)`.

Dans le cadre général de la structure statistique "`S*T`", qu'on rencontre ici, la décomposition additive "`S*T=Z+S(Z)+T(Z)+S.T`" comporte deux termes systématiques `Z` et `T(Z)`. Les deux autres termes correspondent à deux aspects distincts des variations inter-individuelles : `S(Z)` représente les variations de niveau moyen de performance des sujets et sert de terme adjoint à `Z` ; "`S.T`" représente les variations des effets individuels de `T` et sert de terme adjoint à `T(Z)`.
On notera, de plus, que si on s'intéresse uniquement à la question `Z`, il suffit de seulement considérer le protocole dérivé `S` puisque sa décomposition fournit les deux termes nécessaires : `Z` et `S(Z)`. De même, l'étude de l'effet `T(Z)` ne nécessite que le protocole `T(S)` puisque celui-ci se décompose en `T(Z)` et `S.T`. Ainsi, le protocole `S` est le protocole dérivé pertinent pour la question `Z`, et `T(S)` est pertinent pour `T(Z)`.

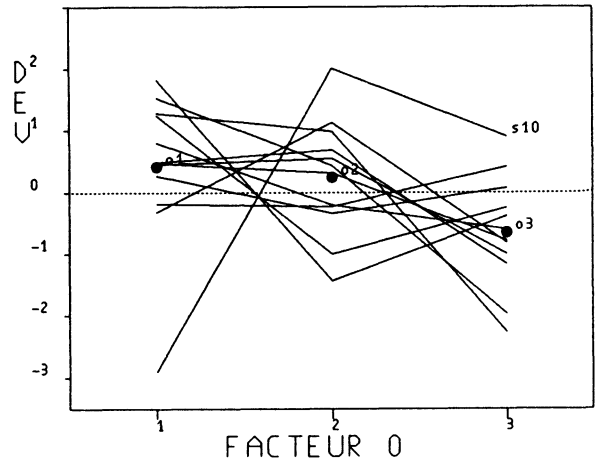
[27] Ceci se généralise aux interactions d'ordre supérieur ; par exemple, un protocole dérivé "`A.B.C`" est centré selon, à la fois, `A*B`, `A*C` et `B*C`, d'où son ddl : $(A-1)(B-1)(C-1)$.

Revenons à notre exemple. Le protocole initial S*O est figuré dans le graphique 8 ; chaque "profil" y représente un sujet, avec ses 3 valeurs selon O. Le graphique 9 du protocole O(S) s'obtient, à partir de S*O en enlevant la moyenne de chaque sujet. Le caractère plus ou moins parallèle des profils (*i.e.* l'interaction "S.O") ne change pas ; chaque profil est simplement translaté verticalement pour amener sa moyenne à *0*. L'effet moyen de O n'est pas affecté par la dérivation O(S) et est donc identique dans les deux graphiques : il se caractérise, en particulier, par une décroissance de o1, à o2, puis à o3. Mais on notera que seuls 4 profils individuels, parmi 12, présentent cette même caractéristique. Enfin, les deux composantes des variations individuelles, S(Z) et "S.O", sont présents dans le graphique 8 ; mais seule la composante "S.O" subsiste dans le graphique 9.

Graph S*O -> O,DEV (gr.8)



Graph O(S) -> O,DEV (gr.9)



En ce qui concerne l'importance des effets, pour comparer la moyenne générale Z à la valeur *0*, on calibre "Rsdcor Z" par "Rsdcor S(Z)" ; pour se prononcer sur l'effet global du facteur O, on calibre "Rsdcor O(Z)" par "Rsdcor S.O". L'analyse de ces deux effets est résumée dans le tableau d'analyse de la variance qui suit. L'écart de Z à *0cm*, près de 1*cm* en moyenne, apparaît important (Ec= 1.439), et l'effet global du facteur O est intermédiaire (Ec= 0.450).

Source de variation	Rss	Rdf	Rms	Mean	Rsdcor	Ec
Z	35.701	1	35.701	0.996	0.996	1.439
S(Z)	15.799	11	1.436		0.692	
O(Z)	7.878	2	3.939		0.573	0.450
S.O	35.665	22	1.621		1.273	

t Si, au lieu de la question sur l'effet global de O, on considère seulement l'effet partiel "o1_o2,o3", le protocole dérivé pertinent est alors "[o1_o2,o3](S)", et se décompose en "[o1_o2,o3](Z)" et "S.o1_o2,o3". Mais, comme l'opposition "o1_o2,o3" a 1 ddl, tout peut être retraduit de façon équivalente en termes de protocoles dérivés par différence "Diff o1_o2,o3" : le protocole pertinent devient le *protocole des différences* "S Diff o1_o2,o3" qui se décompose en "Z Diff o1_o2,o3" et "S(Z) Diff o1_o2,o3".

Cette remarque a une portée générale : dans le cadre de la structure statistique S*T, toute comparaison à 1 ddl sur T, peut se faire de façon équivalente en deux étapes : *a)* construire d'abord un protocole dérivé par différence, de structure S ; *b)* analyser l'écart de la moyenne de ce protocole à la valeur *0*, en le décomposant en Z et S(Z) comme on l'a fait au §2..

Selon qu'on adopte l'une ou l'autre manière de procéder, certains indices seront différents, comme par exemple les Rsdcor ; par contre l'effet-calibré Ec ne sera pas affecté.

5. ANALYSE DES EFFETS LIÉS À LA STRUCTURE "A<B>*C"

Considérons à nouveau le dossier "Négligence", dont le protocole de base a pour structure : S12<C2>*M2*03->DEV (cf. I.§2.1. p. 10). On désire étudier, au niveau individuel, les effets conjoints de la condition (facteur C) et de la main (facteur M) sur la déviation (variable DEV) en se restreignant à l'orientation o1. Pour répondre à cette question, on considère, comme protocole initial, le protocole dérivé "S<C>*M/o1->DEV" de structure S12<C2>*M2.

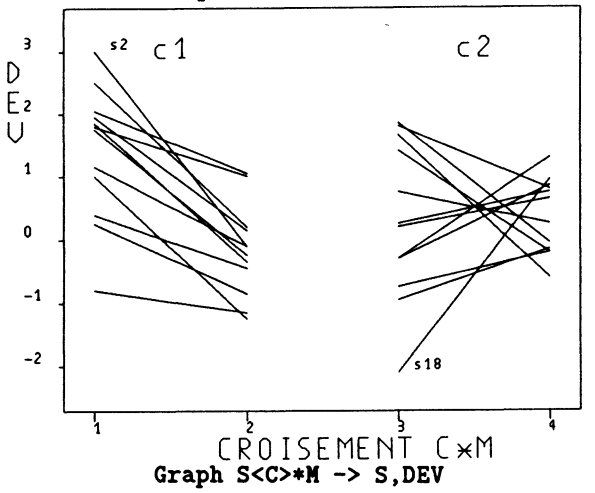
5.1. Protocole initial

Le tableau des valeurs du protocole initial est donné ci-après (les poids omis sont tous égaux à 1). On l'a également représenté sous forme graphique de deux façons différentes (graphiques 10 et 11). Pour faciliter la compréhension des dérivations que nous allons aborder, dans chaque graphique on a identifié les sujets s2 et s18.

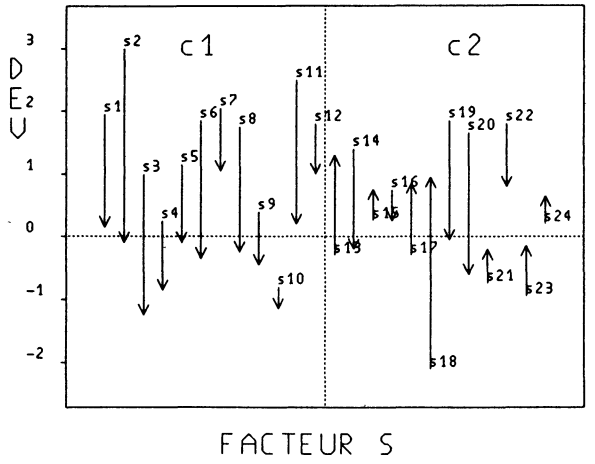
Table S<C>*M -> DEV

		m1	m2
	s1	1.95	0.15
	s2	3.00	-0.10
	s3	1.00	-1.25
	s4	0.25	-0.85
	s5	1.15	-0.10
c1	s6	1.85	-0.35
	s7	2.05	1.05
	s8	1.75	-0.25
	s9	0.40	-0.45
	s10	-0.80	-1.15
	s11	2.50	0.20
	s12	1.80	1.00
	s13	-0.30	1.30
	s14	1.40	-0.20
	s15	0.25	0.75
	s16	0.75	0.25
	s17	-0.30	0.85
c2	s18	-2.10	0.95
	s19	1.85	-0.05
	s20	1.65	-0.60
	s21	-0.75	-0.20
	s22	1.80	0.80
	s23	-0.95	-0.15
	s24	0.20	0.65

Graph S<C>*M -> CM,DEV (gr.10)



Graph S<C>*M -> S,DEV (gr.11)



Les deux graphiques sont équivalents en ce qui concerne les y (variable DEV). La seule différence tient aux facteurs figurant sur l'axe des x : CM, *i.e.* le croisement C*M pour le graphique 10 [28], et S pour le graphique 11. Chaque segment du graphique 10 représente un sujet et relie l'unité s<c>*m1 (à gauche) à s<c>*m2 (à droite) ; l'effet individuel de M se lit par la pente du segment. Dans le graphique 11, ces segments sont représentés par des

[28] A la description du protocole initial par les facteurs S, C et M, on a ajouté un "*facteur technique*" CM confondu avec C*M qui facilite les représentations graphiques. Le facteur CM a 4 modalités qui correspondent respectivement dans l'ordre à c1m1, c1m2, c2m1, c2m2.

flèches ; l'effet individuel de M se lit par la longueur et l'orientation des flèches. Il ressort de ces deux représentations que :

- L'effet de M est très homogène pour les sujets du groupe c1 : il y a toujours une diminution de DEV quand on passe de m1 à m2. Par contre, pour les sujets du groupe c2, les effets individuels de M sont beaucoup plus dispersés : on trouve à la fois des diminutions et des augmentations de DEV.
- En conséquence, l'effet moyen de M apparaît nettement plus élevé pour c1 que pour c2 : l'interaction entre C et M semble importante.
- L'effet moyen de C se lit en comparant le niveau moyen de déviation entre la partie gauche (c1) et la partie droite (c2) de chaque graphique. Il est difficile de bien l'apprécier sur le graphique 11, mais sur le graphique 10, il apparaît peu important. Cependant, du fait de l'importance de l'interaction entre C et M, cet effet moyen est de peu d'intérêt : il "mélange" un effet faible pour m2 à un effet relativement important pour m1.

5.2. Protocole dérivé "C*M"

On étudie l'effet conjoint de C et M à un niveau moyen, à l'aide du protocole dérivé des moyennes C*M donné ci-après. A partir de celui-ci, on peut analyser, comme on l'a fait au §4., les effets principaux de C et de M et l'effet d'interaction "C.M". Comme on l'a remarqué au vu des données individuelles, l'interaction "C.M" est élevée : pour la main gauche (m1), la condition active conduit à des déviations de $1.117cm$ plus à droite que la condition passive, alors qu'on observe un effet de sens opposé $-0.538cm$ pour m2. De ce fait l'effet moyen C(Z) n'a pas grand sens ; il correspond à une moyenne de deux valeurs très hétérogènes. Ainsi, à la décomposition symétrique "C*M=Z+C(Z)+M(Z)+C.M", on préférera plutôt la décomposition dissymétrique : C*M=Z+M(Z)+C(m1)+C(m2) (cf. éq. (12) p. 42). Les Rss et Rdf des protocoles associés à ces deux décompositions sont également indiqués ci-après.

Raw C*M -> DEV

	m1		m2	
	x^u	w_u	x^u	w_u
c1	1.408	12	-0.175	12
c2	0.292	12	0.363	12

Protocole	Rss	Rdf
C*M	26.766	4
Z	10.688	1
C(Z)	1.006	1
M(Z)	6.863	1
C.M	8.209	1

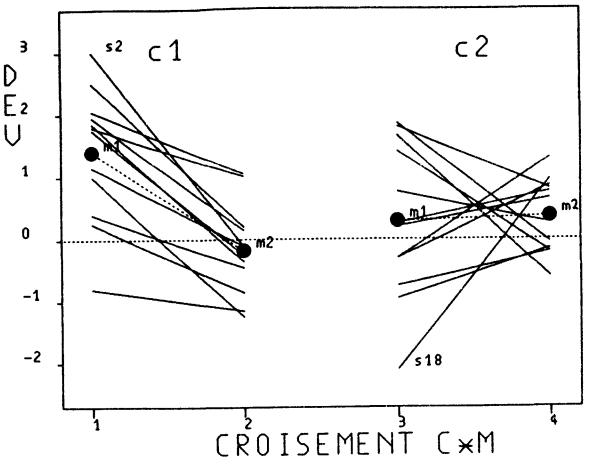
Protocole	Rss	Rdf
C*M	26.766	4
Z	10.688	1
M(Z)	6.863	1
C(m1)	7.482	1
C(m2)	1.733	1

5.3. Protocoles résiduels "S(C)" et "S(C).M"

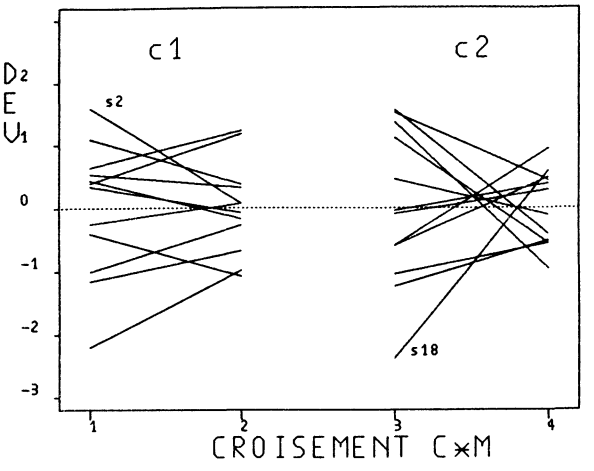
Si on désigne par F le facteur composé de C et M, F=C*M, on peut décomposer de façon additive le protocole S&F en "F+S(F)", *i.e.* en "C*M+S(C*M)" (cf. eq. (12) p. 42 et note 23 p. 42). Le premier terme vient d'être étudié et représente l'effet conjoint des facteurs C et M. Le second terme, S(C*M), représente les variations inter-individuelles, une fois enlevé ce qui est dû à C*M. Nous allons étudier dans le détail ce protocole résiduel.

Les graphiques 12 et 13 permettent de comprendre la décomposition qui vient d'être énoncée. Le graphique 12 représente le protocole initial S<C>*M auquel on a superposé le protocole des moyennes C*M. Le graphique 13 représente le protocole résiduel S(C*M), obtenu en ôtant, à chacun des points de S<C>*M, la moyenne correspondante de C*M. Graphiquement, cela revient simplement à déplacer (verticalement, et avec une rotation) chacun des deux ensembles de "segments" (les sous-protocoles S<c1>*M et S<c2>*M) de façon à amener les 4 points moyens de C*M à l'ordonnée *0*.

Graph S<C>*M + C*M -> CM,DEV (gr.12)

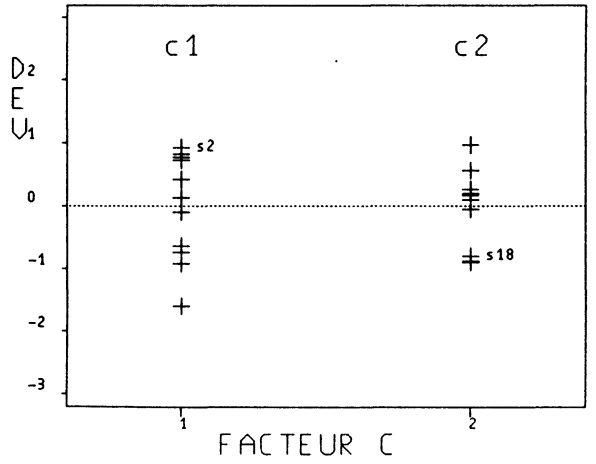


Graph S(C*M) -> CM,DEV (gr.13)

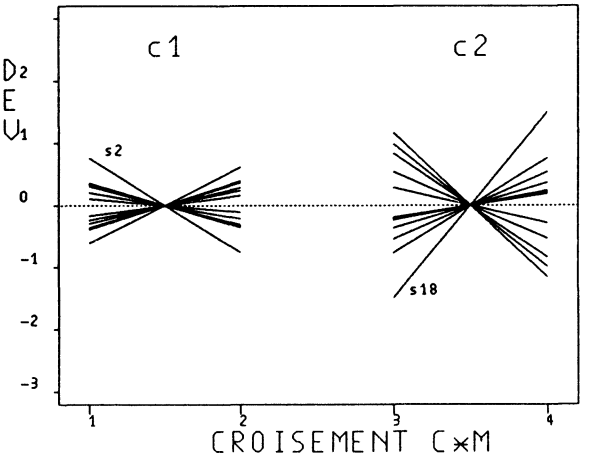


Le protocole résiduel général, S(C*M), peut lui-même être décomposé en deux protocoles résiduels, qui représentent deux facettes différentes de la variabilité inter-individuelle : *a)* dans S(C*M), cf. graphique 13, les sujets n'ont pas même moyenne (milieu d'un segment), d'où la composante S(C) ; *b)* l'effet de M (pente d'un segment) varie entre les sujets, d'où la composante "S(C).M" [29]. Ces deux protocoles sont figurés graphiquement ci-après.

Graph S(C) -> CM,DEV (gr.14)



Graph S(C).M -> CM,DEV (gr.15)



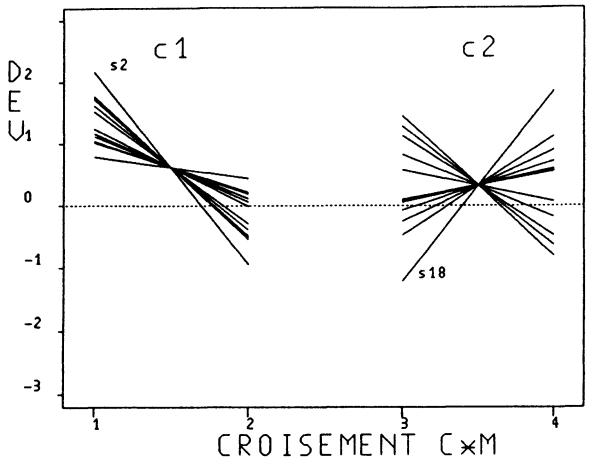
Le protocole S(C) (graphique 14) représente les variations inter-individuelles quant au niveau moyen de performance des sujets (à l'intérieur de chacun des deux groupes C=c1,c2). On passe de S(C*M) à S(C) en moyennant sur le facteur M ; graphiquement cette opération revient à remplacer chaque segment du graphique 13 par son milieu.

Le protocole "S(C).M", graphique 15, représente les variations inter-individuelles quant à l'effet du facteur M (à l'intérieur de chaque condition). On passe du protocole S(C*M) à "S(C).M" en enlevant le niveau moyen des sujets ; graphiquement ceci revient simplement à "saisir" chaque segment du graphique 13 par son milieu, et à le déplacer (verticalement et sans rotation) pour l'amener à l'ordonnée *0*. De même que S(C) est la réunion de S(c1) et S(c2), le protocole "S(C).M" apparaît comme la réunion des protocoles d'interaction "S.M/c1" et "S.M/c2". Sur le graphique 15 on visualise clairement la plus grande homogénéité des sujets de c1 quant à l'effet de M déjà évoquée.

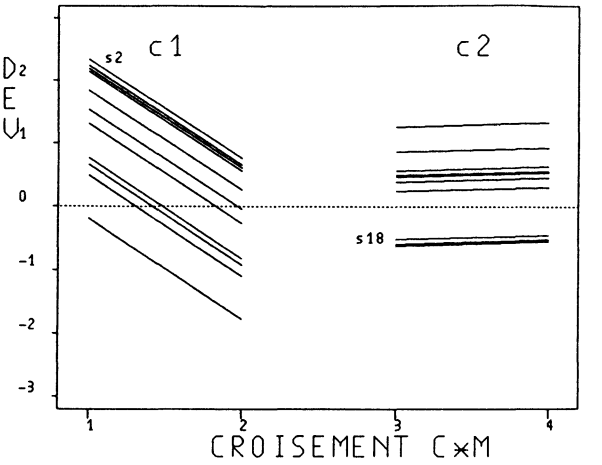
[29] La décomposition "S(C*M) = S(C) + S(C).M" n'est additive que sous certaines conditions sur la pondération (cf. eq. (19) p. 50).

On peut interpréter intuitivement les protocoles dérivés résiduels S(C) et "S(C).M" par une autre approche : regarder comment se présenterait le protocole initial si une de ces deux composantes était absente des données (*i.e.* si le protocole résiduel concerné était constamment nul) ; c'est ce qu'illustrent les graphiques 16 et 17. Dans le graphique 16 on a enlevé la composante S(C) ; ainsi tous les sujets de chaque groupe ont même moyenne de déviation : tous les segments se croisent en leur milieu. Dans le graphique 17 on a enlevé la composante S(C).M; ainsi tous les sujets de chaque groupe présentent le même effet du facteur M : dans chaque groupe tous les segments sont parallèles [30].

Graph S<C>*M - S(C) -> CM,DEV (gr.16)



Graph S<C>*M - S(C).M -> CM,DEV (gr.17)



5.4. Décompositions de la structure "A<B>*C"

On peut toujours, comme on vient de le signaler, décomposer A<B>*C en "B*C+A(B*C)", puis décomposer à son tour B*C par (12) ou (13). On obtient ainsi deux décompositions additives sans aucune condition restrictive (ces décompositions sont également valables pour un facteur composé A&B&C quelconque, en remplaçant '<>' et '*' par '&') :

$$A<B>*C = Z + B(Z) + C(B) + A(B*C) \quad (17)$$

$$A<B>*C = Z + C(Z) + B(C) + A(B*C) \quad (18)$$

Le protocole A(B*C) se décompose à son tour en deux protocoles disjoints A(b1*C) et A(b2*C). Chacun de ces deux protocoles, par exemple A(b1*C), est de structure A(C), et, à la condition A⊥C/b1, se décompose encore par (14) en "A(Z)/b1+A.C/b1", *i.e.* "A(b1)+A(b1).C", d'où la décomposition de A(B*C) :

$$A(B*C) = A(B) + A(B).C \quad \text{ssi } A \perp C/b1 \text{ et } A \perp C/b2 \quad (19)$$

Enfin, si B⊥C, le terme C(B) de (17) se décompose en "C(Z)+B.C", et en injectant (19) dans (17), on obtient la *décomposition canonique* de A<B>*C :

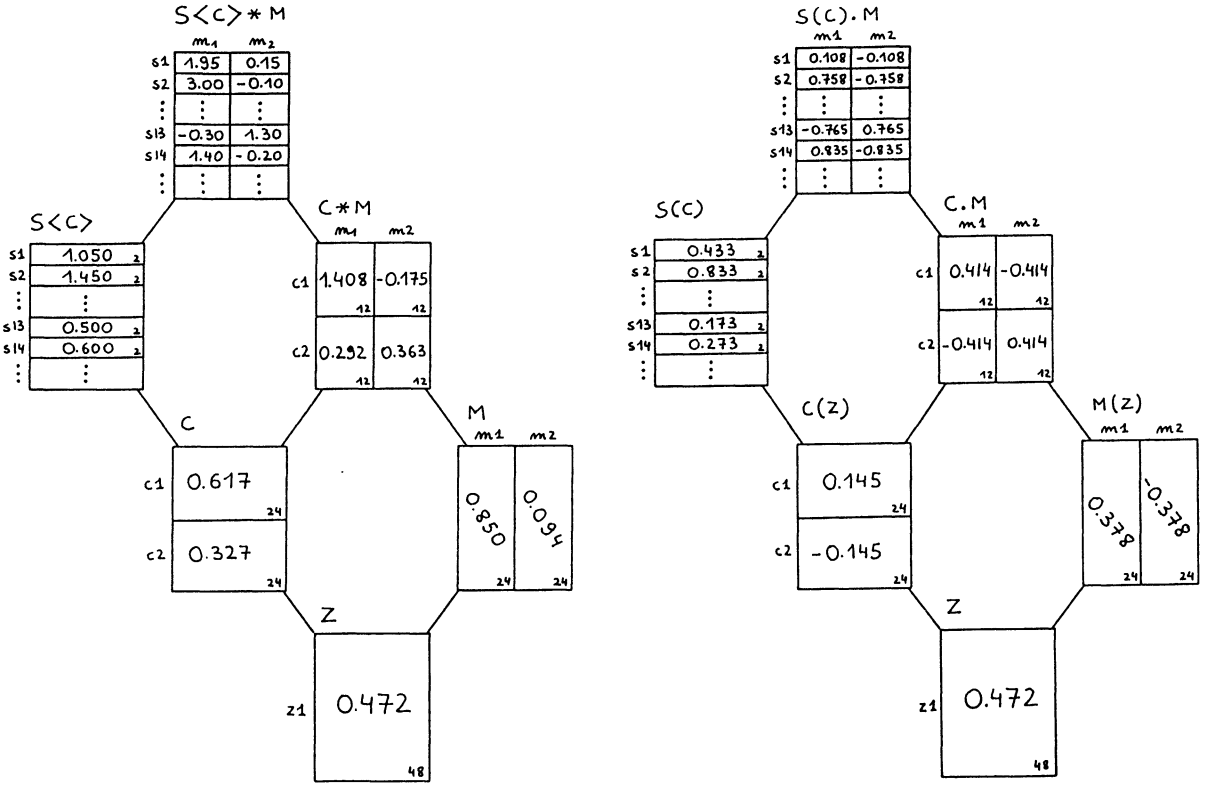
$$A<B>*C = Z + B(Z) + C(Z) + B.C + A(B) + A(B).C \quad (20)$$
$$\text{ssi } A \perp C/b1,\ A \perp C/b2 \text{ et } B \perp C$$

Explicitons cette décomposition sur notre exemple (où les conditions d'additivité sont vérifiées). Le diagramme 18 représente le treillis de finesse associé à S<C>*M (cf. I.§2.3.4.

[30] Le symbole '-', qui apparaît dans les titres des graphiques, ne fait pas partie du langage LID de EyeLID ; il indique qu'on a soustrait canoniquement deux protocoles entre eux (cf. chapitre VII.).

p. 12). Pour chaque élément du treillis, par exemple C*M, on a figuré le protocole dérivé (par moyennage) correspondant. De haut en bas du treillis, on trouve des résumés de plus en plus grossiers des données, jusqu'à Z, la moyenne générale. Le diagramme 19 représente les termes de la décomposition canonique ; on constate qu'ils correspondent terme à terme aux éléments du treillis. Avec ces deux diagrammes, on retrouve toutes les décompositions canoniques de ce chapitre : par exemple C*M du diagramme 18 se décompose avec tous les termes du diagramme 19 situés en dessous de "C.M" (en incluant celui-ci) : "Z + C(Z) + M(Z) + C.M".

Protocoles dérivés par moyennage (gr.18) Décomposition canonique (gr.19)



5.5. Structure statistique "S<G>*T"

Pour un protocole de structure statistique S<G>*T, où S désigne un facteur de groupe, la décomposition canonique (20) fournit quatre sources de variation systématiques (*i.e.* "Z", "G(Z)", "T(Z)" et "G.T") auxquelles correspondent les sources adjointes "S(G)", pour les deux premières, et "S(G).T", pour les deux dernières.

En appliquant ce résultat au protocole initial S<C>*M analysé dans cette section, dans lequel S est un facteur de groupe, on obtient la partie supérieure du tableau d'analyse de la variance qui suit. Celle-ci comporte les six sources de variation mentionnées, avec les statistiques habituelles ; la colonne "Effet" correspond à Mean pour "Z", à Diff pour "C(Z)" et "M(Z)", ou à "Diff C Diff M", pour "C.M" [31]. On y retrouve, entre autres, les termes de la décomposition de C*M déjà analysés au §5.2. p. 48.

[31] L'additivité de cette décomposition est ici vérifiée : l'addition des six termes redonne "Rss S<C>*M" = 65.878 et "Rdf S<C>*M" = 48. Mais, même dans le cas contraire, la décomposition (20) servirait toujours de base pour le choix des sources adjointes.

Source de variation	Rss	Rdf	Rms	"Effet"	Rsdcor	Ec
Z	10.688	1	10.688	0.472	0.472	0.667
C(Z)	1.006	1	1.006	0.290	0.205	0.290
S(C)	22.002	22	22.002		0.707	
M(Z)	6.863	1	6.863	0.756	0.535	0.606
C.M	8.208	1	8.208	1.654	0.827	0.938
S(C).M	17.109	22	0.778		0.882	
C(Z)/m1	7.482	1	7.482	1.117	0.790	0.683
S(C)/m1	29.408	22	1.337		1.156	
C(Z)/m2	1.733	1	1.733	-0.538	0.380	0.572
S(C)/m2	9.703	22	0.441		0.664	

On constate que la déviation moyenne "Z" ($0.472cm$) correspond à un écart-calibré à $0cm$ important : Ec= 0.667 > 0.6. L'effet moyen de la condition "C(Z)" apparaît faible : différence de $0.290cm$ en faveur de la condition active c1, soit un Ec= 0.290 < 0.4. L'effet de la main "M(Z)" (écart positif de $0.756cm$ en faveur de la main gauche) est également important : Ec= 0.606. L'effet d'interaction "C.M" est extrêmement important : effet d'interaction de $1.654cm$, soit un effet-calibré Ec=0.937.

Du fait de l'importance de cette interaction, déjà signalée, mais désormais attestée en terme d'effet-calibré, l'étude de l'effet du facteur C n'a de sens qu'au niveau local, *i.e.* conditionnellement à chaque main. D'où la partie inférieure du tableau : à chaque question spécifique "C(Z)/m" (qui peut aussi s'exprimer "C(m)"), on associe la source adjointe spécifique "S(C)/m". Avec cette analyse complémentaire, on conclut descriptivement que l'effet de C est important, dans le sens c1>c2, pour m1 (Ec= 0.683), et intermédiaire, mais dans le sens opposé, pour m2 (Ec= 0.572). Ces deux effets-calibrés sont plus proches l'un de l'autre que ne le sont les valeurs absolues des effets "bruts" ($1.117cm$ et $0.538cm$), dû au fait de la plus forte dispersion inter-individuelle en m1 qu'en m2.

6. ANALYSE DES EFFETS POUR DES STRUCTURES COMPLEXES

6.1. La décomposition standard

La démarche adoptée dans les sections précédentes s'étend à tout protocole dont le plan est quasi-complet (PQC). Des décompositions canoniques d'un facteur A (3), d'un emboîtement A<B> (10) et d'un croisement A*B (15), et de règles pour les combiner que nous détaillerons pas ici, on déduit la décomposition canonique d'un PQC quelconque (voir *e.g.* Duquenne, 1977). Pour un PQC, cette décomposition canonique épouse la structure du treillis de finesse des partitions, on en a vu une illustration en III.§5.4. p. 51 [32]. La décomposition canonique du protocole de base pris dans son ensemble constitue la *décomposition standard* de l'ANOVA évoquée au début de ce chapitre.

6.2. L'approche spécifique

Du fait de sa construction, la décomposition standard ne prend pas en compte le statut des facteurs. Cependant, par rapport aux objectifs de l'expérience, certains facteurs sont principaux, d'autres sont secondaires ; ceci se traduit par une dissymétrie des questions à leur égard, en particulier dans le cas où les facteurs interagissent de façon importante. On rencontre ainsi fréquemment la situation, où étant donnés un facteur principal A et

[32]Pour l'exemple du dossier "Négligence", la décomposition canonique du protocole de base "S<C>*M*O" comporte les termes : "Z", "C(Z)", "M(Z)", "O(Z)", "C.M", "C.O", "M.O", "C.M.O", "S(C)", "S(C).M", "S(C).O", "S(C).M.O" ; liste qu'on pourra confronter au treillis du chapitre I.§p. 13.

un facteur secondaire B, les questions d'intérêt sont du type "A/b" et où la question de l'effet moyen de B n'a pas d'intérêt. Dans un tel cas, l'analyse standard répond très imparfaitement aux objectifs de la recherche : elle ne fournit pas de réponse directe à certaines des questions d'intérêt, alors qu'elle en fournit à d'autres questions qui ne se posent pas. On la complètera alors par l'analyse de questions spécifiques.

6.2.1. *Principe de l'approche spécifique*

L'approche spécifique, dans son principe, peut être décrite selon les étapes suivantes :

- A toute question d'intérêt dans le cadre d'un protocole complexe, P, à tout effet, noté "eff", on peut associer *une ou plusieurs* sources adjointes, représentant chacune un aspect plus ou moins spécifique de la variabilité inter-individuelle.
- Ces diverses sources adjointes diffèrent quant au nombre d'unités du protocole de base qu'elles font intervenir et quant aux "échangeabiblités" entre sujets qu'elles supposent. Parmi ces sources adjointes, on retient *la source adjointe minimale*, notée "*adj*", *i.e.* celle qui a le plus petit Rdf ou, autrement dit, qui repose sur des conditions d'échangeabilité minimales [33].
- Etant donné les deux effets "eff" et "*adj*", on considère les protocoles dérivés qui les contiennent tous deux dans leur décomposition, et parmi ceux-ci celui qui a le plus petit Rdf, qu'on appelle "*protocole dérivé pertinent minimal (PDPM)*", P'.

A ce stade, l'analyse de la question "eff" sur le protocole de base P peut être ramenée à celle de "eff" sur le protocole dérivé minimal pertinent P' ; celui-ci est alors pris comme protocole initial sur lequel portera seul l'analyse. Ce protocole initial plus simple a souvent pour structure une des structures remarquables envisagées dans les sections précédentes, d'où justement leur caractère "remarquable".

6.2.2. *Source adjointe et protocole dérivé pertinent minimaux*

t On n'exposera pas ici dans le détail les règles qui permettent de déterminer la source adjointe et le protocole dérivé minimaux (le lecteur pourra se reporter à Hoc, 1983 ; Lecoutre, 1984) [34]. On se contentera d'en donner l'esprit pour la classe générale de questions à laquelle appartiennent toutes celles que nous avons envisagées dans ce chapitre : les *g-comparaisons* (notion de *comparaison* étendue aux moyennes, cf. chapitre V.) de "*type produit sur* G*T" pour la structure statistique S<G>*T.

Chaque g-comparaison de ce type se décompose en un produit de V, une g-comparaison sur G portant sur une partie G' de G, et de W, une g-comparaison sur T portant sur une partie T' de T. Dans la détermination de la source adjointe, les facteurs G et T interviennent de façon dissymétrique : pour G c'est G' qui est "minimal", mais pour T c'est W. Cette dissymétrie découle des relations différentes existant entre le facteur de groupe S et les facteurs G et T : S<G> et S*T. Ainsi, dans le cas particulier où la g-comparaison W a un seul ddl (moyennage sur T, sur T', différence entre t1 et t2, *etc.*), il suffit d'appliquer la dérivation correspondante à chaque sujet s, d'où une "valeur résumée" unique par sujet. Le PDPM sera alors le protocole de ces valeurs résumées, de structure S<G'>.

Par exemple, pour l'effet d'interaction partiel (eff) "g1_g2,g3.t1_t2,t3" dans le cadre d'un protocole de base S<G4>*T4, la source adjointe minimale (*adj*) est "S(g1,g2,g3).t1_t2,t3". Le PDPM (P') est alors "S<g1,g2,g3> Diff t1_t2,t3" de structure "S<G'>" ; au niveau de P' les deux sources eff et *adj* s'exprimeront respectivement par "g1_g2,g3" et "S(g1,g2,g3)".

[33] Ce point est encore plus déterminant lorsqu'on aborde l'analyse inductive puisqu'à l'échangeabilité des sujets viennent s'ajouter d'autres hypothèses (*e.g.* normalité, homogénéité) (cf. chapitre VI.).

[34] Ces règles font notamment partie intégrante des logiciels VAR3 et PAC (mais non de EyeLID dans sa version actuelle 2.04).

7. CONCLUSION

En guise de conclusion à ce chapitre, nous voudrions insister sur quelques aspects essentiels de l'ANACOMP, telle que nous l'avons présentée ici, par rapport à l'ANOVA traditionnelle : les trois premiers sont plutôt de nature pédagogique, les deux derniers plutôt d'ordre méthodologique.

Sources de variation, protocoles dérivés : Toutes les sources de variation de l'ANOVA sont assimilables à des protocoles dérivés. Pour chacun, on calcule, à l'aide de formules uniques, les statistiques usuelles du tableau de l'ANOVA (`Rss`, `Rdf` et `Rms`). Il n'est pas nécessaire d'exhiber une nouvelle batterie de formules à chaque nouveau plan (comme dans beaucoup de textes sur l'ANOVA). A un apprentissage technique des formules "magiques" de l'ANOVA, on a substitué l'apprentissage de concepts véhiculés par un langage des questions, le langage LID.

Visualiser les protocoles : Disposer d'un protocole dérivé pour chaque source de variation a une autre implication importante. Chacun peut être visualisé graphiquement, et ce, éventuellement, à l'aide de plusieurs graphiques, chacun en faisant ressortir tel ou tel aspect particulier. Ainsi l'interprétation intuitive des sources de variation complexes (interactions, sources "intra", termes résiduels, *etc.*) en est grandement facilitée.

Centrer ou ne pas centrer ? Dans les textes standard sur l'ANOVA, les questions envisagées ne concernent que des écarts entre valeurs ; l'analyse revient alors en fait à ne considérer du protocole de base que le protocole centré associé : le terme `Z` en est exclu dès le départ. L'analyse d'une question du type "comparer une moyenne à *0* ou à une valeur de référence" figure en général dans un chapitre séparé. L'approche présentée ici permet de traiter les deux types de questions dans un cadre unique.

L'importance des effets : Dans les tableaux d'analyse de la variance de ce chapitre, on a systématiquement indiqué l'effet-calibré associé à chaque source de variation, c'est-à-dire un indice *descriptif* mesurant l'importance de l'effet. Cet usage, s'il se répandait, permettrait d'éviter les interprétations erronées, qu'on lit malheureusement encore parfois : *a)* "Le test est non-significatif, *donc* il n'y a pas d'effet", *alors qu'en fait l'effet observé est important* ; *b)* "Le test est significatif, *donc* l'effet est important", *alors qu'en fait l'effet observé est faible* (pour plus de détails, voir chapitre VI.).

Moyennes ou individus ? Considérer le facteur "sujets" comme un facteur de groupe conduit, on l'a dit, à ne calculer sur les "sujets" que des statistiques de groupe. Parmi celles-ci, figure bien entendu la moyenne, mais cette dernière est tellement privilégiée en ANOVA qu'on tend à en oublier que les autres statistiques descriptives peuvent aussi constituer des critères de comparaison dignes d'intérêt : médiane, quartiles, étendue, fréquence des valeurs dépassant une valeur de référence, *etc.*.
En fin de compte, un protocole, dans lequel on a conservé les données individuelles mais en omettant les identificateurs des "sujets", constitue également une statistique de groupe, dont toutes les autres se déduisent. Lorsque ce protocole est représenté graphiquement, ces diverses statistiques "sautent aux yeux", ainsi que, parfois, des propriétés des données insoupçonnables au seul vu des tableaux de chiffres. Regarder les données individuelles est donc essentiel.

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 55–60)

IV. LES "AFFRES" DU DÉSÉQUILIBRE

RÉSUMÉ — *On considère ici le cas d'un croisement "A2*B2" non-équipondéré qui permet d'illustrer des problèmes plus généraux liés au déséquilibre du plan. On insiste sur les implications méthodologiques du déséquilibre.*

SUMMARY — The difficulties of the unbalanced case
*We consider here the case of an unbalanced 2 by 2 crossing "A2*B2", that enables to illustrate some general issues related to unbalanced designs. Methodological implications of unbalanceness are stressed.*

1. INTRODUCTION

Comme on l'a dit au chapitre II., le cas d'un protocole non-équipondéré nécessite, lorsqu'on en regroupe des unités, un choix entre moyennage pondéré et moyennage équipondéré. Une conséquence de ceci est qu'un certain nombre d'effets envisagés au chapitre III. ne sont plus définis de façon intrinsèque et peuvent parfois être grandement affectés par ce choix.

C'est une des forces de l'ANACOMP que d'avoir, d'abord, explicité les difficultés liées au déséquilibre, puis, permis de les résoudre en termes de *points de choix* qui s'offrent au chercheur : ce sont notamment les "options" pondérée et équipondérée proposées par les logiciels conçus dans ce cadre : VAR3, PAC et EyeLID. Un résultat majeur en a été d'établir une jonction entre les méthodes usuelles d'analyse des données expérimentales et celles de l'analyse des données d'observation pour lesquelles le déséquilibre est la règle (voir *e.g.* Le Roux, Rouanet, 1984a).

Dans ce chapitre, on considère un protocole initial de structure A*B en étudiant plus particulièrement le cas A2*B2. Au chapitre III.§4., on a étudié divers types d'effets qu'on peut définir dans le cadre d'une telle structure (moyenne générale, effets partiels, effets principaux, effet d'interaction) et donné sa décomposition additive en effets principaux et effet d'interaction. Que reste t-il de cet échafaudage, où tout est "d'équerre", lorsque le protocole n'est pas équipondéré ? On étudiera d'abord la non-unicité de certains des effets mentionnés (§2.), puis ses conséquences sur la détermination de décompositions additives (§3.). (Ces aspects sont étudiés en détail dans Le Roux, Rouanet, 1983 et 1984b). On conclura par l'examen des conséquences méthodologiques du déséquilibre (§4.).

2. NON-UNICITÉ DE CERTAINS EFFETS POUR UN PROTOCOLE PONDÉRÉ "A*B"

Pour pouvoir comparer avec le cas équipondéré, on prendra ici le même (valeurs identiques) protocole D2*C2 qu'au chapitre III. §4.1. p. 40, mais avec des poids différents.

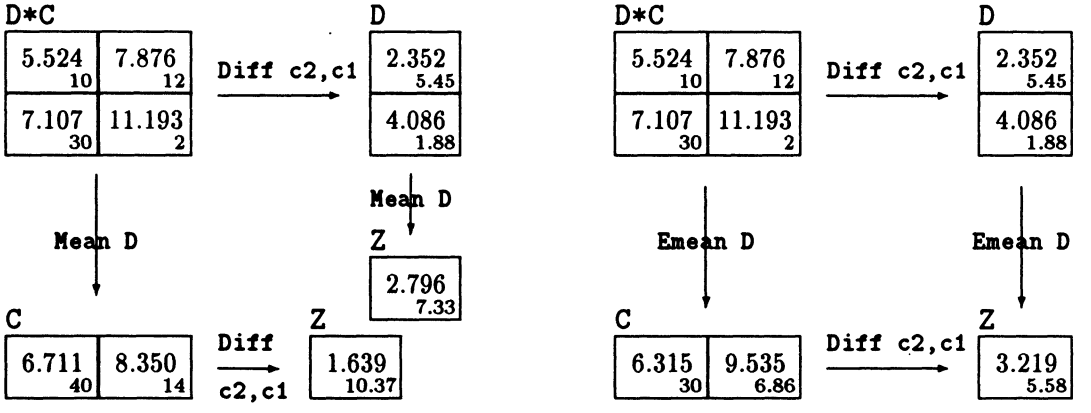
Raw D*C -> RT

	c1		c2	
	x^u	w_u	x^u	w_u
d1	5.524	10	7.876	12
d2	7.107	30	11.193	2

Le calcul des effets partiels, *e.g.* les effets de C conditionnellement à chaque modalité d de D, ne fait pas intervenir la pondération : `Diff c2,c1/d1`= 2.352 et `Diff c2,c1/d2`= 4.086. Ces effets sont insensibles à la disparité des poids et donc "intrinsèques".

2.1. Les divers effets principaux : pondéré, équipondéré, harmonique

Par contre, l'effet principal (ou moyen) de C fait intervenir la pondération, et n'est donc pas défini de façon univoque. En effet, calculer l'effet principal du facteur C implique deux dérivations : moyenner sur D et calculer par différence sur C. Un premier point de choix est, du fait du non-équilibre, celui entre "**Mean D**" et "**Emean D**" en ce qui concerne le moyennage sur D. Un deuxième point de choix vient du fait que l'*ordre dans lequel les deux dérivations sont effectuées peut affecter le résultat* : dans ce cas, on dit que la paire de dérivations *ne commute pas* [1]. Les diagrammes ci-après fournissent les diverses définitions possibles de l'effet de C :



Ces deux diagrammes appellent plusieurs remarques :

- La paire de dérivations "**Mean/Diff**" ne commute pas. Selon l'ordre dans lequel on les applique, on obtient l'effet "`Diff c2,c1 Mean D`"= 1.639 (l'"*effet pondéré*", qu'on désigne par `C[P]`) ou "`Mean D Diff c2,c1`"= 2.796 (l'"*effet harmonique*", qu'on désigne par `C[H]`). On note que l'effet harmonique est, par construction, nécessairement intermédiaire entre les deux effets partiels : ici 2.352 < 2.796 < 4.086. Par contre, ce n'est pas nécessairement vrai de l'effet pondéré : ici 1.639 < 2.352 et 1.639 < 4.086 [2].
- La paire de dérivations "**Emean/Diff**" commute, et on trouve l'"*effet équipondéré*", désigné par `C[E]` : "`Emean D Diff c2,c1`"= "`Diff c2,c1 Emean D`"= 3.219, qui, par construction, est la moyenne simple des deux effets partiels [3].

Sur le graphique 1, on a représenté simultanément : le protocole initial D*C (avec des "disques" dont la taille représente le poids associé) et trois segments dont les pentes respectives sont les trois effets que nous venons de définir. Le segment "`C[P]`" relie

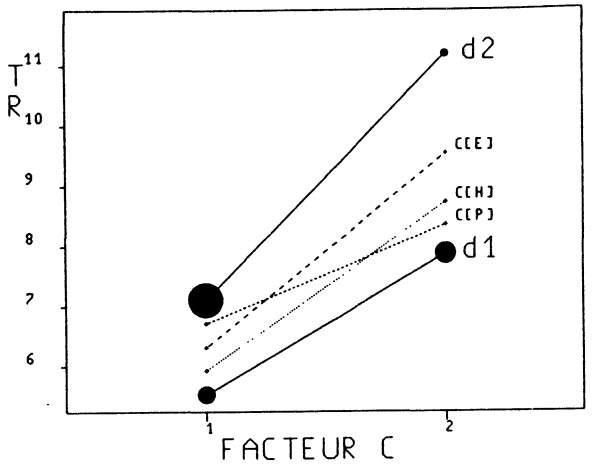
[1] Rappelons que le langage LID est récursif et permet ainsi des dérivations "en cascade" (*e.g.* "`Z Diff c2,c1 Emean D`") ; dans de telles demandes, l'ordre des dérivations se lit de *droite à gauche*.

[2] Ceci peut conduire à des résultats perçus comme "paradoxaux" : un effet pondéré nul alors que les effets partiels sont non-nuls et de même sens, voire un effet pondéré de sens opposé au sens des effets partiels. Ces "paradoxes" sont parfois désignés par l'expression "effets de structure"..

[3] Les écritures `C[P]`, `C[E]` et `C[H]` ne sont pas des formules de EyeLID.

les points moyens (pondérés) c1 et c2. Le segment "C[H]" représente un "segment moyen" entre les segments d1 et d2 (pondérés par les poids associés aux différences "Diff c2,c1/d", *i.e.* $5.45 = \frac{10\times12}{10+12}$ pour d1 et $1.88 = \frac{30\times2}{30+2}$ pour d2). Le segment "C[E]" représente l'effet équipondéré et peut se voir soit comme un segment reliant des points moyens (équipondérés), soit comme un segment moyen (équipondéré), du fait de la commutativité de "Emean/Diff".

Graph D*C->C,TR + C[P] + C[E] + C[H] (gr.1)



On visualise clairement l'"attraction" de l'effet pondéré C[P] par les unités de fort poids, et l'insensibilité aux poids de l'effet équipondéré C[E] (on pourra comparer avec le graphique 6 du chapitre III.§4.1. p. 41). On notera que ces deux effets seraient différents même en absence d'interaction entre les facteurs D et C. Par contre, en absence d'interaction, les effets équipondéré et harmonique coïncideraient.

2.2. Les autres effets : moyenne générale et interaction

La décomposition (III. (15) p. 44) d'un croisement D2*C2 en "Z+D(Z)+C(Z)+D.C" fait intervenir trois sortes de termes, selon les dérivations impliquées sur D et sur C : Z, la moyenne générale s'obtient par deux moyennages ; D(Z) et C(Z), les effets principaux, par un moyennage et une différence ; et D.C, l'effet d'interaction, par deux différences. Dans la section précédente, on a vu les conséquences du déséquilibre sur les effets principaux. Qu'en est-il pour la moyenne générale et l'interaction ?

- Il y a deux façons de moyenner, Mean et Emean, d'où d'abord quatre possibilités de définir Z, puis huit selon l'ordre des dérivations. Les paires de dérivations "Mean/Mean", "Emean/Emean" commutent et, de ce fait, on peut simplement parler de *la* moyenne pondérée, désignée par Z[P], et de *la* moyenne équipondérée, désignée par Z[E]. Par contre, les paires du type "Mean/Emean" ne commutent pas. On se retrouve donc avec en tout *six* moyennes générales différentes.
- Par contre, il y a une seule façon de calculer une différence, qui du fait de la commutativité de "Diff/Diff" conduit à *un unique* effet d'interaction.

2.3. Résumé des choix possibles ; conditions d'équivalence des effets

Le tableau 1 résume les conséquences de la non-équipondération sur les différents types d'"effets" envisagés ; les cases divisées en deux correspondent aux cas non-commutatifs.

Ce qu'on vient d'énoncer, pour un croisement D2*C2, est général et vaut pour le croisement de facteurs à un nombre quelconque de modalités. Dans ce cas, les

	Mean C	Emean C	Diff c2,c1
Mean D	Z[P] = 7.136	Emean C Mean D = 7.530 Mean D Emean C = 7.327	C[P] = 1.639 C[H] = 2.796
Emean D	Mean C Emean D = 6.914 Emean D Mean C = 7.085	Z[E] = 7.925	C[E] = 3.219
Diff d2,d1	D[H] = 1.906 D[P] = 0.556	D[E] = 2.450	D.C = 1.734

Tableau 1: *Résumé des effets pour un croisement "D2*C2" : six moyennes "générales", trois effets principaux pour* C *et* D*, un effet d'interaction.*

effets possibles sont encore ceux donnés par le tableau 1, mais les effets principaux et d'interaction représentent alors des effets multidimensionnels qui ne se ramènent pas au calcul d'une différences ou d'une différence de différence.

En passant d'une pondération quelconque à l'équipondération, certaines cases du tableau 1 s'agrègent et on obtient alors seulement les 4 effets de l'équation (III. (15) p. 44) (soit les 4 cases inférieures droites du tableau 1). Entre ces deux extrêmes, il y a plusieurs intermédiaires, selon les propriétés particulières de la pondération. On donne ci-après les conditions pour que, dans le cadre d'un croisement quelconque A*B, les effets pondéré A[P], équipondéré A[E] et harmonique A[H] coïncident (voir Le Roux, Rouanet, 1983 et 1984b). En notant w_{ab} la pondération de A*B, $w_a = \sum_b w_{ab}$, $w_b = \sum_a w_{ab}$ et $w = \sum_{ab} w_{ab}$, on a les propriétés :

- ssi la pondération est *équilibrée harmoniquement sur* B, *i.e.* $\forall b, b'$, $\sum_a 1/w_{ab} = \sum_a 1/w_{ab'}$, alors A[E]=A[H] ;
- ssi la pondération est *équilibrée sur* B *pour chaque* a, *i.e.* $\forall a\ \forall b, b'$, $w_{ab} = w_{ab'}$, alors A[E]=A[P] ;
- ssi la pondération est *orthogonale*, *i.e.* $\forall a\ \forall b$, $w_{ab} = (w_a w_b)/w$, alors A[P]=A[H].
- ssi la pondération est *équilibrée*, *i.e.* w_{ab} constant, alors A[P]=A[H]=A[E].

3. DÉCOMPOSITIONS D'UN CROISEMENT NON-ÉQUIPONDÉRÉ

t Revenons à la décomposition du croisement D*C en "Z+D(Z)+C(Z)+D.C". Dans le cas équilibré considéré au chapitre III.§4., cette décomposition est additive. Que devient-elle hors de ce cas, comment est-elle affectée par les choix précédents ?

On a vu (cf. III. (5) p. 33) qu'un facteur "D2" se décompose additivement en "Z Mean D+Z Diff D" ; de même "C2" se décompose en "Z Mean C+Z Diff C". De ces relations, on déduit deux décompositions additives de D2*C2 :

$$\texttt{D2*C2} = \texttt{Z Mean D Mean C} + \texttt{Z Diff D Mean C} + \texttt{Z Mean D Diff C} + \texttt{Z Diff D Diff C} \quad (1)$$

$$\texttt{D2*C2} = \texttt{Z Mean C Mean D} + \texttt{Z Mean C Diff D} + \texttt{Z Diff C Mean D} + \texttt{Z Diff C Diff D} \quad (2)$$

Puisque "Mean/Mean" et "Diff/Diff" commutent, les premiers et derniers termes de ces deux décompositions désignent respectivement le même effet : respectivement "Z[P]" et "D.C". Par contre, du fait que "Mean/Diff" ne commute pas, les termes intermédiaires (effet pondéré D[P] et effet harmonique C[H], dans (1), et inversement dans (2)) sont différents. On vérifie l'additivité, en termes de Rss et de Rdf, des décompositions (1) et (2) dans le tableau suivant :

Protocole	x^u	w_u	Rss	Rdf
D*C			2815.320	4
Z Mean D Mean C, *i.e.* Z	7.136	54	2749.803	1
Z Diff D Mean C, *i.e.* D[P]	-0.556	13.04	4.026	1
Z Mean D Diff C, *i.e.* C[H]	-2.796	7.33	57.295	1
Z Diff D Diff C, *i.e.* D.C	1.734	1.40	4.196	1
Z Mean C Mean D, *i.e.* Z	7.136	54	2749.803	1
Z Mean C Diff D, *i.e.* D[H]	-1.906	9.21	33.474	1
Z Diff C Mean D, *i.e.* C[P]	-1.639	10.37	27.847	1
Z Diff C Diff D, *i.e.* D.C	1.734	1.40	4.196	1

La généralisation des équations (1) et (2) à un croisement A*B quelconque donne les deux décompositions additives suivantes :

$$A*B = Z[P] + A[H] + B[P] + A.B \quad (3)$$

$$A*B = Z[P] + A[P] + B[H] + A.B\,. \quad (4)$$

L'effet pondéré Z[P] correspond, en langage LID de EyeLID au protocole "Z", de même que les effets pondérés A[P] et B[P], correspondent respectivement aux protocoles centrés "A(Z)" et "B(Z)". L'effet d'interaction "A.B" est toujours défini de manière intrinsèque bien que moins immédiat à obtenir [4]. La particularité des deux décompositions précédentes est donc l'intervention des termes harmoniques A[H] et B[H] [5]. Lorsque le croisement A*B est orthogonal, on a A[H]=A[P] et B[H]=B[P], et les équations (3) et (4) redonnent alors toutes deux la décomposition (III. (15) p. 44).
Dans les textes et les logiciels standard (*e.g.* SAS) sur l'ANOVA, on rencontre une batterie de décompositions du même type que (3) selon qu'on choisit pour l'un ou l'autre effet l'option [P], [E] ou [H] mais sous des appellations qui véhiculent mal les choix opérés ("types I, II, III, IV") et surtout qui ne considèrent qu'un ensemble restreint d'options. Par exemple, pour l'étude d'un croisement A*B dans le cadre d'un plan S<A*B>, on a les équivalences suivantes : I (A[H] et B[P]), II (A[H] et B[H]), III et IV (A[E] et B[E]).

4. CONSÉQUENCES MÉTHODOLOGIQUES DU DÉSÉQUILIBRE

Résumons les deux difficultés majeures entraînées par le déséquilibre, dont découlent les implications méthodologiques détaillées plus loin :

- La décomposition (III. (15) p. 44) d'un croisement A*B, en effets principaux pondérés et interaction, n'est pas additive dans le cas non-orthogonal. Ceci tient à ce que les facteurs A et B sont corrélés entre eux. Du coup, la question "A(Z)" n'est pas indépendante de la question "B(Z)" [6].
- Les effets principaux ne sont pas définis de façon intrinsèque et nécessitent un choix. Il en est de même pour la moyenne générale. Cette difficulté n'est pas spécifique de la structure de croisement, et se pose, dans le cadre d'un plan quelconque dès qu'un moyennage correspond à un regroupement d'unités non équipondérées.

[4] Lorsque le croisement A*B n'est pas orthogonal, le protocole dérivé d'interaction, ne s'obtient plus en centrant deux fois, d'abord selon A, puis selon B, comme on l'a fait au chapitre III.§4.3.2.. En effet, si on procède ainsi, le centrage selon B détruit le centrage selon A et, ainsi, le protocole dérivé obtenu n'est pas doublement centré. Cependant, en procédant *itérativement* à des paires de centrages "selon A, selon B", il y a convergence vers un protocole doublement centré, *i.e.* le protocole d'interaction (Guigues, 1981).

[5] En termes de protocoles dérivés, les effets harmoniques A[H] et B[H] sont moins immédiats à obtenir. Dans le cas général, l'effet A[H], par exemple, ne peut pas être représenté par un protocole de support A, mais seulement par un protocole de support A*B qui s'obtient par différence (terme à terme) entre les protocoles dérivés "A(B)" et "A.B". Dans le cas particulier d'un croisement A2*B, l'effet A[H] peut être représenté par le protocole "Z Mean B Diff A".

[6] On voit clairement cela sur le graphique 1 : l'effet C[P], *i.e.* C(Z), est "attiré" par les forts poids des points c1d2 et c2d1 et il en est de même pour l'effet D(Z). Ainsi les deux effets reflètent surtout tous deux la même différence "Diff c1d2,c2d1".

Quel effet choisir ? La réponse à cette question ne peut être générale : elle dépend des raisons de la disparité des poids. Lorsque cette disparité reflète une certaine réalité (*e.g.* les fréquences des modalités sont proches de leur fréquences dans la population d'intérêt), les effets pondéré et équipondéré correspondent à deux questions différentes posées aux données dont chacune a une certaine pertinence. Dans le cas contraire, la question correspondant à l'option pondérée n'aura, en général, pas de sens et on préfèrera le calcul d'effets équipondérés [7].
Supposons, par exemple, qu'on veuille comparer les salaires (variable) selon le sexe (facteur A) pour la population française à partir d'un échantillon représentatif. Si on introduit le facteur "niveau d'études" (facteur B), les deux questions évoquées seront : "Y-a t-il une différence de salaire selon le sexe ?", *i.e.* "**A Mean B**", qui revient à "oublier" l'existence du facteur B ; et "Y-a t-il une différence de salaire selon le sexe, à niveau d'étude constant ?", *i.e.* "**A Emean B**".

Interaction et non-équipondération : La non-unicité des effets principaux est indépendante de la présence ou absence d'interaction. La pondération d'un protocole dérivé découle de propriétés structurelles du protocole de base, alors que l'interaction est une propriété des valeurs observées. Les deux notions relèvent donc de niveaux différents. Par contre, la présence d'une interaction et la non-équipondération ont une même implication méthodologique, à savoir l'*importance à accorder aux effets partiels* : dans le premier cas, parce que ceux-ci étant différents, on a moins envie de les moyenner ; dans le second cas, parce que ceux-ci sont définis de façon intrinsèque.

Plans équilibrés et déséquilibre : Le recours à un plan équilibré, quand cela est possible, n'élimine pas entièrement le problème de la non-équipondération : dès qu'on procède à des comparaisons entre plusieurs sous-ensembles de modalités, de tailles différentes, on tombe à nouveau sur des protocoles non-équipondérés, et donc sur le choix entre moyennage pondéré et équipondéré.

Le problème des facteurs non-contrôlés : On a noté que l'effet d'un facteur A, calculé par moyennage pondéré sur B, peut ne pas être intermédiaire entre les effets partiels A/b, jusqu'à des situations extrêmes d'annulation ou d'inversion de l'effet (cf. note p. 56). Mais imaginons que l'expérimentateur n'a pas inclus le facteur B dans son plan d'expérience et a seulement contrôlé le facteur A (parce que, par exemple, le facteur B n'a pas encore été reconnu pertinent pour le domaine expérimental en question). Le non-contrôle du facteur B peut aboutir à la non-orthogonalité du croisement A*B. Mais, puisque B est ignoré, l'effet de A qu'on calcule revient à procéder par moyennage pondéré sur B, et la possible non-orthogonalité de A*B, peut conduire à conclure que A a peu d'effet, alors qu'il en a beaucoup pour chaque b ; ou au contraire conclure qu'il a un effet important, alors qu'il a un effet faible pour chaque b [8].

[7] On a implicitement considéré dans ce chapitre que les poids utilisés pour le calcul des effets pondérés proviennent seulement d'effectifs différents. Mais l'idée du calcul d'une moyenne ou d'un effet pondérés est plus générale et peut se faire à l'aide de poids définis extérieurement aux données : c'est ce qu'on fait lorsque, par exemple, on "redresse un échantillon".

[8] Cette remarque pourrait amener à douter de l'intérêt de toute expérimentation. Mais on y verra plutôt un moteur d'évolution du champ expérimental. Une expérience $E1$ révèle un effet de A ; suit l'expérience $E2$ qui ne confirme pas les résultats de $E1$, et l'analyse des facteurs non-contrôlés, intervenant avec des pondérations différentes en $E1$ et $E2$, suggère qu'un facteur B est peut-être responsable du désaccord ; l'expérience $E3$ est alors élaborée en incluant le facteur B, et amène à amender les résultats de $E1$ et $E2$, *etc.*.

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 61-70)

V. INTRODUCTION À LA NOTION DE COMPARAISON

RÉSUMÉ — *L'analyse d'un ensemble de données planifiées conduit à deux objets clés, qui correspondent respectivement à deux approches complémentaires : les "protocoles dérivés" (cf. chapitres II., III. et IV.) et les "comparaisons". Ce chapitre constitue une introduction au point de vue des comparaisons et explicite le lien entre les deux approches.*

SUMMARY — Introduction to the notion of comparison
The analysis of a planned data-set leads to two key objects, that correspond respectively to two complementary approaches : the "derived protocols" (cf. chapters II., III. and IV.) and the "comparisons". This chapter introduces the viewpoint of comparisons and provides links between the two approaches.

1. INTRODUCTION

Comme on l'a dit au chapitre I., l'analyse des comparaisons (ANACOMP) constitue une reconstruction complète de l'ANOVA, élaborée par Rouanet & Lépine (1976 et 1977), puis poursuivie par Lecoutre (1984 et 1991) et Rouanet & Le Roux (1993). Ce court chapitre ne saurait être un résumé exhaustif de cet imposant travail et nous renvoyons aux textes cités et aux références qui y sont incluses pour de nombreux compléments.

On aura remarqué que la notion de comparaison, bien qu'évoquée régulièrement, a été repoussée à ce chapitre tardif de ce texte. C'est justement un de nos "paris" que d'avoir exploité à fond la ligne des protocoles dérivés, en évitant le recours au point de vue des comparaisons. Pourtant, ces deux aspects sont en fait indissociables et c'est, du coup, une sorte de "tour de passe-passe" dont le lecteur a été victime.

N'y voyez aucune malice; les raisons de ce choix sont tout à fait avouables. Pour reprendre une métaphore déjà utilisée, une question posée aux données se traduit par : une comparaison, *i.e.* un *angle sous lequel on regarde les données* ; et un protocole dérivé, *i.e. ce qu'on voit des données sous cet angle*. En donnant à cette métaphore une "coloration" photographique, on peut dire que : le protocole de base constitue le "sujet du photographe" situé dans l'espace; une comparaison correspond à la fois à l'appareil photo avec son système de "lentilles", et à sa position dans l'espace par rapport au "sujet" ; et le protocole dérivé est le résultat, la "photographie". Dans notre initiation à la "noble discipline de la photographie", nous avons choisi de d'abord montrer des "photos" avant d'introduire la théorie et la technique de la "prise de vue".

L'approche des comparaisons est, on l'aura préssenti avec la métaphore précédente, plus fondamentale mais aussi plus abstraite que celle des protocoles dérivés. Elle repose sur une *formalisation linéaire* qui fait appel à des notions mathématiques d'*algèbre linéaire* (*e.g.* espaces vectoriels, métrique euclidienne) dans laquelle la notion essentielle est celle

de la *dualité* entre l'*espace des mesures* et l'*espace des protocoles* (cf. Rouanet, Le Roux, 1993, chap. II). Son caractère abstrait tient aussi à ce que toute une partie de son discours peut se faire indépendamment des données. Ici on adoptera, dans la mesure du possible, une approche plus intuitive que formelle en insistant sur les aspects géométriques (les fameuses "lentilles", en particulier) [1].

Le point de vue des comparaisons est une clé qui permet de retrouver et d'expliquer un certain nombre de résultats énoncés dans les chapitres précédents, principalement : la notion d'indépendance entre questions posées aux données, les problèmes liés au cas non-équipondéré, la justification des dérivations des poids. Une fois rappelés un certain nombre de notions de base sur les comparaisons (§2. et §3.), qui établiront ainsi progressivement des "jonctions" entre les deux points de vue, nous serons en mesure d'expliciter leur caractère complémentaire et indissociable (§4.).

2. MESURES ET PROTOCOLES SUR UN SUPPORT PONDÉRÉ

2.1. Mesures et contrastes sur un support pondéré

Prenons l'exemple d'un protocole univarié de support `U` de 3 unités, et calculons, sur ce protocole, les statistiques : moyenne pondérée "`Mean U`", moyenne équipondérée "`Emean U`", et les deux effets partiels correspondant aux différences "`Diff u1,u2`" et "`Diff u1_u2,u3`" (par la suite, on appellera "effet" chacune de ces statistiques). Le protocole, avec ses valeurs (variable x^U) et ses poids (pondération w_U), et ces divers effets sont donnés ci-après :

`Raw U`

	x^u	w_u
u1	$7/4 = 1.750$	**4**
u2	$2/3 = 0.667$	**9**
u3	$8/15 = 0.533$	**15**

Statistique	Détail du calcul de l'effet					Effet
`Mean U`	$4/28 \times 7/4$	+	$9/28 \times 2/3$	+	$15/28 \times 8/15$	0.750
`Emean U`	$1/3 \times 7/4$	+	$1/3 \times 2/3$	+	$1/3 \times 8/15$	0.983
`Diff u1,u2`	$1 \times 7/4$	+	$-1 \times 2/3$	+	$0 \times 8/15$	1.083
`Diff u1_u2,u3`	$4/13 \times 7/4$	+	$9/13 \times 2/3$	+	$-1 \times 8/15$	0.467

Le calcul de chacun de ces effets se fait en appliquant aux 3 valeurs de x^U une *famille de coefficients*, a_U : par exemple a_U =[4/28, 9/28, 15/28] pour l'effet "`Mean U`". Une telle famille constitue une *mesure* sur le support `U` [2]. Parmi toutes les mesures possibles, deux catégories de mesures jouent des rôles privilégiés :

- Une *mesure-moyenne*, ou simplement une *m-mesure*, $m_U = (m_u)_{u \in U}$ est telle que : $\forall u, m_u > 0$ et $\sum_u m_u = 1$. L'effet associé à une m-mesure est une certaine moyenne. Les mesures correspondant à "`Mean U`" et "`Emean U`" sont des m-mesures. Pour le moyennage pondéré "`Mean`", on parle de "w_U-moyennage".
- Une *mesure-contraste*, ou simplement un *contraste* $c_U = (c_u)_{u \in U}$ est telle que : $\sum_u c_u = 0$. Intuitivement, un contraste *oppose* les modalités affectées de coefficients $c_u > 0$, aux modalités affectées de coefficients $c_u < 0$. Les mesures associées à "`Diff u1,u2`" et "`Diff u1_u2,u3`" sont des contrastes.

On notera R_U l'ensemble de toutes les mesures sur le support `U`. Cet ensemble est un espace vectoriel de dimension `U`. Une mesure $a_U = (a_u)_{u \in U}$ est un vecteur de cet espace

[1] La métaphore "photographique", bien que nécessairement imparfaite comme toute métaphore, pourra cependant rester à l'esprit et jettera, j'espère, quelques "lumières" sur certains concepts abstraits.

[2] Une *mesure* sur `U`, $a_U = (a_u)_{u \in U}$, est une famille de coefficients $a_u \in R$, qui, lors d'un regroupement d'unités, se dérivent naturellement par sommation. Une mesure est repérée par l'utilisation d'*indices en bas*. En revanche, les *variables* notées avec *indices en haut*, x^U, se dérivent naturellement par moyennage.

et a_u sa coordonnée pour la dimension u. L'*effet* de la mesure a_U appliquée au protocole x^U est par définition : $\text{Eff}(a_U, x^U) = \sum_u a_u x^u$.

2.2. Comparaisons, g-comparaisons à 1, à plusieurs, degré(s) de liberté (ddl)

Si on multiplie tous les coefficients d'un contraste c_U par une constante $k \neq 0$, on obtient un contraste $c'_U = k \times c_U$ qui traduit la même opposition que c_U. La *comparaison à 1 ddl*, C, associée à c_U, est *cette opposition*, indépendante du *représentant* particulier choisi : c_U ou c'_U. La comparaison C à 1 ddl est ainsi l'ensemble des contrastes proportionnels à c_U ($C = \{k \times c_U\}_{k \in R}$) et constitue un sous-espace vectoriel de R_U de dimension 1, *i.e.* une droite vectorielle.

Une *comparaison à plusieurs ddl* est un sous-espace vectoriel de R_U engendré par *plusieurs contrastes linéairement indépendants*. Par exemple 2 contrastes non-proportionnels c_U et d_U engendrent une comparaison à 2 ddl, *i.e.* un plan vectoriel : $C = \{k \times c_U + k' \times d_U\}_{(k,k') \in R^2}$. En particulier, la comparaison globale de n modalités constitue un cas typique de comparaison à $(n-1)$ ddl.

Nous élargirons la notion de comparaison, sous l'appellation de "*g-comparaison*"("g" pour "généralisée"), à tout sous-espace vectoriel de R_U engendré par une ou plusieurs *mesures quelconques* (et non uniquement des contrastes) : une *g-comparaison à 1 ddl* est une droite vectorielle de R_U ; une *g-comparaison à 2 ddl*, un plan vectoriel ; *etc.*[3].

Le nombre de ddl, *i.e.* la dimension, d'une g-comparaison est aussi le nombre minimum de mesures nécessaires pour la représenter. En particulier, la *comparaison globale* sur U est engendrée par tous les contrastes sur U et peut être représentée par seulement $(U-1)$ contrastes : elle a $(U-1)$ ddl. La *g-comparaison globale* sur U est engendrée par toutes les mesures sur U et peut être représentée par seulement U mesures (*e.g.* $[1,0,0,\cdots]$, $[0,1,0,\cdots]$, *etc.*) : elle a U ddl.

2.3. Espace euclidien R_U des mesures sur un support pondéré U

Les notions qui viennent d'être introduites sont purement *vectorielles*, *i.e.* ne dépendent pas de l'échelle de chacune des dimensions u. En plus de cette structure vectorielle, le choix d'un *produit-scalaire* entre mesures munit l'espace des mesures R_U d'une *métrique euclidienne* ; celle-ci permet de donner un sens aux notions de *norme* d'une mesure (longueur d'un vecteur) et d'*angles* entre mesures. En notant $a_U.b_U$ le produit-scalaire de a_U et b_U, $\theta(a_U, b_U)$ l'angle entre a_U et b_U, et $\|a_U\|$ la norme de a_U, on a :

$$a_U.b_U = \sum_u (a_u b_u)/w_u \quad (1)$$

$$\|a_U\|^2 = a_U.a_U = \sum_u (a_u)^2/w_u \quad (2)$$

$$\cos(\theta(a_U, b_U)) = (a_U.b_U)/(\|a_U\| \times \|b_U\|) \quad (3)$$

On remarquera le rôle particulier que joue la pondération w_U dans la définition de cette métrique[4]. La métrique choisie consiste à solidariser les échelles des U dimensions à l'aide

[3]Les premiers textes sur l'ANACOMP privilégient les contrastes et les comparaisons qui généralisent la notion d'écart. Mais la construction élaborée autorise des mesures plus générales (voir Rouanet, Le Roux, 1993, définition de l'application-effet, p. 136) ; cette ligne a également été suivie par Lecoutre (1991). Les appellations "m-mesure", "g-comparaisons" sont spécifiques au texte présent.

[4]La pondération w_U est elle aussi une mesure (d'où les indices en bas) ; elle est qualifiée de "fondamentale" parce qu'elle sert à définir la métrique. Les coefficients d'une mesure représentent eux aussi des poids, mais pour éviter les confusions, on réserve le terme de "poids" aux w_u.

des poids w_U : chaque poids joue le rôle d'une "lentille" qui multiplie (par rapport à la métrique euclidienne élémentaire) l'échelle de chaque dimension par $1/\sqrt{w_u}$ [5] [6].

Une mesure a_U est *normée* si sa norme $\|a_U\|$ vaut 1, *normalisée* si $\sum_{u,a_u>0} a_u = 1$. Deux mesures a_U et b_U sont *orthogonales* si leur angle vaut 90°, *i.e.* si $a_U.b_U = 0$, ce qu'on note $a_U \perp b_U$. Par exemple, les mesures correspondant à "Mean U", "`Diff u1,u2`" et "`Diff u1_u2,u3`" sont deux-à-deux orthogonales. La notion d'orthogonalité s'étend aux g-comparaisons : deux g-comparaisons sont orthogonales si toute mesure de l'une est orthogonale à toute mesure de l'autre. De façon générale, la m-mesure "`Mean U`" est orthogonale à tout contraste, donc à toute comparaison, et par conséquent, à la comparaison globale sur `U`. La comparaison globale sur `U` se décompose elle-même de façon orthogonale, par exemple par l'ensemble des $(U-1)$ contrastes "`Diff u1,u2`", "`Diff u1_u2,u3`", "`Diff u1_u2_u3,u4`", *etc.*.

2.4. Espace euclidien des protocoles R^U sur un support pondéré `U`

Le protocole observé x^U constitue aussi un vecteur de `U` coordonnées (ses valeurs). On désigne par R^U l'ensemble des protocoles de support `U`. L'espace R^U est aussi un espace vectoriel, qu'on munit également d'une structure euclidienne en définissant produit-scalaire, norme et angle entre deux protocoles x^U et y^U comme suit :

$$x^U.y^U = \sum_u (x^u y^u) \times w_u \tag{4}$$

$$\|x^U\|^2 = x^U.x^U = \sum_u (x^u)^2 \times w_u \tag{5}$$

$$\cos(\theta(x^U, y^U)) = (x^U.y^U)/(\|x^U\| \times \|y^U\|) \tag{6}$$

On note que la pondération w_U intervient encore pour définir la métrique, mais de façon inverse pour R^U par rapport à R_U : dans le monde des protocoles, les mêmes "lentilles" sont utilisées, mais à l'envers, et chacune multiplie l'échelle de chaque dimension u par : $\sqrt{w_u}$. *De plus, le carré de norme $\|x^U\|^2$, n'est autre que la somme des carrés bruts (`Rss`) du protocole pondéré (x^U, w_U).*

2.5. Dualité des espaces R_U et R^U

A toute mesure a_U de R_U, on associe sa *w_U-densité* $a^U = a_U/w_U$, qui est un protocole de R^U. Réciproquement, à un protocole x^U de R^U, on associe la mesure $x_U = x^U \times w_U$, dont la w_U-densité est x^U [7]. Ces formules de passage d'un espace à l'autre s'accompagnent d'un transfert mutuel de toutes les propriétés vectorielles et euclidiennes : l'angle entre x^U et y^U dans R^U est l'angle entre x_U et y_U dans R_U ; la norme de x^U dans R^U est la norme de x_U dans R_U, *etc.*. Les deux espaces R^U et R_U sont en *dualité* : tout concept,

[5] Pour faire un parallèle avec notre espace de tous les jours, qui lui aussi est euclidien (tridimensionnel), tout se passe ici comme si les différentes dimensions de base (hauteur, largeur, longueur) avaient des coefficients différents lorsqu'il s'agit de calculer des distances. Ceci pourra sembler étrange à l'architecte, mais beaucoup moins à l'enseignant (l'habitude de comparer des élèves sur la base de notes globales obtenues à partir de matières diversement pondérées).

[6] Si le protocole est élémentaire ($w_u = 1$), la métrique est la métrique euclidienne élémentaire ; c'est aussi le cas (à un changement d'échelle près) si le protocole est équipondéré.

[7] Pour rendre plus intuitives ces notions, on pensera à l'exemple de Rouanet & Le Roux (1993, p. 19) : des pays (`U`) dont chacun est de superficie w_u et de population a_u ; si on regroupe des pays, les superficies et les populations s'ajoutent, mais les densités de population a_u/w_u se "w_u-moyennent".

toute statistique, définis dans R_U ont leur contrepartie dans R^U. Voici quelques exemples de sous-espaces équivalents par ce transfert réciproque [8] :

Espace des mesures R_U	Espace des protocoles R^U	Dimension (ddl)
g-comparaison globale	espace des protocoles	U
comparaison globale	espace des protocoles centrés	U-1
comparaison à 1 ddl	droite de l'espace des protocoles centrés	1
g-comparaison engendrée par **Mean**	espace des protocoles constants	1

On peut comprendre le lien entre les deux espaces ainsi : *a)* "comparer", c'est ne s'intéresser qu'aux écarts, et donc aux protocoles centrés ; *b)* "moyenner", c'est ne s'intéresser qu'à ce qui est commun, et donc aux protocoles constants.

En conséquence, on peut représenter toute mesure, et plus généralement toute g-comparaison, dans l'espace des protocoles R^U. En prenant comme origine le protocole 0^U (de coordonnées toutes nulles), l'espace vectoriel (composé de vecteurs) R^U est assimilable à un espace affin (composé de points) : ainsi, les protocoles de R^U sont représentés par des points, et le protocole observé par un point particulier ; dans R^U une g-comparaison à 1 ddl se représente par une droite ; à 2 ddl, par un plan, *etc.*.

Avec cette représentation simultanée des protocoles et des g-comparaisons, *interroger le protocole observé*, consiste à : *a)* exprimer la question par une g-comparaison (d'où une droite, un plan, *etc.* de R_U) ; *b)* représenter cette g-comparaison dans l'espace des protocoles (d'où la droite, le plan, *etc.* correspondant de R^U) ; *c)* en se plaçant désormais uniquement dans R^U projeter orthogonalement (au sens de (6)) le protocole observé sur cette droite, ce plan. Le protocole projeté obtenu constitue la *réponse à la question qu'exprime la g-comparaison* [9]. Le protocole projeté y^U de x^U, par une g-comparaison à 1 ddl représentée par la mesure a_U, ainsi que sa norme $\|y^U\|$ sont donnés par :

$$y^u = \frac{a_u}{w_u} \times \frac{\mathrm{Eff}(a_U, x^U)}{\|a_U\|^2} = \frac{a_u}{w_u} \times \frac{\sum_u a_u x^u}{\sum_u a_u^2/w_u} \tag{7}$$

$$\|y^U\|^2 = \frac{\mathrm{Eff}(a_U, x^U)^2}{\|a_U\|^2} = \frac{(\sum_u a_u x^u)^2}{\sum_u a_u^2/w_u} \tag{8}$$

2.6. Représentations graphiques de R^U

2.6.1. *Premier exemple*

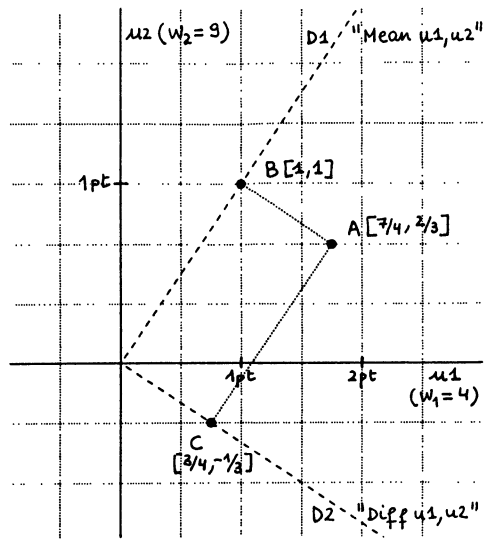
Considérons la restriction du protocole du §2.1. aux unités {u1,u2} : x^U=[7/4,2/3] pondéré par w_U=[4,9] (on notera toujours "U" le support qui n'a plus ici que deux unités). L'espace R^U est de dimension 2, soit un plan. Considérons les deux questions, "**Mean u1,u2**" et "**Diff u1,u2**", auxquelles correspondent dans l'espace R_U, respectivement les deux mesures orthogonales [4/13,9/13] et [1,-1] ; ou encore dans l'espace R^U respectivement la droite des protocoles constants, notée $D1$, et celle des protocoles centrés, notée $D2$ (du fait qu'il y a ici seulement deux unités ($U = 2$), l'espace des protocoles centrés est ici de dimension ($U - 1 = 1$), c'est-à-dire une droite) [10] :

[8] On rappelle qu'un protocole est w_U-centré si sa w_U-moyenne vaut *0*, et constant si ses valeurs sont toutes égales.

[9] Ceci justifie la métaphore photographique du §1. : le protocole projeté est ce qu'on voit du protocole de base si on regarde orthogonalement à la question (le "projecteur") qu'on pose.

[10] La mesure [4/13,9/13] de R_U engendre la droite de R_U des points $[4k/13, 9k/13]_{k\in R}$. La w_U-densité associée [1/13,1/13] de R^U engendre la droite $D1$ de R^U des points $[k/13, k/13]_{k\in R}$, *i.e.* des protocoles constants. De même, $D2$ est la droite des points $[k/4, -k/9]_{k\in R}$, *i.e.* des protocoles w_U-centrés.

Dans le graphique (1) de l'espace R^U, on a figuré : le protocole observé (point A), les deux droites $D1$ et $D2$, et les protocoles projetés correspondants (points B et C).
Les valeurs des protocoles se lisent sur les axes $u1$ et $u2$ (on considère qu'elles s'expriment dans l'unité physique "*pt*"). L'unité de mesure intrinsèque de l'espace euclidien R^U, *i.e.* celle qui sert à calculer des distances, est représentée par $1cm$. Mais en terme de l'unité "*pt*", les échelles de chaque axe sont différentes et s'obtiennent en *appliquant les "lentilles"* $\sqrt{w_u}$: l'échelle est de $2cm$ pour $1pt$ sur l'axe `u1` ($2 = \sqrt{4}$), et de $3cm$ pour $1pt$ sur l'axe `u2` ($3 = \sqrt{9}$).



(gr.1)

L'équation (7) donne les coordonnées des protocoles projetés B et C : on vérifie que B est constant (sa valeur commune pour $u1$ et $u2$, vaut 1, soit la moyenne pondérée "`Mean u1,u2`"), et que C est w_U-centré. (On verra que A, B et C sont assimilables aux protocoles "`U`", "`Z`" et "`U(Z)`".) Les normes des protocoles A, B et C représentent les distances des points à l'origine O (OA, OB, et OC) et s'obtiennent par (5) (ou encore par (8) pour OB et OC). Ces statistiques sont données dans le tableau ci-après.

Point	Projection sur	Protocole x^U ou projeté y^U	Norme $\|\|x^U\|\|$ ou $\|\|y^U\|\|$	$\|\|x^U\|\|^2$ ou $\|\|y^U\|\|^2$
A		$[7/4, 2/3]$	$OA = \sqrt{65/4} = 4.031$	$65/4 = 16.25$
B	`Mean u1,u2`	$[1, 1]$	$OB = \sqrt{13} = 3.606$	13
C	`Diff u1,u2`	$[3/4, -1/3]$	$OC = \sqrt{13/4} = 1.803$	$13/4 = 3.25$

Les deux mesures sont orthogonales, d'où l'angle droit entre $D1$ et $D2$, et, par le *théorème de Pythagore* : $OA^2 = OB^2 + OC^2$, *i.e.* $16.25 = 13 + 3.25$. Du point de vue des protocoles projetés, cette décomposition des carrés de normes n'est autre qu'une décomposition des `Rss`. C'est une première "jonction" entre le point de vue des comparaisons et celui des protocoles dérivés : *l'orthogonalité de plusieurs g-comparaisons, se traduit par une décomposition additive des* `Rss` *en termes de protocoles projetés.*

2.6.2. *Deuxième exemple*

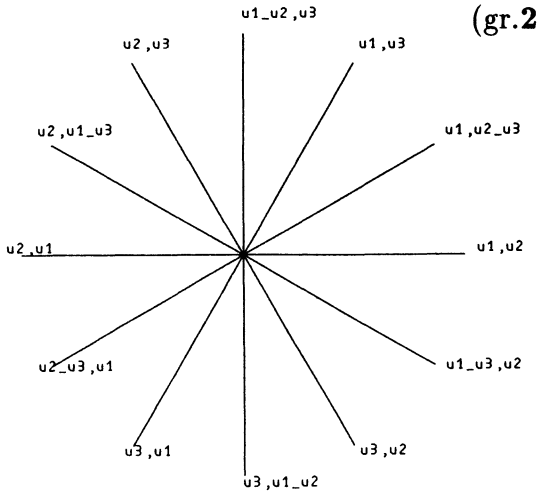
Revenons au support initial `U` de 3 unités, et plaçons nous dans le sous-espace de R^U associé à la comparaison globale sur `U`, *i.e.* l'espace des protocole centrés : il a $(U-1) = 2$ ddl et est donc un plan de R^U, noté P.

Supposons d'abord, contrairement à notre exemple, que le support `U` soit équipondéré. Le graphique (2) représente les droites associées aux contrastes du type : "`Diff u1,u2`", "`Diff u1,u2_u3`", *etc.*. Deux contrastes adjacents font entre eux un angle de 30°. On retrouve l'orthogonalité de contrastes tels que : "`Diff u1,u2`" et "`Diff u1_u2,u3`" (cf. chapitre III.§2.3 p.33).

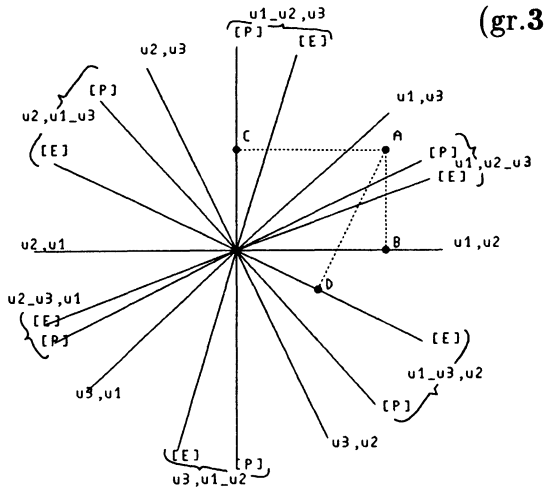
Revenons à l'exemple initial non-équipondéré ; dès qu'il s'agit de regrouper deux modalités, il y a maintenant deux possibilités : pondérée "`Diff u1_u2,u3 Mean`" (notée `[P]`), ou équipondérée "`Diff u1_u2,u3 Emean`" (notée `[E]`), d'où maintenant 9 droites, figurées dans le graphique (3). Dans ce graphique, on a également figuré plusieurs protocoles projetés orthogonalement de x^U : sur P (point A), et sur trois droites

particulières (points B, C et D) [11]. La solution pondérée "`Diff u1_u2,u3 Mean`" est encore orthogonale à "`Diff u1,u2`" [12]. Les deux solutions pondérée et équipondérée pour "`Diff u1_u2,u3`" sont intermédiaires entre "`Diff u1,u3`" et "`Diff u2,u3`" ; le poids $w_2 = 9$ est supérieur à $w_1 = 4$, et ainsi la solution pondérée fait jouer un rôle plus grand à x^{u2} qu'à x^{u1}, d'où son rapprochement avec le contraste "`Diff u2,u3`" ; comparativement, la solution équipondérée est plus proche de "`Diff u1,u3`".

(gr.2)



(gr.3)



A l'extrême, si le poids w_2 était énorme par rapport à w_1, l'angle entre "`Diff u1_u2,u3 Mean`" et "`Diff u2,u3`" serait presque nul. Les questions qu'expriment ces deux contrastes, et partant les réponses, seraient ainsi presque les mêmes. Cette remarque exprime, de façon intuitive, le fait suivant : un angle nul entre deux g-comparaisons correspond à des questions identiques, un angle non-nul à des questions différentes, et un angle droit à des questions indépendantes, d'où une deuxième "jonction" : *l'indépendance des questions se traduit par l'orthogonalité des g-comparaisons associées et par une additivité des* `Rss` *des protocoles projetés correspondants.*

2.7. Statistiques associées à une g-comparaison

Parmi les statistiques associées à une g-comparaison, certaines dépendent seulement de la g-comparaison, et non des mesures particulières choisies pour la représenter, *e.g.* l'effet ; certaines dépendent du protocole observé, d'autres non :

- Le *nombre de ddl*, *i.e.* la dimension, d'une g-comparaison ne dépend que de celle-ci.
- La *somme des carrés* associée à une g-comparaison C, appliquée au protocole pondéré (x^U, w_U) est le carré de la norme du protocole projeté y^U de x^U sur C, *i.e.* la `Rss` du protocole pondéré (y^U, w_U) ; elle ne dépend pas du choix des mesures pour représenter la g-comparaison. Mais, représenter une g-comparaison à n ddl par n mesures orthogonales, conduit à une simplification technique : sa somme des carrés s'obtient alors par l'addition des n sommes des carrés composantes. En particulier, la somme des carrés de la g-comparaison globale sur `U`, est la `Rss` de (x^U, w_U), *i.e.* "`Rss U`" ; celle de la comparaison globale sur `U` est la `Rss` du protocole centré associé à (x^U, w_U), *i.e.* "`Rss U(Z)`".

[11] On notera la propriété d'"hérédité" de la projection : les points B, C et D sont également les projections orthogonales de A sur les droites concernées.

[12] Le point A correspond au protocole centré associé à x^U de coordonnées [1,-0.083,-0.217] dans R^U, et de norme $\sqrt{4.767}$. A l'aide des équations (7) et (5) on calculerait coordonnées et normes des protocoles B, et C, et on retrouverait la décomposition additive des `Rss` : $OA^2 = OB^2 + OC^2$, *i.e.* $4.767 = 3.250 + 1.517$.

- A partir des statistiques précédentes et de U et $W = \sum_u w_u$, on définit, comme pour un protocole dérivé, les indices `Rms`, `Rvar`, `Rvarcor`, `Rsd` et `Rsdcor`, qui dépendent à la fois de la g-comparaison et du protocole observé.

2.8. Représentants privilégiés d'une g-comparaison

Au niveau méthodologique, les décompositions orthogonales sont privilégiées parce qu'elles signifient l'*indépendance entre elles des questions posées aux données* ; parmi toutes les décompositions orthogonales possibles, on choisit de préférence une décomposition qui comporte les questions d'intérêt. Le plus souvent, ainsi qu'on l'a fait au §1., une question sera exprimée par des mesures normalisées, dont l'effet a une interprétation directe : $[1,0,0]$ une valeur, $[1/2,1/2,0]$ une moyenne, $[1,-1/2,-1/2]$ une différence de moyennes, *etc.*.

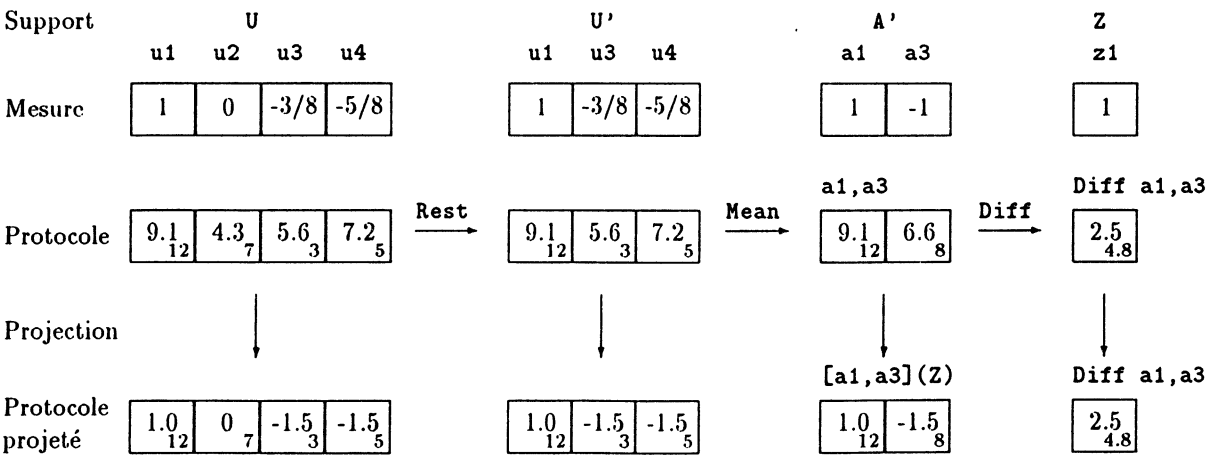
3. DE L'ESPACE DES PROTOCOLES R^U À L'ESPACE R^A

3.1. Dérivation et remontée

Considérons un protocole de support `U` emboîté dans un facteur `A` : `U<A>`. Souvent une question s'exprime naturellement au niveau du support dérivé `A`, et non au niveau de `U` lui-même. On dispose alors de deux objets "primitifs" : un protocole sur `U` et une g-comparaison sur `A`. Comment appliquer cette dernière au premier alors qu'ils ne sont pas définis sur le même support ?

La première approche consiste à construire le protocole dérivé "`A`" par moyennage sur `U`, puis à l'interroger par une g-comparaison sur `A` : c'est l'approche des *dérivations*. L'autre approche consiste à traduire la question initiale sur `A` par une g-comparaison sur `U` : c'est l'approche de la *remontée* (de `A` vers `U`). Des règles générales de dérivation et de remontée permettent d'assurer l'équivalence des deux approches (cf. 3.2.).

Nous ne ferons ici qu'illustrer ces notions à l'aide d'un exemple miniature : un protocole sur `U4`, décrit par le facteur `A3` (`u1<a1>`, `u2<a2>` et `[u3,u4]<a3>`), et la question "`Diff a1,a3 Mean`". Le diagramme ci-dessous montre comment cette même question peut être envisagée, de façon équivalente, à différents niveaux.



La partie supérieure du diagramme correspond à l'approche des comparaisons : *poser une question, i.e. une g-comparaison (ici une comparaison à 1 ddl), à un protocole*. La question s'exprime naturellement au niveau du support "`A'=a1,a3`", comme l'application du contraste [1,-1] sur les deux moyennes `a1` et `a3`. On passe d'un niveau à un autre

en *dérivant* (de gauche à droite) ou en *remontant* (de droite à gauche). La remontée de la question est ici décomposée en deux étapes : un "r-moyennage" (pondéré ici), *i.e.* la remontée d'un moyennage, et une "r-restriction", *i.e.* la remontée d'une restriction. La question peut également ici être dérivée vers le support `Z` par une dérivation par "différence".

La partie inférieure du diagramme correspond à l'approche des projections : *considérer la réponse à la question, i.e. un protocole projeté* (obtenu par (7)). Le protocole à l'extrême gauche correspond à la projection du protocole de base sur la comparaison (exprimés sur le support `U`). Pour chaque couple "comparaison-protocole" (en haut), on a un protocole projeté correspondant, défini sur un support de plus en plus petit lorsqu'on va de gauche à droite. A l'extrême droite, la projection est triviale.

3.2. Règles de dérivation/remontée pour **Mean**, **Emean** et **Diff**

t A chaque couple de "dérivation-remontée" sont associées des formules de transition pour les valeurs x^U, les poids w_U et les coefficients des mesures a_U. Ces formules assurent un passage "harmonieux" d'un support à un autre : conservation du produit scalaire, des angles et des normes, donc des orthogonalités et des sommes des carrés). En général, on a seulement recours à la dérivation des protocoles (valeurs et poids), et à la remontée des mesures.

Calculer une des statistiques **Mean**, **Emean**, ou **Diff** sur un protocole pondéré (x^U, a_U) consiste à lui appliquer une certaine mesure a_U, d'où l'effet $x' = \sum_u a_u x^u$ (dérivation du protocole). Pour toute statistique ainsi définie par l'application d'une mesure, la dérivation des poids est *canonique* : $w' = \sum_u (a_u^2/w_u)$. Ce choix assure l'identité pour les **Rss** entre, d'une part, l'application de a_U au protocole initial (x^U, w_U), et d'autre part, l'application de $a' = [1]$ au protocole dérivé d'une unité (x', w'). Ceci justifie les règles de dérivation des poids du chapitre II. (cf. tableau 2 p. 23).

Pour le moyennage pondéré (**Mean**), la dérivation des mesures se fait par *sommation*, et leur remontée par *w-répartition* (répartition des coefficients proportionnellement aux w_u).

Pour le moyennage équipondéré (**Emean**), la dérivation des mesures se fait par *sommation-harmonique*, leur remontée par *équi-répartition*.

3.3. Qu'entendre par "équivalence" ? Dans quel espace se situer ?

Les règles de dérivation/remontée qui précèdent permettent de calculer la `Rss` associée à la g-comparaison dans n'importe quel espace, et ce soit au niveau des couples "mesure-protocole", soit au niveau des protocoles projetés : dans le diagramme p.68, on vérifiera ainsi que la `Rss` vaut 30.0, à quelque niveau qu'on se situe. On peut ainsi parler de la "`Rss` associée à la question" sans avoir à préciser à quel niveau celle-ci été envisagée.

Mais ceci n'est pas vrai pour toute statistique associée à une question : le calcul de l'effet ne peut se faire qu'avec un couple "mesure-protocole" ou au niveau du protocole des différences. De même le point de vue des comparaisons est nécessaire pour le calcul des `Rdf`. Il en est de même pour la statistique `Rsdcor` caractérisant la grandeur de l'effet. Par contre, l'indice d'importance de l'effet `Ec`, défini comme le rapport des `Rsdcor` associées à deux questions (cf. III.§2.5), est défini de façon univoque *si on se place dans le même espace pour les deux questions.*

Pour juger de l'orthogonalité de plusieurs g-comparaisons, il faut se placer dans le même espace. Si deux g-comparaisons sont respectivement définies sur les supports `A` et `B`, il est nécessaire de remonter chacune sur un support commun : `A&B` au minimum.

4. COMPARAISONS ET PROTOCOLES DÉRIVÉS

4.1. Les deux points de vue

Le diagramme p. 68 permet de bien comprendre la différence entre les deux approches complémentaires, *comparaisons* et *protocoles dérivés*, ainsi que leurs connexions.

Du point de vue des comparaisons (en haut), on *pose une question aux données*, d'où toujours deux objets à considérer simultanément : une g-comparaison et un protocole. Par dérivation ou remontée, plusieurs couples de ce type sont envisageables et conduisent à la même réponse (pour les statistiques **Eff**, **Rdf** et **Rss**). A gauche, le protocole est simple (c'est le protocole de base), et la complexité de la question s'exprime dans la g-comparaison. A droite, la g-comparaison est simple, et la complexité de la question apparaît dans les dérivations qu'a subies le protocole. Comme on voit, cette approche fait déjà intervenir des protocoles dérivés (par des dérivations qui réduisent le support).

Du point de vue des projections (en bas), on considère seulement la *réponse à la question* : un protocole projeté qui intègre protocole de base et g-comparaison (c'est la fameuse "photographie"). Bien sûr il y a à nouveau plusieurs "photographies" selon le couple g-comparaison/protocole envisagé. Les projections sont aussi des dérivations, le prototype en étant le *centrage*. Ces dérivations ne changent pas le support.

On comprend ainsi pourquoi le seul point de vue des protocoles dérivés est toujours incomplet, lorsqu'on le dissocie de celui des comparaisons : dans un protocole projeté, on ne peut savoir si ce qu'on voit résulte des données elles-mêmes ou de l'angle de vue adopté. C'est aussi pour cela que le **Rdf** d'un protocole dérivé ne dépend ni des valeurs ni des poids, mais des dérivations dont il a été l'objet, qui, elles, traduisent la g-comparaison.

Enfin, on a noté l'équivalence entre *m-mesures* et *protocoles constants* et celle entre *comparaisons* et *protocoles centrés*. De ce fait, la décomposition additive, *i.e.* orthogonale, en termes de protocoles dérivés, de **A** en **Z** + **A(Z)** (cf. III.§ 2.3 p.33) est, en termes de comparaisons, celle de la g-comparaison globale sur **A** en m-mesure "**Mean A**" et comparaison globale sur **A**. De même, toutes les autres décompositions énoncées aux chapitres III. et IV. ont leurs "alter-ego" en termes de g-comparaisons.

4.2. Langage LID et comparaisons

Avec ce qui vient d'être dit, on comprend pourquoi le langage LID peut être envisagé des deux points de vue. Dans les premiers chapitres, on l'a d'abord présenté comme un *langage de protocoles dérivés* : une formule génère une succession de dérivations, dont éventuellement des projections. Mais, une formule du langage peut aussi servir à générer une g-comparaison à appliquer au protocole de base : le langage LID constitue donc également un *langage de demandes de g-comparaisons* Nous avons en fait constamment utilisé cette double signification des formules LID depuis le début de ce chapitre.

Cependant ce *langage de formules*, seul introduit jusqu'ici, ne permet d'exprimer que certaines g-comparaisons (*e.g.* les droites figurées dans le graphique 3) ; le langage LID plus général comporte également des *demandes par vecteur*, qui permettent de considérer n'importe quelles comparaisons ou m-mesure [13].

[13]Les demandes par formules incluent, dans les logiciels VAR3, PAC et EyeLID, certains contrastes standard intervenant dans la régression (contraste linéaire, *etc.*.). Les demandes par vecteur ne sont implémentées, à l'heure actuelle, que dans VAR3 et PAC.

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 71–80)

VI. INTRODUCTION À L'ANALYSE INDUCTIVE

RÉSUMÉ — *On présente d'abord la méthode standard d'analyse inductive pour les données expérimentales : l'analyse de la variance. On en rappelle les faiblesses en ce qui concerne la question de l'importance des effets. Dans la ligne de Rouanet (1994), on insiste sur l'avantage majeur des méthodes d'inférence bayésienne : intégrer description et induction.*

SUMMARY — Introduction to inductive analysis
We first present the standard inductive method for experimental data, the ANOVA. We recall its weaknesses as far as the question of the importance (or the "size") of the effect is concerned. Following Rouanet (1994), we stress the main advantage of the bayesian methods : integrate description and induction.

1. INTRODUCTION : DE LA DESCRIPTION À L'INDUCTION

Dans ce texte, nous avons choisi d'insister sur les aspects descriptifs de l'analyse des données planifiées. Ce chapitre en évoque certains des aspects inductifs. Pour le cas des données expérimentales, seul envisagé ici, la méthode inductive standard est l'analyse de la variance (ANOVA). On en trouvera une présentation, dans un cadre classique, dans *e.g.* Hays (1981) et Scheffé (1959). Nous nous plaçons ici dans le cadre, plus large, de l'ANACOMP développé dans Hoc (1983), Lecoutre (1984), Rouanet *et al.* (1991).

On rappellera d'abord les grandes lignes de l'ANOVA traditionnelle (tests F et t de Student) ainsi que ses insuffisances (§2.). Les méthodes d'inférence bayésienne, que nous présenterons ensuite (§3.), répondent à ces insuffisances, et, de ce fait, constituent un outil privilégié pour la *généralisation des conclusions descriptives*. Les points abordés dans le §3. sont discutés en détail dans Rouanet (1994). Dans le but de mettre en avant les principes de l'inférence, nous considérerons ici le seul cas de questions à 1 ddl.

Au chapitre III., on a énoncé deux idées-forces de l'ANACOMP : *a)* les objectifs du chercheur, exprimés par des *questions spécifiques*, doivent guider l'analyse ; *b)* les réponses à ces questions doivent, dans un premier temps, s'appuyer sur l'*étape descriptive*. La dernière idée-force, sur laquelle nous voudrions insister ici, consiste à *réconcilier, ou, mieux, intégrer les deux étapes de l'analyse, descriptive et inductive*, dont l'articulation est presque "démantelée" dans l'ANOVA traditionnelle.

1.1. Un exemple

Reprenons l'exemple du protocole "`S*O/c1m1->DEV`" dérivé du dossier "Négligence" déjà traité en III.§4.4. p. 45. A partir de ce protocole initial, de structure "`S12*O3`", on se pose

les questions suivantes : *a)* g-comparaison "Z", *i.e.* "la déviation moyenne diffère-t-elle de *0*?" ; *b)* comparaison "o1,o2", *i.e.* "la déviation moyenne est-t-elle différente entre les orientations o1 et o2 ?" [1].

Pour la question "Z", on a trouvé une déviation moyenne vers la droite $d_{obs} = 0.996cm$ correspondant à un écart-calibré à *0* de $ec_{obs} = +1.439$, d'où la conclusion descriptive d'écart à *0* (vers la droite) "*important*". Pour la question "o1,o2", on trouve une différence (en faveur de o1) $d_{obs} = 0.171cm$ correspondant à un effet-calibré $ec_{obs} = +0.085$, d'où la conclusion descriptive d'un effet "*faible*" [2].

Cependant, porter des conclusions sur les seuls sujets observés constitue rarement le seul but de l'analyse ; c'est souvent sur une population plus vaste de "sujets potentiels" — dont on considère les sujets observés comme un échantillon — qu'on veut pouvoir les porter. L'échantillon conduit à l'*effet observé* d_{obs}, et on cherche à se prononcer sur l'*effet "parent" ou "vrai"* δ dans la population : on dit que δ est le *paramètre de l'inférence.* (De même, à l'effet-calibré observé $ec_{obs} = \pm s_{eff}/s_{adj}$, correspond un effet-calibré parent $\pm\sigma_{eff}/\sigma_{adj}$.) L'effet observé est positif (ou important ou faible). Peut-on arriver à des conclusions analogues sur l'effet parent ? En d'autres termes, l'étape descriptive permet d'énoncer une *propriété observée* sur l'échantillon, et le but de l'étape inductive est de *généraliser (dans la mesure du possible) cette propriété à la population.*

1.2. Cadres de justification et d'interprétation (CJI) des procédures inductives

C'est la *structure statistique du protocole* qui fonde la possibilité d'inférence, puisque l'inférence consiste en une tentative de *généralisation sur le facteur de groupe* S. Mais, pour pouvoir généraliser, il faut que l'échantillon soit *représentatif* de la population. Pour les méthodes considérées ici, cette représentativité est assurée (de façon probabiliste) si on suppose que l'échantillon a été *extrait au hasard* de la population (Rouanet, Bernard, Le Roux, 1990, p. 197).

La notion de "structure statistique du protocole" s'élargit alors en celle de "*modèle-cadre*", qui comprend un ensemble d'hypothèses qui ne sont pas remises en cause lors de l'inférence, on parlera de "*présuppositions*" : parmi celles-ci, l'échantillonnage au hasard qui confère alors la statut de *facteur aléatoire* au facteur de groupe "S". Chaque méthode inductive repose sur un modèle-cadre particulier qui précise la portée de la généralisation qui est faite, et qui en constitue ainsi le *cadre de justification et d'interprétation (CJI)* (voir *e.g.* Rouanet *in* Rouanet *et al.*, 1991). Nous présenterons ici deux modèles-cadres, qui reposent tous deux, entre autres choses, sur la présupposition d'échantillonnage au hasard [3] :

- Le modèle-cadre *fréquentiste* conduit aux tests de signification et permet de se prononcer sur la *vraisemblance* d'hypothèses concernant le paramètre.
- Le modèle-cadre *bayésien* permet d'aboutir à des *énoncés probabilistes* concernant les valeurs possibles du paramètre.

---

[1] Rappelons que la comparaison "o1,o2", peut s'exprimer par le protocole centré "[o1,o2](Z)".

[2] L'*écart-calibré* ou *effet-calibré* "Ec" est un indice descriptif d'importance de l'écart ou de l'effet qui a été défini comme le rapport de deux "Rsdcor", qu'on peut noter en abrégé s_{eff}/s_{adj} (cf. III. eq. (6) p. 35) ; cet indice est donc toujours positif. Ici, pour un effet à 1 ddl, on lui adjoindra le sens de l'effet, d'où un indice signé : $ec_{obs} = \pm s_{eff}/s_{adj}$. Avec cet amendement, les conventions du chapitre III. deviennent : $ec_{obs} > 0.6$ ou $ec_{obs} < -0.6$ pour effet important, et $|ec_{obs}| < 0.4$ pour effet faible.

[3] Un autre modèle-cadre est celui de l'*inférence combinatoire* ou *ensembliste* qui conduit aux *tests non-paramétriques*. Ce CJI repose uniquement sur l'*échangeabilité* des modalités de "S", sans hypothèse d'échantillonnage au hasard (Rouanet, Bernard, Lecoutre, 1986 ; Rouanet, Bernard, Le Roux, 1990).

1.3. L'inférence spécifique

L'ANOVA est souvent présentée dans le cadre d'un *modèle général*, qui porte sur le protocole de base pris dans son ensemble et assure la validité de *toutes* les inférences réalisables. Mais ainsi, plus le plan du protocole de base est complexe, plus l'ensemble des présuppositions s'accroît, et devient alors irréaliste. On évite cet écueil en adoptant l'approche de l'*inférence spécifique*. Cette approche, dont nous avons présenté le pendant descriptif au chapitre III., est décrite dans Rouanet & Lecoutre (1983) et mise en oeuvre dans les logiciels VAR3 et PAC. Le principe en est le suivant : Pour une question spécifique d'intérêt, on ne considère, du protocole de base, que le protocole dérivé pertinent minimal (PDPM), *i.e.* celui qui comporte dans sa décomposition source d'intérêt et source adjointe minimale ; le modèle-cadre est posé au seul niveau du PDPM, et constitue ainsi un *modèle spécifique*, d'où un ensemble de présuppositions plus faible.

2. L'INFÉRENCE FRÉQUENTISTE : TESTS F ET t DE L'ANOVA

2.1. Principes généraux des tests de signification

Tout test de signification — et en particulier ceux intervenant dans l'ANOVA — peut être décrit selon un certain nombre d'étapes de raisonnement, que nous illustrons sur l'exemple de la comparaison "`o1,o2`". En adoptant l'approche spécifique, il suffit de considérer comme protocole pertinent, le protocole des différences individuelles d_s (= `Diff o1,o2/s`) de moyenne d_{obs}, estimation du paramètre δ. La population sur laquelle on infère est caractérisée par sa *distribution parente* de telles différences.

(i) Outre le caractère aléatoire du facteur "`S`", le modèle-cadre comprend des *hypothèses techniques* : ici, d'*indépendance* des observations relatives aux différents sujets, et de *normalité* de la distribution parente des différences.

(ii) A la question d'intérêt, "La différence parente est-elle positive ($\delta > 0$) ?", on associe une *hypothèse dite "nulle"*, et notée $H_0 : \delta = 0$. L'hypothèse H_0 représente un modèle (qu'on voudrait "valider" ou "infirmer" dans un sens donné) que le test met à l'épreuve.

(iii) On choisit une certaine statistique D (ici "`Diff o1,o2`") qui mesure l'écart (d_p) d'un protocole quelconque p à l'hypothèse H_0. Sur le protocole observé, cette statistique vaut d_{obs}, soit ici $d_{obs} = 0.171$ [4].

(iv) L'ensemble des hypothèses (présuppositions + H_0) définit un ensemble de protocoles (ici d'échantillons) possibles P, sur chacun desquels on pourrait aussi calculer la statistique D. En contruisant cet ensemble de protocoles, et en calculant pour chacun la valeur d_p, on obtient la *distribution d'échantillonnage (DE)* de la statistique D : ici la distribution de la moyenne de 12 valeurs extraites au hasard d'une distribution normale (présupposition) de moyenne *0* (H_0). Par l'hypothèse d'échantillonnage aléatoire, cette DE fournit une *probabilité* (sur P) à tout intervalle de valeurs de D.

(v) On *situe* le protocole observé par rapport à l'ensemble P, *i.e.* d_{obs} par rapport à la DE de D : on calcule d'abord les probabilités $p_{sup} = P(D > d_{obs})$ et $p_{inf} = P(D < d_{obs})$; puis, on définit le *seuil observé unilatéral*, $p_{obs} = min(p_{inf}, p_{sup})$; ainsi que le *seuil observé*

[4] Plusieurs statistiques peuvent être envisagées pour mesurer l'écart à H_0. Pour la statistique "`Diff o1,o2`", un fort écart à H_0 correspond à des valeurs soit fortement positives, soit fortement négatives : cette statistique est *orientée*. Pour les statistiques *non-orientées*, un fort écart à H_0 correspond seulement à des valeurs fortement positives. Pour une question à 1 ddl, on a toujours le choix entre les deux types de statistiques ; pour plusieurs ddl, les statistiques envisageables sont non-orientées.

bilatéral par $p_{bil} = 2 \times p_{obs}$ [5]. Deux cas peuvent alors se présenter :
Si p_{obs} est petit, le protocole observé est extrême et fait partie des échantillons fortement improbables. Deux conclusions sont alors possibles : ou bien les hypothèses sont vraies et on a observé un "cas rare", ou bien les hypothèses sont fausses. La *logique du probable* sur laquelle s'appuient les tests conduit à choisir la seconde alternative, et puisqu'on ne remet pas en cause le modèle-cadre, c'est l'hypothèse H_0 qu'on doit alors *rejeter*.
Si, à l'opposé, p_{obs} est grand, le protocole observé est très central, et il n'y a pas lieu de mettre en cause les hypothèses, dont H_0 : H_0 est dite *compatible* avec les données.

(vi) Plutôt que de trancher radicalement entre "rejet" et "non-rejet" de H_0, on préfère situer p_{obs} par rapport à une grille de seuils-repères unilatéraux, notés $\alpha/2$ (ou encore, situer p_{bil} par rapport à α). Par convention, on utilise usuellement la grille suivante pour α : 0.05, 0.01 et 0.001. Dans le *test standard*, les conclusions possibles sont les suivantes : écart à H_0 *non-significatif* au seuil bilatéral 0.05 ; ou écart *significatif* (dans le sens de l'effet observé) au seuil unilatéral 0.05/2, 0.01/2 ou 0.001/2 [6].

2.2. Test F et tableau d'analyse de la variance

L'usage en ANOVA a conduit à privilégier les statistiques F (non-orientée) et t (orientée), plutôt que la statistique "Différence" D. Mais le raisonnement n'est pas affecté : au lieu de situer d_{obs} par rapport à la distribution de D, on situe la statistique F_{obs} par rapport à une distribution du F.

A toute source de variation d'intérêt, "eff", on associe une source adjointe, "adj", qui sert de terme de référence pour calculer l'effet-calibré (cf. III. eq. (6) p. 35). Mais elle sert également de terme de référence pour l'inférence. La statistique F est définie comme le *rapport des carrés-moyens (bruts) de ces deux sources de variation* [7] :

$$F_{obs} = \texttt{Rms}_{eff}/\texttt{Rms}_{adj} \qquad (1)$$

Sous l'hypothèse H_0 et les présuppositions du modèle-cadre, la DE de la statistique F est distribuée selon un F *de Fisher-Snedecor* à $[\texttt{Rdf}_{eff}, \texttt{Rdf}_{adj}]$ ddl.

Pour l'effet "`[o1,o2](Z)`", la source adjointe est "`S.o1,o2`", et on trouve $F_{obs} = 0.087$ à situer par rapport à un F à [1, 11] ddl. On trouve $p_{bil} = P(F > F_{obs}) = 0.773$, soit $p_{obs} = 0.387$. Puisque $p_{bil} > 0.05$, le résultat est déclaré "non-significatif au seuil-repère bilatéral $\alpha = 0.05$" et on conclut que H_0 est compatible avec les données. On procède de même pour l'effet "`Z`", auquel est associé la source adjointe "`S(Z)`" : $F_{obs} = 24.86$ avec [1, 11] ddl, d'où $p_{obs} = 0.0002$. Puisque $p_{obs} < 0.001/2$, le résultat est déclaré "significatif (dans le sens de l'effet observé) au seuil-repère unilatéral $\alpha/2 = 0.001/2$". On rejette l'hypothèse nulle H_0, et on conclut à l'existence d'une déviation moyenne parente vers la droite ($\delta > 0$). Tous ces résultats sont consignés dans le *tableau d'ANOVA* :

Source de variation	Rss	Rdf	Rms	F_{obs}	p_{obs}	p_{bil}	Conclusion
`Z`	35.701	1	35.701	24.86	0.0002	0.0004	S. à 0.0005 (unil.)
`S(Z)`	15.799	11	1.436				
`[o1,o2](Z)`	0.175	1	0.175	0.087	0.387	0.773	NS. à 0.05 (bil.)
`S.o1,o2`	22.061	11	2.006				

[5] Ceci ne vaut que parce que la statistique D choisie est "orientée". Pour une statistique "non-orientée", on définirait $p_{bil} = p_{sup}$ (et également $p_{obs} = p_{bil}/2$ pour une question à 1 ddl).

[6] On notera la dissymétrie des conclusions : le rejet de H_0 est une conclusion orientée (on rejette H_0 dans le sens indiqué par l'effet observé), alors que la compatibilité est non-orientée (on n'a pas pu rejeter H_0 ni dans un sens, ni dans l'autre).

[7] Pour un effet d'intérêt donné, il peut y avoir une ou plusieurs sources adjointes, selon qu'on adopte le modèle général ou un modèle plus ou moins spécifique, d'où les divers tests F proposés par VAR3. Adopter l'approche spécifique revient à prendre la source adjointe minimale, *i.e.* le test F'_1 de VAR3.

2.3. Lien entre test F et test t de Student

Lorsque la source de variation d'intérêt a 1 ddl, le test F précédent est équivalent à un test du t *de Student*, en vertu des égalités : $t^2_{obs} = F_{obs}$ et $t^2_{[\nu]} = F_{[1,\nu]}$, où $t_{[\nu]}$ désigne la distribution du t de Student à $[\nu]$ ddl.

2.4. Conditions de validité des tests F et t

Revenons sur les présuppositions du modèle-cadre. Quelles présuppositions sont mises en jeu pour une question et un protocole de base quelconque ? En quoi ces présuppositions sont-elles restrictives en ce qui concerne la validité de l'inférence ?

2.4.1. *Echantillonnage au hasard : par construction ou par hypothèse ?*

Les deux méthodes inductives décrites dans ce chapitre se placent dans le cadre de l'échantillonnage aléatoire. Dans la pratique, cependant, cette présupposition est souvent irréaliste. On est en fait souvent dans la situation où, disposant de données, on cherche une population dont les données *pourraient avoir été* extraites au hasard, et la généralisation réalisée lors de l'étape inductive porte alors sur cette population. Mais il suffira d'avoir oublié une caractéristique particulière de l'échantillon pour que l'inférence ait en fait une portée réduite. A cet égard, l'examen des facteurs constants du protocole est essentielle.

2.4.2. *Hypothèses techniques*

Les tests F et t reposent sur diverses hypothèses techniques, concernant la population de sujets (ou les populations lorsque `S` est emboîté dans des groupes) :

- Indépendance des observations individuelles.
- Normalité de la (ou des) distribution(s) parente(s).
- Homogénéité des variances des distributions parentes, lorsqu'on compare plusieurs groupes de sujets (*e.g.* question "`G(Z)`" dans le cadre de la structure `S<G>`).
- Conditions "de circularité" pour une comparaison à plusieurs ddl sur des traitements (*e.g.* question "`T(Z)`", pour un protocole "`S*T3`") (voir Rouanet, Lépine, 1970).

Après un tel déballage d'hypothèses, plus d'un chercheur pourrait être tenté de ranger au placard son livre préféré de Statistique inductive et de se retourner vers la seule approche descriptive. Le point est que, dans beaucoup de contextes expérimentaux — et pas seulement en Psychologie —, toutes ces présuppositions, en tout cas prises au pied de la lettre, sont vraisemblablement toujours *fausses*.

Heureusement, celles-ci ne sont pas toutes cruciales pour la validité des tests F et t. Ces tests sont extrêmement *robustes* à la non-normalité des distributions parentes, même dans les cas extrêmes de variables structurellement non-normales (*e.g.* temps de réaction, réponse en $[0, 1]$), dès que les effectifs ne sont pas trop petits et pas trop déséquilibrés. Les hypothèses d'homogénéité et de circularité sont plus cruciales, mais, lorsqu'elles sont suspectées, il existe des solutions de rechange qui n'y font pas appel (voir Lecoutre (1991) et le logiciel PAC). Enfin, les hypothèses d'indépendance sont cruciales, mais en général assurées grâce aux précautions expérimentales.

Mais surtout, l'approche spécifique réduit considérablement le poids de ces présuppositions, puisque celles-ci ne portent plus alors que sur le seul protocole dérivé pertinent. En particulier, la condition de circularité est automatiquement remplie pour toute comparaison à 1 ddl.

2.5. Insuffisance des tests de signification

Des expressions générales de la statistique F_{obs} (1) et de l'effet-calibré ec_{obs} (cf. III. eq. (6)), on tire la relation :

$$F_{obs} = ec_{obs}^2 \times \frac{N_{eff} W_{adj}}{N_{adj} W_{eff}} = ec_{obs}^2 \times \hat{n} \ , \qquad (2)$$

où N et W représentent respectivement nombre d'unités et poids total du protocole dérivé concerné. La quantité $\hat{n}$ est assimilable à un nombre de sujets. Pour une question à 1 ddl, cette relation peut aussi s'exprimer, $t_{obs} = ec_{obs} \times \sqrt{\hat{n}}$.

Le résultat du test F est ainsi fonction de deux choses : l'importance descriptive de l'effet (ec_{obs}), et le potentiel inductif des données ($\hat{n}$). Si l'effet parent est petit sans être nul, on peut parvenir à un test significatif, pour autant qu'on recueille suffisamment de données. A l'inverse, on peut obtenir un test non-significatif, avec un effet important mais peu de données. En fait, le test F est conçu pour détecter *n'importe quel* écart à H_0, aussi petit soit-il : il ne répond qu'à la question "$ec_{obs} > 0$?" soit "$\delta > 0$?".

Le problème est que, en général, le modèle que représente H_0, *e.g.* $\delta = 0$, n'est en général qu'approché (il y a peu de facteurs dont on puisse penser qu'ils n'ont *strictement* aucun effet). Dans ce cas, l'expérimentateur qui recueille beaucoup d'observations aboutira irrémédiablement, au bout du compte, à rejeter H_0 [8]. Plutôt que de mettre à l'épreuve un modèle *ponctuel*, la question est en fait plus souvent de montrer :

- soit que le modèle est *à peu près* vrai, *i.e.* que l'écart au modèle est *négligeable* ; cette conclusion généralisera la conclusion descriptive d'effet faible ;
- soit que le modèle est largement faux, *i.e.* que l'écart au modèle est *notable* ; cette conclusion généralisera la conclusion descriptive d'effet important.

Le test de signification ne répond à aucune de ces deux questions : significatif ne signifie pas notable, non-significatif ne signifie pas négligeable. Ces insuffisances des tests, souvent soulignées, amènent à se tourner vers les méthodes bayésiennes pour lesquelles, on va le voir, ces problèmes sont sans objet.

3. L'INFÉRENCE BAYÉSIENNE

3.1. L'approche bayésienne de l'induction

L'approche bayésienne peut être décrite comme un *modèle d'apprentissage* (du chercheur face à la réalité) : l'*état de connaissance initial* sur le paramètre, décrit par une *distribution initiale*, est révisé grâce aux données, en un *état de connaissance final*, décrit par la *distribution finale*. Dans l'approche bayésienne *non-informative*, la distribution initiale exprime un état initial d'*ignorance* et l'état final traduit alors l'apport propre aux données (voir *e.g.* Jeffreys, 1961 ; Box, Tiao, 1973). On se placera ici uniquement dans ce cadre et la distribution finale sera alors appelée "*distribution standard*" [9].

Dans l'approche bayésienne, le paramètre est considéré comme une variable, et l'incertitude à son sujet est exprimée par une distribution de probabilité. Mais ces

[8] Si on voulait forcer le trait à l'extrême, on pourrait dire que le test de H_0 répond à une question *sans intérêt* : H_0 est *toujours* fausse, et il suffirait de recueillir suffisamment de données pour le montrer !

[9] Plusieurs solutions peuvent être envisagées pour répondre à cette motivation "non-informative". On adoptera ici la solution de Lecoutre (1984 et 1991). Les choix qui subsistent reflètent la difficulté de formaliser ce qu'est l'"ignorance", mais ne sont pas inhérents à l'approche bayésienne (voir *e.g.* Bernard, 1993). Par la suite, on s'autorise à simplifier en parlant de "la" distribution standard.

probabilités sont bien différentes de celles du cadre fréquentiste. Alors que les probabilités de l'inférence fréquentiste vont de l'inconnu — on fait une hypothèse sur le paramètre — vers le connu — on probabilise l'ensemble des échantillons qu'on aurait pu obtenir, les probabilités de l'inférence bayésienne sont naturelles : elles vont du connu — les données observées — vers l'inconnu — le paramètre.

Hormis la distribution initiale, le modèle-cadre de l'inférence bayésienne repose sur les mêmes présuppositions que l'inférence fréquentiste ; à tout test F ou t valide, général ou spécifique, correspond une solution bayésienne aux conditions de validité identiques.

3.2. Quelques distributions standard

Pour une question à 1 ddl, considérons les variables d'intérêt suivantes : l'effet parent δ (dont l'effet observé d_{obs} est l'estimation), et la variable effet-calibré $ec_{par} = \pm\sigma_{eff}/s_{adj}$ (à laquelle correspond l'effet-calibré observé $ec_{obs} = \pm s_{eff}/s_{adj}$) [10]. Les distributions standard des variables δ et de ec_{par} se déduisent de la distribution du t élémentaire à ν ddl, t_ν (on note $x \sim t_\nu(a, b^2)$ pour indiquer que $\frac{x-a}{b} \sim t_\nu$) :

$$\delta \sim t_\nu(d_{obs},\ d^2_{obs}/F_{obs}) \tag{3}$$

$$ec_{par} \sim t_\nu(ec_{obs},\ 1/\hat{n})\ , \tag{4}$$

où ν désigne le ddl brut de la source adjointe : "$\mathtt{Rdf}_{adj}$".

La figure 1 donne les distributions standard de δ pour les deux questions "Z" et "o1,o2" ; les distributions standard de ec_{par} apparaissent à la figure 2. Chaque distribution est centrée sur l'effet observé correspondant. La dispersion de chacune de ces distributions exprime l'incertitude relative à chaque variable concernée, qui dépend des variations inter-individuelles et de $\hat{n}$ en ce qui concerne δ, et de $\hat{n}$ seulement en ce qui concerne ec_{par}.

A partir de la distribution standard, la réponse au problème de l'induction est directe. L'effet observé pour "Z" est important : $ec_{obs} > 0.6$. Cette propriété est-elle également vraie pour ec_{par} avec une probabilité, une "*garantie*", suffisante ? Autrement dit, la probabilité "$Prob(ec_{par} > 0.6)$" est-elle suffisamment élevée ? De même, pour "o1,o2", on cherchera à généraliser la propriété $|ec_{obs}| < 0.4$ sur la base de la probabilité "$Prob(|ec_{par}| < 0.4)$". Par convention, on considère, comme pour les tests, une grille de *garanties-repères* notées génériquement γ : 0.95, 0.99, 0.999.

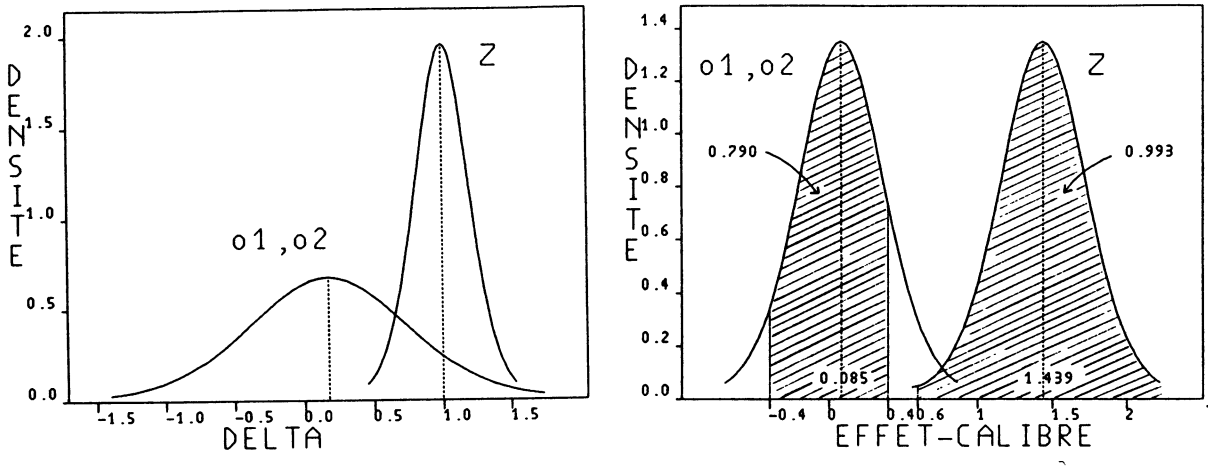
Pour "Z", on trouve : $Prob(ec_{par} > 0.6) = 0.993$. On peut généraliser la conclusion descriptive d'effet important, avec la garantie 0.993, d'où une *conclusion d'écart notable*. Pour "o1,o2", on trouve : $Prob(|ec_{par}| < 0.4) = 0.790$. La conclusion descriptive d'effet faible ne peut être généralisée avec une garantie suffisante : on ne peut parvenir à une *conclusion d'effet négligeable*.

Comme on voit, il n'est pas possible de généraliser n'importe quelle propriété observée. Si pour la question Z, on cherche à généraliser la propriété d'"écart énorme" $ec_{obs} > 1.4$, on trouve une faible garantie (0.553) du fait que la valeur 1.4 est proche de d_{obs} et le nombre de sujets petit : la propriété est de forte *teneur* ($ec_{par} > 1.4$) mais de faible *portée* (garantie 0.553). En assouplissant la teneur de la propriété, *e.g.* ($ec_{par} > 1.0$) ou ($ec_{par} > 0.6$), on en augmente la portée : 0.922 ou 0.993. Pour parvenir à une garantie suffisamment élevée, il y a une "taxe" à payer sur la teneur, taxe d'autant plus élevée

[10] C'est pour simplifier l'exposé qu'on considère la variable $ec_{par} = \pm\sigma_{eff}/s_{adj}$ (assimilée à un paramètre), au lieu de l'effet-calibré parent $\pm\sigma_{eff}/\sigma_{adj}$ (le véritable paramètre) dont la distribution standard n'est pas une distribution élémentaire ; ce faisant, on néglige l'incertitude relative à la grandeur de la source adjointe.

que l'échantillon est petit. Si l'échantillon est trop petit, la taxe peut même dépasser le capital initial et on ne pourra conclure ni dans un sens ni dans l'autre.

Distribution standard de δ (gr.1) Distribution standard de ec_{par} (gr.2)



3.3. Approche fréquentiste et approche bayésienne standard

L'inférence bayésienne standard procure des réinterprétations aux procédures fréquentistes [11]. Pour les inférences envisagées ici, ce lien est assuré par les deux relations : $p_{obs} = Prob(\delta < 0) = Prob(ec_{par} < 0)$, et $1 - p_{bil} = Prob(0 < \delta < 2d_{obs}) = Prob(0 < ec_{par} < 2ec_{obs})$. Ces relations permettent de traduire, bayésiennement, les conclusions "significatif" et "non-significatif".

Un résultat est significatif quand p_{obs} est petit, *e.g.* la question "Z" avec $p_{obs} = 0.0002$. Bayésiennement, ceci signifie que $Prob(ec_{par} > 0) = 1 - p_{obs} = 0.9998$ est grand : on peut dire, avec une bonne garantie, que l'effet va dans le même sens que l'effet observé, *i.e.*"l'*existence* d'un effet est *bien* établie", énoncé de *teneur minimale* avec une *grande* portée. La recherche de conclusion d'effet notable apparaît alors comme une extension de la conclusion de résultat significatif : on augmente la teneur de la conclusion en en diminuant la portée.

Un résultat est non-significatif lorsque p_{obs} n'est pas assez petit : *e.g.* la question "o1,o2" avec $p_{obs} = 0.387$. Du coup, aucun des énoncés, $Prob(ec_{par} < 0) = p_{obs} = 0.387$ et $Prob(ec_{par} > 0) = 1 - p_{obs} = 0.613$, n'a une garantie importante, pas plus que l'énoncé $Prob(0 < ec_{par} < 2ec_{obs}) = 1 - p_{bil} = 0.227$. "Non-significatif" apparaît ainsi comme un simple constat d'ignorance ; dans le cas extrême d'un effet observé nul, le dernier énoncé devient $Prob(0 < \delta < 0) = 0$, dont on percevra aisément le caractère peu (sic) informatif.

En approchant t_ν par une distribution normale, les conclusions du test t et les énoncés bayésiens ne dépendent plus que de ec_{obs} et $\hat{n}$, d'où les figures 3 et 4. La figure 3 correspond à l'approche bayésienne : pour de petites valeurs de $\hat{n}$, le potentiel inductif des données est faible et aucune conclusion "positive" ne peut être atteinte ; si $\hat{n}$ est suffisamment grand et l'effet-calibré ec_{obs} important (resp. faible), on parvient à une conclusion d'effet notable (resp. négligeable) [12]. Au contraire, le test peut être significatif ou non-significatif que l'effet-calibré observé soit faible ou important (figure 4).

En superposant les deux graphiques, deux faits ressortent :

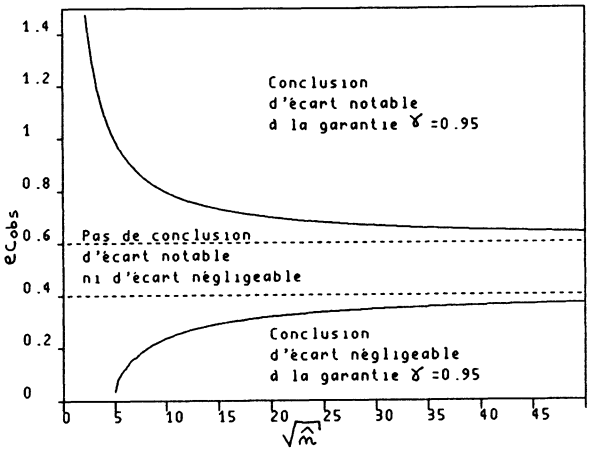
[11] La réciproque n'est pas vraie ! Certains énoncés bayésiens utiles n'ont pas de réinterprétation fréquentiste. De ce point de vue, on peut dire que l'approche bayésienne domine l'approche fréquentiste.

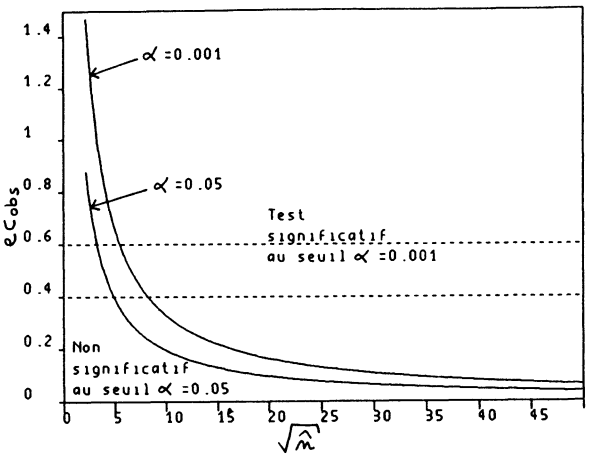
[12] On pourrait aussi ajouter à la figure 3 une région d'écart intermédiaire : $0.4 < ec_{par} < 0.6$.

- On ne peut espérer pouvoir conclure à un écart notable que si *(i)* l'écart-calibré observé est important, et *(ii)* le test est significatif (à condition de choisir une garantie et un seuil compatibles) ; ces conditions sont nécessaires mais non suffisantes ;
- Pour pouvoir conclure à un écart négligeable, il est nécessaire que l'écart-calibré soit faible ; mais la conclusion "non-significatif" n'est ni nécessaire, ni suffisante, pour cela.

En bref, le test peut être un premier pas pour conclure "notable", mais n'est pas pertinent pour conclure "négligeable".

Notable/Négligeable (gr.3) Significatif/Non-significatif (gr.4)





3.4. Mise en oeuvre pratique de l'inférence bayésienne

Les distributions et énoncés bayésiens précédents s'obtiennent à l'aide du logiciel PIF (Poitevineau, Lecoutre, 1986), à partir du seul tableau d'analyse de la variance. Le logiciel PAC (Lecoutre, Poitevineau, 1992) permet la mise en oeuvre de l'inférence bayésienne sur divers paramètres (dont la grandeur de l'effet et l'effet-calibré) pour des comparaisons à plusieurs ddl. Mentionnons aussi le logiciel IBF (Bernard, 1986) pour l'inférence bayésienne sur des données catégorisées.

4. REMARQUES CONCLUSIVES

A l'égal de beaucoup d'autres domaines scientifiques, la Statistique n'est ni une science pure, ni une science achevée. Son évolution est, comme ailleurs, le fruit d'un complexe mélange : de considérations scientifiques certes, mais aussi de contraintes techniques, de phénomènes sociologiques, historiques, et parfois individuels. Tout cela peut sembler banal, mais la prise de conscience de certains de ces moteurs extra-scientifiques d'évolution permet de mieux comprendre le décalage actuel entre les outils disponibles, les pratiques des utilisateurs, et les objectifs réels de l'analyse statistique.

4.1. Le passé

Une conception physiciste de la réalité : Les méthodes inductives, telle que l'ANOVA, ont d'abord été développées dans le contexte des expérimentations agronomiques, avec une *conception "physiciste" de la réalité* : la réalité suit un modèle, les écarts à ce modèle constituent des erreurs de mesure, d'où le fait qu'on s'intéresse surtout à des moyennes de façon à "gommer" ces erreurs. Ainsi, dans plus d'un ouvrage sur l'ANOVA, les sources adjointes sont qualifiées d'"*error term*". Le plaquage de cette conception sur

l'étude des phénomènes psychologiques aboutit à ceci : les différences inter-individuelles sont assimilées à des erreurs de mesure ; il y a de quoi se faire dresser les cheveux sur la tête à plus d'un psychologue différentialiste !

"Décider" ou "Analyser des données" ? : La théorie des tests s'est placée dans une perspective décisionniste, en assignant à l'analyse inductive le rôle suivant : "rejeter" ou "accepter" le modèle. Hors de certains domaines (*e.g.* recherche médicale), l'objectif réel de l'activité de recherche n'est, en général, pas de décider, mais, plutôt, d'apprendre sur la réalité par l'expérience.

L'ère des tables : Les contraintes techniques ont également joué un rôle déterminant dans le développement des outils statistiques. La théorie statistique de l'ANOVA s'est en effet développée avant l'ère informatique avec le problème majeur suivant : toute distribution nécessaire devait être calculée à la main puis tabulée. D'où la recherche de simplifications techniques, permettant de traiter des problèmes variés à l'aide d'un jeu restreint de distributions et de tables. Les créateurs de ces méthodes ont ainsi privilégié des statistiques *moins naturelles*, mais qui possédaient de *bonnes propriétés mathématiques* : *e.g.* les statistiques F et t dans le cadre de l'ANOVA.

Splendeur de la description, et misère de l'induction : Enfin il y a eu ces dernières années un essor considérable des méthodes descriptives, grandement stimulé par les progrès de l'informatique. Ces méthodes se sont développées dans des directions relativement naturelles qui n'ont pas, en général, été suivies des développements inductifs correspondants.

4.2. Le présent et le futur

Le présent : Le bilan de tout cela est que la panoplie usuelle des outils inductifs est mal assortie aux outils descriptifs : on utilise des méthodes faisant appel à des statistiques non-naturelles, répondant à des questions non-pertinentes, ou, en tout cas, de teneur bien limitée par rapport aux conclusions désirées. D'où la tentation d'utiliser ces outils inductifs en les détournant de leur cadre initial, pour les amener à étayer des conclusions plus signifiantes. On aboutit à toutes sortes d'interprétations fausses, dont les deux principales : "Le F est très significatif, donc l'effet est important", "Le F est non-significatif, donc il n'y a pas d'effet".

Le futur : On est désormais à un tournant dans l'évolution des méthodes statistiques. On le voit par l'insistance de plus en plus fréquente sur les insuffisances des méthodes usuelles et l'intérêt croissant pour le problème de l'importance des effets (voir *e.g.* Rouanet, 1994, et les références qui y sont citées). De plus l'évolution considérable des moyens informatiques rend désormais accessibles (ou en voie de l'être) des méthodes, aux objectifs ambitieux, mais naguère inenvisageables sur le plan pratique, et en premier lieu les méthodes bayésiennes. En voici quelques exemples seulement : la possibilité de se débarrasser des présuppositions techniques ; des méthodes inductives plus naturelles, "... de l'inférence sans Khi^2, sans t et sans F", comme le proposent Corroyer & Bert (1990) ; des méthodes permettant de généraliser des conclusions complexes (Bernard, 1991) ; et la voie de l'inférence prédictive qu'offre l'approche bayésienne.

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 81–96)

VII. ANALYSE DE DONNÉES PLANIFIÉES PAR LE LOGICIEL EyeLID

RÉSUMÉ — *Ce dernier chapitre traite des aspects pratiques. On présente d'abord brièvement le logiciel* EyeLID. *Des exemples de données expérimentales sont ensuite analysés et commentés.*

SUMMARY — Analysis of planned data with EyeLID
This last chapter deals with some practical aspects. We first briefly present the EyeLID *software. We then analyse in details several experimental data sets.*

1. INTRODUCTION

Ce dernier chapitre aborde la pratique de l'analyse des données planifiées, en insistant surtout sur les aspects descriptifs et leur mise en oeuvre avec le logiciel EyeLID. Il est composé de deux parties distinctes. Au §2., on décrit succintement le logiciel. Les sections qui suivront sont consacrées à l'analyse des deux dossiers expérimentaux présentés au chapitre I, pour chacun desquels on a envisagé plusieurs questions qui n'ont pas la prétention de constituer une analyse exhaustive de chaque dossier [1].

2. DESCRIPTION RESUMÉE DU LOGICIEL EyeLID

Nous ne pouvons décrire en quelques pages toutes les fonctionnalités de EyeLID. Les concepts sur lesquels il s'appuie ont été présentés aux chapitres I, II et III. Ceux-ci sont en fait apparents dans le fonctionnement du logiciel et dans le dialogue avec l'utilisateur. On trouvera le détail des commandes, mots-clés, options, *etc.* dans la documentation du logiciel (Bernard, Rouanet, Baldy, 1993). Il suffira qu'on donne ici des précisions concernant les trois "aliments" de base de EyeLID : le *protocole de base* (les données), les *demandes d'analyse* LID (les questions), et les *commandes graphiques*.

2.1. Les données : le protocole de base

2.1.1. *Fichiers de données* ".LID"

Le protocole de base est donné au logiciel EyeLID, sous la forme d'un fichier ASCII (d'extension ".LID") qui comprend les éléments principaux suivants :

- On indique le nombre de *facteurs* et de *variables*. Chacun d'eux est désigné par un *code*, par exemple S ou SUJ pour les "sujets" ; pour chaque facteur on indique, en

[1] On trouvera d'autres exemples, dans un contexte d'analyses post-factorielles dans Bernard *et al.* (1989) et Faye, Bernard (1993).

plus, le numéro de modalité maximum qu'il peut comporter. Ces codes seront utilisés dans les demandes d'analyse : *e.g.* `SUJ` pour le facteur, et `suj1`, `suj2`, *etc.*, pour ses modalités [2].

- Les données elles-mêmes se présentent comme un tableau qui pour chaque unité (une ligne) donne dans l'ordre (en colonnes) : un poids, les valeurs des variables, et les numéros de modalités des facteurs. Chaque ligne est en "format libre" (les champs sont séparés par un ou plusieurs espaces).

Aucune indication concernant les *relations* entre facteurs n'est nécessaire : celles-ci peuvent être quelconques, chaque unité étant décrite par toutes les modalités qui lui correspondent.

2.1.2. *Statut des facteurs et des variables pour* EyeLID

La distinction entre facteurs et variables est essentielle dans EyeLID, les questions étant toujours, de façon schématique, du type : "Quel est l'effet des facteurs sur les variables ?". Les facteurs et les variables constituent tous deux des *variables au sens large*, mais les facteurs sont des variables catégorisées, alors que les variables sont numériques (les variables en $\{0,1\}$ en étant un cas particulier).

Un fichier LID contient toujours plusieurs facteurs qui expriment la planification (des données et/ou des questions), auxquels est automatiquement adjoint le facteur constant `Z`. A ces facteurs "essentiels", on ajoute parfois des *facteurs techniques* qui facilitent certaines analyses ou certaines représentations graphiques [3]. Les relations entre ces divers facteurs peuvent être quelconques — croisement, emboîtement ou simple composition —, EyeLID se contentant de contrôler si telle ou telle relation est vérifiée lorsqu'on utilise certains opérateurs (*e.g.* '`*`','`<>`').

Enfin, un facteur `A` peut n'être pertinent que pour une partie des unités de base, *i.e.* constituer un *facteur partiel* : pour cela, dans le fichier "`.LID`" on affecte le numéro de modalité *0* aux unités non concernées. Toute demande qui fera intervenir `A` ne portera que sur les unités de base concernées.

En conséquence, la structure des données qu'accepte EyeLID est extrêmement souple et peut s'adapter, aussi bien aux données expérimentales aux plans les plus complexes, qu'aux données multidimensionnelles issues de méthodes factorielles.

2.1.3. *Entrée des données ; interfaces*

EyeLID ne possède pas de module d'entrée des données propre. Celles-ci pourront, soit être saisies par traitement de texte (en mode ASCII), soit être exportées à partir d'autres logiciels. Plusieurs interfaces "clés en main" sont disponibles, pour des données déjà saisies pour les logiciels suivants : ADDAD, DS3 (de D. Corroyer), PAC, SAS et VAR3. Mais on pourra facilement en exporter à partir d'autres logiciels qui permettent de créer des fichiers ASCII.

---

[2] En addition du fichier `.LID`, on peut fournir un fichier `.DIC` (pour "dictionnaire") qui associe, à certains des codes, un *libellé en clair*. A l'affichage, on pourra choisir de faire apparaître soit les codes soit les libellés.

[3] Toute liberté peut être ici prise, par exemple : ajout d'un facteur "`AB`" confondu avec "`A*B`", d'un facteur dont les modalités expriment un niveau de performance pour une des variables, d'un facteur issu d'une classification, d'un facteur exprimant un niveau de contribution à un axe (pour les données multidimensionnelles issues de méthodes factorielles), *etc.*.

2.2. Les questions : langage LID et protocoles dérivés

2.2.1. *Les demandes d'analyse* LID

On trouvera la description complète de la syntaxe du langage LID dans la documentation du logiciel. On rappelle seulement ici la syntaxe générale des *demandes d'analyse* de LID les plus communes : "*motclé formulederiv* `->` *varderivlist*", par exemple, "`Table A*B Diff c1,c2 -> V1,V2`". Une telle demande d'analyse a deux fonctions : générer un *protocole dérivé*, et indiquer une *procédure* qu'on lui applique. Détaillons les éléments qu'elle comprend :

- Une formule de dérivation, qui définit le support du protocole dérivé (ici "`A*B`") et le mode de dérivation des valeurs à partir du protocole de base (ici "`Diff c1,c2`"). Celle-ci peut contenir des codes (des facteurs, des modalités) des *mots-clés à droite* prédéfinis (`Diff`, `Mean`, `Var`, *etc.*) et divers *opérateurs* (*e.g.* '`*`', '`,`').
- Une liste de variables dérivées, introduite par le symbole '`->`' [4], qui peuvent être soit des variables du protocole de base, soit des variables définies (*e.g.* par des formules mathématiques) à partir d'une ou plusieurs d'entre elles. La *formule* que constituent ces deux premiers éléments définit un protocole dérivé.
- Un *mot-clé à gauche* de la formule, qui indique la procédure à appliquer au protocole dérivé, par exemple : "`Graph`", "`Table`", "`Mean`", *etc.*. Tous les mots-clés à droite peuvent être utilisés comme mots-clés à gauche (la réciproque n'est pas vraie). A chaque mot-clé correspond un module du logiciel qui peut avoir besoin de la spécification d'*options* (*e.g.* format des valeurs pour "`Table`"). Lorsque nécessaire, on désignera ces options entre crochets, *e.g.* [*option*].

2.2.2. *Récursivité du langage*

Tout protocole dérivé, une fois construit, peut être stocké, pour une utilisation ultérieure lors de la même session, (mot-clé "`Store`") sous un nom commençant par le caractère '`$`'. C'est par cet intermédiaire que se fait notamment la *superposition* graphique de plusieurs protocoles dérivés. Le protocole de base, lui-même, est vu comme un protocole dérivé particulier automatiquement nommé "`$probase`".

Inversement, tout protocole dérivé peut être sauvegardé de façon définitive dans un fichier "`.LID`" (mot-clé "`File`"), et pourra, lors d'une autre session, jouer le rôle d'un protocole de base ; les mot-clés "`Pbase`" et "`Pbtmp`" permettent de considérer, au cours d'une même session, un protocole dérivé comme protocole de base [5].

Ces deux points expriment le premier aspect de la conception *récursive* du logiciel [6] : les données qu'on traite et les résultats de ce traitement ne constituent qu'un *unique type d'objet*, des *protocoles*. L'autre aspect, essentiel bien que moins visible à l'utilisation, est que les formules "en cascade", dont on a vu des exemples, s'interprètent comme la construction de protocoles dérivés successifs [7].

---

[4] Le symbole '`->`' s'obtient à l'aide de deux caractères : '`-`' ("moins") et '`>`' ("supérieur").

[5] Ces trois mots-clés rendent extrêmement aisée la mise en oeuvre l'*approche spécifique*.

[6] Conception récursive grandement facilitée par l'utilisation du langage C.

[7] La formule de dérivation du §2.2.1. génère tout d'abord un protocole dérivé de support "`A*B*c1,c2`", obtenu par moyennage sur le protocole de base (dérivation par défaut) ; le protocole dérivé de support "`A*B`", est obtenu par différence à partir du précédent.

2.2.3. *Opérateurs d'une formule* LID

Les formules LID contiennent un nombre restreint d'opérateurs. C'est leur combinaison récursive qui crée la richesse du langage [8]. Le tableau ci-après décrit succintement ces opérateurs (la plupart d'entre-eux ont été introduits au chapitre III.), en les illustrant par des formules applicables à un protocole de base de structure "`S<A5>*B`" :

Opérateurs	Exemple	Fonction
_ ou _..	`a1_a2` ou `a1_..a4`	Regroupement de modalités ("ou" logique)
	`a1b2`	Conjonction de modalités ("et" logique)
, ou ,..	`a1,a2` ou `a1,..a4`	Enumération de modalités
!	`!a1`	Négation, équivalent à "`a2_a3_a4_a5`"
&	`a1,a2&B`	Composition
*	`S*B`	Composition avec vérification du croisement
<>	`S<a1,a2>`	Composition avec vérification de l'emboîtement
/	`S&B/a1`	Restriction, filtre
()	`S(a1)` ou `S(B)`	Dérivation *intra*
.	`A.B`	Dérivation d'*interaction*
[]	`[A&B](S)`	Symbole parentéthique

2.3. Principales commandes graphiques de EyeLID

Le module graphique (mot-clé "`Graph`") est le plus important (en taille, comme en possibilités offertes). Pour être représenté graphiquement, le protocole dérivé doit être au minimum bivarié, *i.e.* avoir au moins deux variables dérivées à droite de '`->`' [9]. A l'arrivée dans le module graphique le protocole est représenté avec un "habillage" graphique par défaut. Celui-ci peut être modifié à l'aide de "*commandes graphiques*".

Chaque commande graphique comprend un *mot-clé graphique* [10] et une liste d'*arguments*. Les mots-clés graphiques sont construits selon les principes suivants :

- Le graphique est constitué d'un certain nombre d'*objets*. Il y a d'abord les objets relatifs aux *unités* du protocole : celles-ci peuvent apparaître avec une étiquette ("*label*"), un *marqueur* ("*marker*"), et être reliées ("*join*") entre-elles. Les autres objets concernent les caractéristiques graphiques non-liées aux unités : fenêtre ("*window*"), graduations ("*scale*"), axes ("*axes*"), *etc.*. Chaque objet est désigné par une *racine* (d'une ou plusieurs lettres) : `uni`, `lab`, `mk`, `join`, *etc.*.
- Pour chaque objet on peut spécifier certains *attributs* : couleur, taille, type, *etc.* ; Chaque attribut est également désigné par une racine : `col`, `siz`, `typ`, *etc.*.
- Les mots-clé graphiques sont, pour la plupart, composés comme suit : "*objet*" faire apparaître l'objet, "*n-objet*" faire disparaître l'objet, et "*objet-attribut*" changer l'attribut de l'objet. Ainsi, par exemple, "`labcol`" signifie "changer la couleur des étiquettes", et "`njoin`", "supprimer les liens".

On peut schématiquement décrire les principales commandes graphiques d'EyeLID à l'aide du tableau à double-entrée (Objets/Attributs) qui suit (d'autres commandes

[8] Il est souvent possible d'obtenir un même protocole par plusieurs demandes équivalentes. Ces synonymies permettent, comme dans le langage naturel, de traduire au mieux la question qu'on se pose.

[9] Pour un protocole de base univarié '`->V1`', on parvient à des protocoles bivariés en mettant à droite de '`->`' une "pseudo-variable" : un facteur `A`, d'où "`->A,V1`", une constante, *e.g.* "`->0.50,V1`". Pour les protocoles comportant plus de deux variables, le module graphique permet, à tout moment, de sélectionner le couple de variables à représenter.

[10] On note ici les mots-clés graphiques en minuscules (*e.g.* "`labcol`") de façon à les distinguer des mots-clés (*e.g.* "`Mean`"), bien que minuscules et majuscules peuvent être utilisés indifféremment pour les deux.

importantes, mais qui ne s'intègrent pas dans ce schéma seront données à l'occasion des exemples) :

Objets	Attributs Options	Présence Définition (...)	Absence Suppression (n...)	Couleur (...col)	Type (...typ)	Taille (...siz)
Markers	(mk)	mk	nmk	mkcol *i* mkcol *form i* mkcol *form*	mktyp *i* mktyp *form i* mktyp *form*	mksiz *n* mksiz *form n*
Labels	(lab)	lab lab *fac*	nlab nlab *fac*	labcol *i* labcol *form i* labcol *form*	labtyp 1,2ou3	labsiz *n* labsiz *form n*
	(labsel)	labsel *form*	nlabsel *form*			
Join	(join)	join *form*	njoin	joincol *i* joincol *form i* joincol *form*	jointyp *i* jointyp *form i* jointyp *form*	
Arrows	(arr)	arr	narr			arrsiz 0.5
Scale	(sca)	sca 0 1 scax 0 1	nsca	scacol *i*		
Axes	(axe)	axe 0 0 axey 0	naxe	axecol *i*	axetyp *i*	
Window	(win ou w)	win 0 0 10 10				

Quelques-unes des listes d'arguments possibles pour chaque commande sont indiquées : "*fac*" désigne un facteur, "*form*", une formule ensembliste (ne comprenant que des "`, & * <>`"), "*i*", un index (de couleur, de marqueur, de type de lien), et "*n*", une taille entière. Pour une partie d'entre-eux, ces arguments spécifient quelles unités doivent être affectées, et comment elles doivent l'être ; par exemple, "`labcol A*B`" signifie "faire varier la couleur des étiquettes selon le facteur composé `A*B`", "`labcol a1,a2 3`", "choisir la couleur d'index 3 pour les étiquettes des unités dérivées décrites (entre autres) soit par `a1` soit par `a2`".

3. PRÉSENTATION DES EXEMPLES

On reprend ici les deux dossiers décrits au chapitre I, en présentant, sur chacun, plusieurs exemples d'analyses. Chaque exemple est l'occasion de présenter une "stratégie d'analyse" particulière : *e.g.* l'exploration préliminaire des données, l'analyse d'une question spécifique planifiée. Ces diverses stratégies ne sont pas exclusives les unes des autres. Le but poursuivi est d'illustrer quelques situations paradigmatiques, dans chacune desquelles certains traitements sont canoniques du fait de la structure des données et de la question d'intérêt.

4. DOSSIER "NÉGLIGENCE"

4.1. Phase exploratoire : effet conjoint des facteurs systématiques

Considérons le dossier "Négligence" pour lequel la structure du protocole de base est "`S12<C2>*M2*O3->DEV`", où `S` est un facteur de groupe (I.§2.1. p. 10). Rappelons que les facteurs systématiques désignent la condition, active (`c1`) ou passive (`c2`), la main utilisée, gauche (`m1`) ou droite (`m2`), et l'orientation du regard, à gauche (`o1`), au centre (`o2`), ou à

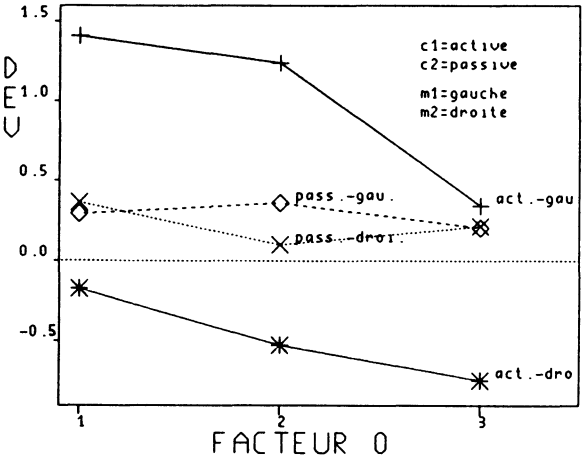
droite (o3) ; la variable DEV représente une déviation positive ou négative, qui est exprimée en *cm* [11].

Une première stratégie d'analyse consiste à étudier globalement les variations de la variable DEV selon l'ensemble des facteurs systématiques C, O et M en moyennant sur le facteur de groupe S. Ceci pourra être fait, dans un premier temps, en mettant de côté les hypothèses ou les questions qui ont présidé au recueil des données et en adoptant une attitude ouverte, qu'on peut qualifier d'exploratoire : regarder comment, en moyenne, les facteurs introduits dans le plan ont "joué".

4.1.1. *Représentation graphique de* C*O*M->DEV

Le protocole dérivé par moyennage sur S peut être désigné au choix par : "C*M*O", "C&M&O", ou "C*M*O Mean S" ; formule à laquelle on adjoint la variable à prendre en compte "->DEV" Pour représenter graphiquement ce protocole dérivé univarié de 12 unités, on choisit de figurer un des facteurs en abscisses, la variable étant mise en ordonnées, d'où par exemple la demande "Graph C*M*O->O,DEV". Ici le choix de O en abscisses est privilégié pour deux raisons : O est ordonné et est celui des trois facteurs qui a le plus grand nombre de modalités [12]. On a indiqué, à droite du graphique 1 ainsi obtenu, l'ensemble des commandes graphiques qui ont permis son habillage ; on a séparé celles-ci en commandes canoniques et commandes "de confort" :

Graph C*M*O->O,DEV (gr.1)



Commandes graphiques canoniques :

```
join O ;
jointyp C*M ; mktyp C*M ; unicol C*M ;
axey O ;
```

Commandes graphiques de confort :

```
nlab O ;
nlabsel ; labsel c1o3,c2o2 ;
mksiz 5 ;
labtyp 2 ;
```

- La commande "join O" signifie "relier o1 à o2 puis à o3, pour tout triplet d'unités *décrites par les mêmes modalités* (à l'exception de O, bien sûr)". Ici, puisque O est croisé avec C*M, on obtient un "profil" selon O pour chaque modalité de C*M. Les commandes "jointyp C*M" ("choisir un type de lien différent selon C*M"), *etc.* ont toutes la même fonction, à savoir, "permettre de distinguer visuellement les 4 profils (par le type de lien, de marqueur, ou la couleur [13] des unités)". Ce qui est paradigmatique ici c'est, du fait du croisement "C*M * O", de *relier sur un facteur*, ici O, et de *distinguer sur*

[11] La valeur *0* pour la variable DEV correspond à l'absence de pseudo-négligence ; les valeurs positives (resp. négatives) indiquent une déviation vers la droite (resp. gauche) ; on dira qu'une déviation est supérieure à une autre en se référant à cette échelle signée.

[12] L'ordre des modalités de O (−30°, 0°, 30°) est une raison structurelle ; relier o1 à o2 puis à o3, dans les graphiques, n'a de sens que lorsque les modalités sont ordonnées. Le grand nombre de modalités de O est une raison de commodité visuelle.

[13] Les commandes sur la couleur sont, malheureusement, mentionnées seulement pour faire travailler l'imagination du lecteur.

le composé des autres facteurs, ici `C*M` [14]. La commande "`axey 0`" permet de faire apparaître l'axe des *y* pour la coordonnée *0*, valeur signifante ici puisqu'elle représente l'absence de pseudo-négligence.

- Les commandes de confort sont : "`nlab 0`" *i.e.* "ne pas faire figurer le facteur `0` dans les étiquettes", "`nlabsel ;...`" pour sélectionner les unités dont on veut l'étiquette, "`mksiz 5`" pour grossir les marqueurs des unités. La commande "`labtyp 2`" permet de passer des "étiquettes formelles", *e.g.* `c1m1`, aux "libellés en clair", *e.g.* "active-gauche".

Quels sont les traits saillants qui se dégagent à l'inspection du graphique 1 :

(i) A l'exception des unités "active-droite", toutes les déviations sont positives, *i.e.* vers la droite, comme attendu dans ce contexte expérimental. Ceci entraîne un niveau moyen de déviation positif, mais dont on constate la grande hétérogénéité d'une situation expérimentale à l'autre.

(ii) En condition passive (`c2`), la déviation moyenne est petite (environ $0.2cm$ vers la droite) et très homogène : il y a peu d'effet conjoint de `M` et de `O`.

(iii) Les différences sont, au contraire, très marquées en condition active (`c1`), avec des déviations positives pour `m1` et négatives pour `m2`, et ce, quelle que soit l'orientation : il y a un effet marqué de `M`, plutôt homogène selon `O`, ce qui signifie peu d'effet d'interaction "`M.O`".

(iv) L'autre vision dissymétrique de cette faible interaction consiste à dire que, pour `c1`, les profils selon `O` pour `m1` et pour `m2` ne sont pas trop éloignés du parallèlisme : ils se caractérisent, tous deux, par une décroissance de la déviation quand on passe de `o1` à `o2`, puis de `o2` à `o3`.

4.1.2. *Du graphique aux chiffres*

Ce qui vient d'être fait revient à essayer de résumer le protocole `C*M*O` en caractérisant l'effet de certaines sources de variation soit comme faible, soit comme important. On peut reconsidérer ce premier résumé "impressionniste" de façon plus méthodique, d'abord, en identifiant par une formule LID chacune des sources de variation commentées, puis ensuite, en quantifiant la grandeur de l'effet de chacune.

- Le point *(i)* a consisté à remarquer que l'effet moyen "`Z`" était positif mais moyenné sur des valeurs hétérogènes : ainsi `Z` seul résume imparfaitement `C*M*O` et il reste beaucoup à expliquer en termes d'*écarts* entre modalités, *i.e.* dans la source de variation "`[C*O*M](Z)`".
- Le point *(ii)* exprime que le protocole "`[M*O](Z)/c2`" des écarts relatifs des modalités de `M*O` pour la condition `c2` est presque constamment nul.
- Le point *(iii)* exprime, entre autres choses, que ces écarts relatifs sont, au contraire, grands pour la condition `c1`.

Ces trois premières remarques peuvent être traduites en terme d'une *décomposition* particulière des effets liés à `C*M*O`, dont les composantes sont : "`Z`", "`[M*O](Z)/c1`", et "`[M*O](Z)/c2`" ; à laquelle il manque le terme "`C(Z)`" pour parvenir à une décomposition additive [15]. Cette décomposition des valeurs est donnée ci-après :

[14] Cette distinction pourrait être introduite en jouant sur des attributs différents en ce qui concerne `C` et `M`, par exemple : `jointyp C ; unicol M`.

[15] Le terme `C(Z)` n'a pas été commenté dans l'analyse "spontanée" du graphique, dans la mesure où s'intéresser à la moyenne de `c1` paraît de peu d'intérêt compte-tenu des différences entre `c1m1` et `c1m2`.

C*M * O

1.408	1.238	0.342
-0.175	-0.529	-0.746
0.292	0.358	0.204
0.363	0.100	0.217

= Z

0.256

+ [M*O](Z)/c1

1.152	0.981	0.085
-0.431	-0.785	-1.002

+ [M*O](Z)/c2

0.036	0.103	-0.051
0.107	-0.156	-0.039

+ C(Z)

Dans cette décomposition, on retrouve, de façon maintenant quantifiée, le faible effet de M*O en c2 (les valeurs sont *toutes* proches de *0*) et le fort effet de M*O en c1 (*certaines* valeurs sont très différentes de *0*).

Les autres remarques des points *(iii)* et *(iv)* reviennent à décomposer, à son tour, le protocole dérivé "[M*O](Z)/c1", en effet de M, effet d'interaction "M.O", puis effet de O ; ce dernier terme est lui-même implicitement décomposé en une *composante linéaire* (la décroissance) et le résidu non-linéaire [16].

Le tableau ci-après résume ces diverses décompositions orthogonales "emboîtées". Avec ce tableau, les conclusions "impressionnistes" issues du graphique, sont maintenant étayées par des chiffres. Les indices de grandeur des effets, Rsdcor (en *cm*), reflètent la discussion précédente.

Source de variation	Rss	Rdf	Rms	Valeur de l'effet	Rsdcor
C*O*M	59.304	12	4.942		0.642
Z	9.430	1	9.430	0.256	0.256
[C*O*M](Z)	49.874	11	4.534		0.615
[C*O*M](Z)	49.874	11	4.534		0.615
C(Z)	0.000	1	0.000	0.001	0.000
[O*M](Z)/c2	0.620	5	0.124		0.102
[O*M](Z)/c1	49.254	5	9.851		0.906
[O*M](Z)/c1	49.254	5	9.581		0.906
M(Z)/c1	39.383	1	39.383	1.479	1.046
O.M/c1	1.481	2	0.741		0.248
O(Z)/c1	8.389	2	4.195		0.418
O(Z)/c1	8.389	2	4.195		0.418
[o1,o3](Z)/c1	8.044	1	8.044	0.819	0.579
[o2,o1_o3](Z)/c1	0.345	1	0.345	0.147	0.098

Concluons en ce qui concerne cet exemple. Nous avons ici cherché à illustrer le fait que, décomposer un protocole en un certain nombre de sources de variation additives, selon les règles définies au chapitre III., peut être envisagé, au niveau descriptif, comme la recherche d'un *résumé efficace* du protocole en question.

4.2. Les niveaux d'analyse d'une question : des moyennes aux individus

Revenons sur l'effet du facteur M en condition active (c1). Dans la section précédente, cet effet a été commenté, à plusieurs *niveaux d'analyse* : au niveau global, en moyennant sur tous les autres facteurs, "M(Z)/c1", et au niveau local, pour chaque orientation (ce qui correspondrait à des demandes comme, par exemple, "M(Z)/c1o1"). Ces effets, global et locaux, peuvent, puisque M a deux modalités, s'exprimer par des différences "Diff m1,m2" qui sont données dans les tableaux ci-après. C'est sur la base de ces résultats, appréciés visuellement dans le graphique 1, qu'on a conclu descriptivement à la relativement faible

[16] Pour un facteur O à 3 modalités, la composante linéaire s'exprime par la comparaison "o1,o3" (protocole dérivé "[o1,o3](Z)") et le résidu non-linéaire par la comparaison "o2,o1_o3". Si le facteur O avait plus de deux modalités, la composante linéaire s'exprimerait par "Z Lin O" (cf. 5.2.).

interaction "M.O/c1", et, par voie de conséquence, à l'intérêt de considérer l'effet moyen de M/c1.

Raw Z Diff m1,m2 -> DEV

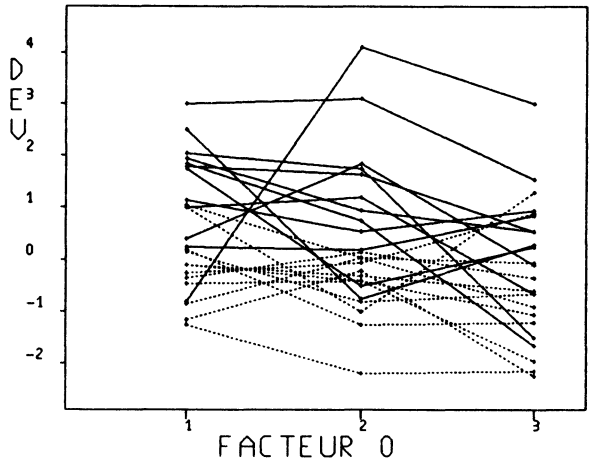
	x^u	w_u
z1	1.479	18

Raw O Diff m1,m2 -> DEV

	x^u	w_u
o1	1.583	6
o2	1.767	6
o3	1.088	6

Mais ces conclusions ne portent encore que sur des moyennes. Qu'en est-il au niveau individuel ? Pour répondre à cette question, la première idée qui vient à l'esprit consiste à réaliser un graphique analogue à 1, en se restreignant à c1, et en figurant tous les sujets, d'où la demande "Graph S*O*M/c1->O,DEV" qui produit le graphique 2 habillé en distinguant les profils sur O en fonction des modalités de M ("join O ; jointyp M"). Les deux sous-ensembles de profils apparaissent assez bien séparés, mais surtout pour les orientations o1 et o2, et nettement moins pour o3.

Graph S*M*O/c1->O,DEV (gr.2)



Commandes graphiques :

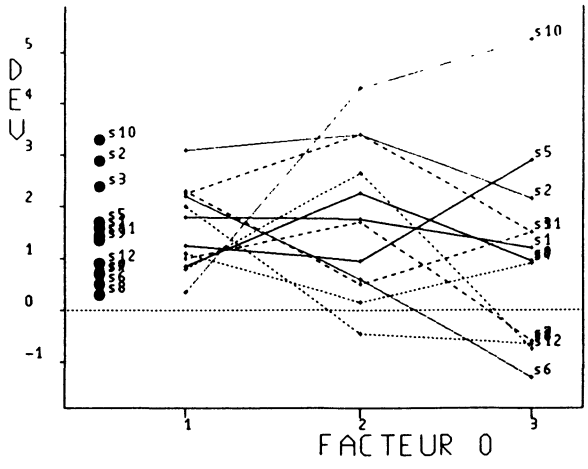
```
join O ; jointyp M ;
nlab ;
```

Cependant ce premier graphique n'est pas le plus pertinent pour notre question initiale : l'appariement des profils, découlant du croisement S*M, n'y est pas apparent. De ce fait, la dispersion interindividuelle qui y est visible est le mélange de deux sources de variation : les sujets diffèrent par leur déviation moyenne, et par leur effet individuel de M [17].

Le protocole pertinent pour étudier l'effet de M au niveau individuel (et en conservant l'orientation) est "S*O Diff M/c1->DEV", qui donne, pour chacun des 12 sujets et chacune des 3 orientations, la différence de déviation entre m1 et m2. Ce protocole est représenté graphiquement dans la partie droite du graphique 3 (avec "->O,DEV" comme variables) ; on constate que seules 5 différences vont dans le sens inverse de l'effet moyen, une en o2 et quatre en o3. A ce protocole dérivé, on a superposé, le protocole des différences moyenné sur O, "S Diff M/c1->DEV" (en lui ajoutant la "pseudo-variable" constante 0.5). Les demandes d'analyse et commandes graphiques nécessaires pour procéder à cette superposition sont indiquées à droite du graphique 3. Le principe général en est le stockage "Store" d'un protocole dérivé, qu'on rappelle ("recall") ensuite dans le module graphique, et qui peut être, lui aussi, habillé par des commandes graphiques, à condition de les faire précéder de la commande "gsup".

[17] On reviendra plus en détail sur ce point au §5.1. ; en quelques mots, le problème soulevé ici est que chaque graphique véhicule un certain *calibrage* visuel, qui n'est pas nécessairement le calibrage le plus pertinent pour l'effet d'intérêt. Ici, on calibre mentalement par "S(Z)+S.M", alors que "S.M" seul suffit ; un tel graphique risque donc d'amener à sous-estimer l'importance de l'effet d'intérêt.

Graph S*O Diff M/c1->O,DEV (gr.3)



Demandes et commandes graphiques :

```
Store S Diff M/c1 -> 0.5,DEV
  [$sdiff]
Graph S*O Diff M/c1 -> O,DEV
  join O ; jointyp S ;
  nlab ; lab S ; nlabsel ; labsel o3 ;
  recall "$sdiff"
  gsup ; mktyp 13 ; mksiz 2 ;
```

De ces deux protocoles superposés il ressort que : *(i)* l'effet de M est plus variable en o2 et o3 qu'en o1 ; *(ii)* en moyennant sur O, les effets individuels de M sont tous positifs. Du fait de ce dernier point, on peut prévoir un *effet-calibré* de M(Z)/c1 élevé. On trouve ici : Rsdcor Z Diff M/c1 = 1.479 et Rsdcor S Diff M/c1 = 0.956, d'où Ec = 1.547, ce qui indique un effet important (Ec > 0.6).

Cet exemple a permis de mettre en lumière le point suivant : à la question générale de l'"effet de M en condition c1", on associe plusieurs questions spécifiques qui se situent, chacune, à un niveau d'analyse différent : global "M/c1", local "M/c1o", voire individuel "M/c1s". A ce dernier titre, la structure de croisement, ici S*M, est privilégiée puisqu'on peut alors caractériser chaque individu par un *effet individuel.*

5. DOSSIER "HORLOGE"

Le dossier "Horloge" a été présenté en I.§2.4. et a été décrit par la structure "S6<D2*O2>*C2*A6*L2->ERR,RT" où S est un facteur de groupe. Il s'agit d'un protocole bivarié : pour chacune de ses 576 unités (24 par sujet), on a mesuré une erreur ERR (en degrés d'angle) et un temps de réponse TR (en secondes).

5.1. Explorer un nuage bivarié à la lumière des facteurs

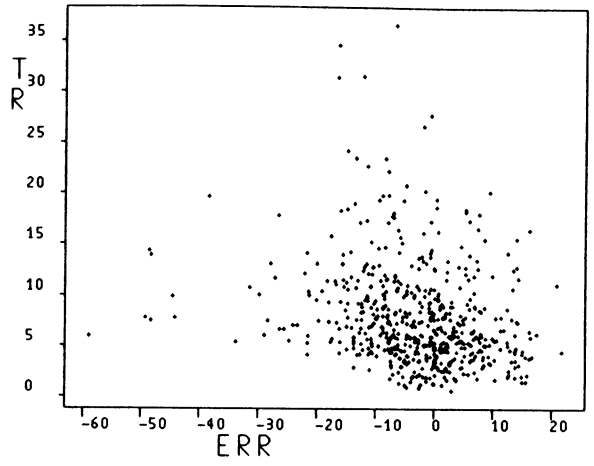
On se situe à nouveau ici lors de la phase exploratoire des données. On a déjà illustré une première stratégie qui, lors de cette phase, consiste à moyenner sur le facteur de groupe et regarder l'effet conjoint des facteurs systématiques (cf. 4.1.). La seconde stratégie, complémentaire de la première, et qui est illustrée ici, consiste à regarder l'ensemble des données en conservant S à la lumière des facteurs du plan. Cela sera particulièrement approprié lorsque, comme ici, le protocole est multivarié.

Les graphiques 4 et 5 correspondent tous deux à la demande d'analyse "Graph $probase" : chacune des 576 unités est représentée par un point dont les coordonnées sont les valeurs des 2 variables du protocole de base ERR et TR. A l'entrée dans le module graphique, on obtient le graphique 4. Pour parvenir au graphique 5, on a utilisé les commandes "mktyp d1 4 ; d2 13 ; mksiz 2", *i.e.* "choix de marqueurs différents pour d1 et pour d2" et "augmentation de leur taille" [18]. Le nuage ainsi obtenu permet de visualiser les variations simultanées de ERR et RT liées au facteur D. Les deux sous-nuages

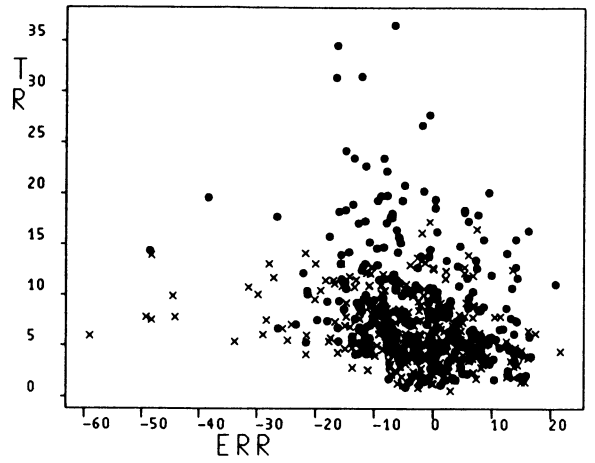
[18] Dans une situation réelle d'exploration, avec un ordinateur muni d'un écran couleur, on utiliserait plutôt la commande "unicol D", plutôt que "mktyp".

ont une large zone de recouvrement, mais on remarque que les unités dont le TR est élévé ($> 20s$) correspondent toujours à la dimension d2 ($30m$). Par contre, il n'y a pas de différences apparentes en ce qui concerne la variable ERR.

Graph $probase (gr.4)



Graph $probase (gr.5)

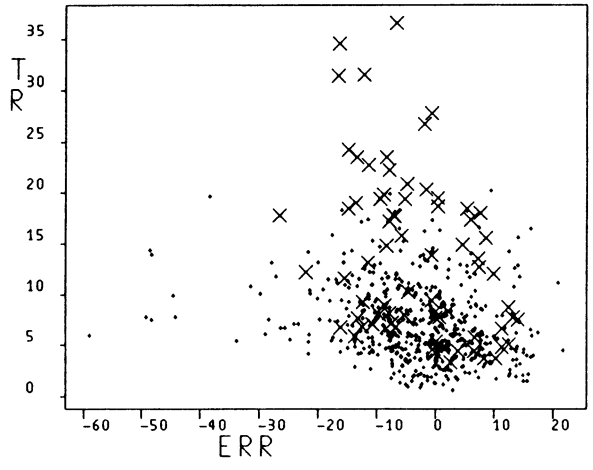


On peut ainsi passer en revue tous les facteurs systématiques du protocole. Pour le côté L, les deux sous-nuages sont très peu discriminables. Pour l'angle (A) on obtient des sous-nuages assez discriminables, à la fois pour ERR et pour RT (les deux sous-nuages pour a1 et a6 ont déjà été montrés en I.§3.2. p. 18). Pour chacun des facteurs O et C, les deux sous-nuages obtenus apparaissent très semblables à ceux observés au graphique 5. On en déduit que les temps de réponses les plus élevés s'observent en conjonction de trois choses : dimension d2 ($30m$), condition c2 (objet-centrée) et ordre o2 (c2 puis c1). Ainsi, la situation d2c2 constitue une tâche plus difficile pour les sujets, et ce particulièrement, lorsqu'elle n'est pas précédée de la situation c1.

Pour aller plus loin dans les possibilités d'EyeLID, les commandes graphiques peuvent porter sur : un facteur composé, *e.g.* "unicol D*C" ou "unicol D*C*O" ; un ou plusieurs sous-nuage particuliers désignés en termes de modalités, *e.g.* "unicol 1 ; d2o2c2 2 ; d1o1c1 3". En bref, dans le monde graphique, on dispose de toute la puissance de la partie *ensembliste* du langage LID. Le graphique 6 donne un exemple de ces possibilités, où on a distingué les unités "d2c2o2" des autres.

Graph $probase (gr.6) Commandes graphiques :

```
mktyp 1 ; d2c2o2 4 ;
mksiz 1 ; d2c2o2 4 ;
```



5.1.1. *Que voit-on dans de tels graphiques ?*

Arrêtons-nous ici un instant pour nous poser la question, "Que peut-on dire relativement à l'importance d'un effet sur la base de tels graphiques ?", qui renvoie à une autre question "Que peut-on voir dans de tels graphiques ?".

Distinguer, dans le nuage global, les sous-nuages associés aux modalités d'un facteur, par exemple `A`, n'est autre qu'une version graphique de l'idée de décomposition *inter-intra* du nuage ou de sa variance : le nuage global figure la variance globale ; chaque sous-nuage indexé par "`a`", la variance *intra*-`a` ; et les écarts entre centres des sous-nuages, la variance *inter*-`A`. En désignant par `U` le support du protocole, la commande graphique "`unicol A`" participe ainsi de la décomposition "`U(Z)=A(Z)+U(A)`".

De ce fait, l'analyse de la plus ou moins grande "séparabilité" des sous-nuages revient à rapporter mentalement l'effet `A(Z)`, aux effets `U(A)` ou `U(Z)`. On retrouve donc ici la même idée que celle de calibrage d'un effet d'intérêt par un effet adjoint, mais avec ici un calibrage qui ne prend pas en compte la structure du plan. De ce fait, la source adjointe implicitement utilisée mélange, entre autres : *(i)* les effets conjoints des autres facteurs systématiques, *(ii)* la variabilité inter-indivuelle globale, et non seulement celle relative à l'effet examiné. Le calibrage mental ainsi réalisé est donc aussi peu spécifique que possible.

Même en ne considérant que des critères internes, on peut envisager diverses manières de définir l'*importance* de l'effet d'un facteur `A` selon la source adjointe par laquelle on le calibre : de la plus générale "`U(Z)`" (la même pour tous les effets), en passant par "`U(A)`", jusqu'à la plus spécifique (dont la formule dépend de la structure statistique du protocole). Nous avons privilégié cette dernière définition dans ce texte. Or, dans les graphiques précédents, la source adjointe spécifique n'est pas dissociée d'un certain nombre d'autres sources de variations non-pertinentes. Mais puisqu'elle est nécessairement de moindre inertie (somme des carrés) que les sources de variation `U(Z)` ou `U(A)`, on peut au moins dire que :

- Une bonne séparabilité visuelle est une indication d'un effet important.
- Une mauvaise séparabilité visuelle n'indique pas forcément un effet faible : il peut s'agir d'un effet important, au niveau spécifique, mais masqué par d'autres effets plus importants, au niveau global.

Ainsi donc, pour juger visuellement de l'importance d'un effet, dans le cadre de l'approche spécifique, cette première phase devra être suivie d'une exploration graphique analogue mais dans laquelle chaque facteur sera examiné en référence à un protocole dérivé plus spécifique, *i.e.* dans lequel on aura éliminé un certain nombre de sources de variation non pertinentes [19].

5.1.2. *Visualisation d'une décomposition inter-intra*

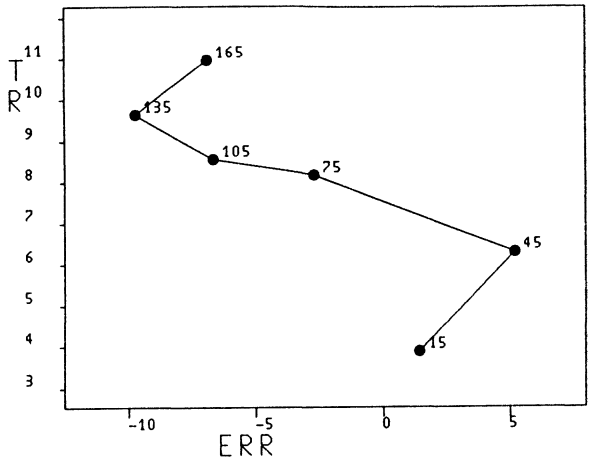
On visualisera encore mieux une décomposition inter-intra de la façon qui suit, illustrée sur l'effet de `A` : on construit d'abord un protocole dérivé moyen sur `A`, représenté dans le graphique 7 ; protocole qu'on stocke sous le nom "`$A`" (au choix, soit par la demande "`Store A->V [$A]`", soit par la commande graphique "`store "$A"`") ; on demande ensuite le graphique d'un protocole jugé pertinent, par exemple "`S<D*O>*A`" [20] ; protocole auquel

[19] "Gommer" des sources de variation d'un protocole, peut se faire dans EyeLID de diverses manières : moyennage, dérivation intra "`()`", dérivation par différence "`Diff`", et "soustraction canonique" de protocoles **(cf. §5.3)**.

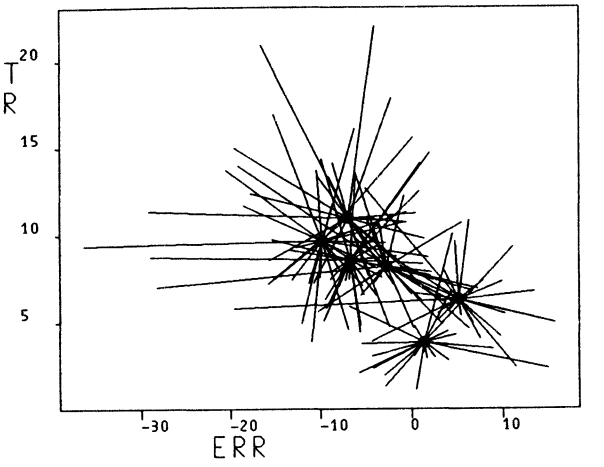
[20] `S<D*O>*A` est plus spécifique que le protocole de base, mais contient encore des sources de variation non spécifiques (en ce qui concerne `A`).

on superpose le précédent (“**recall "$A"**”); enfin, on *joint canoniquement* les deux protocoles par la commande “**pjoin**”. Chaque point moyen “**a**” de “**A**” est alors relié aux points “**S<D*O>*a**” de “**S<D*O>*A**”, comme au graphique 8. Les “étoiles” de ce graphique auront automatiquement la couleur de la modalité “**a**” correspondante.

Graph A->V (gr.7)



Graph S<D*O>*A->V + A->V (gr.8)

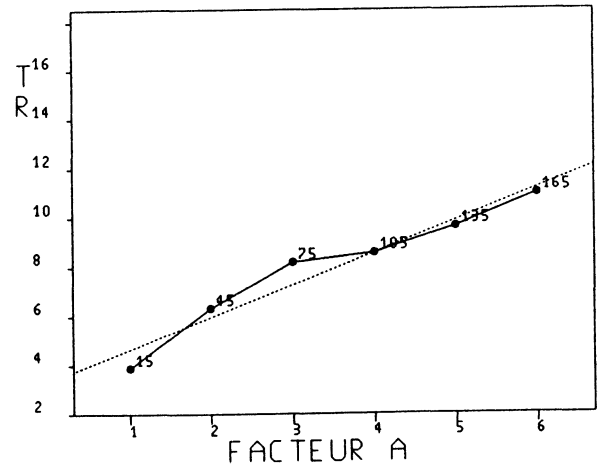


5.2. Analyse d'une hypothèse de recherche : l'effet linéaire de A sur TR

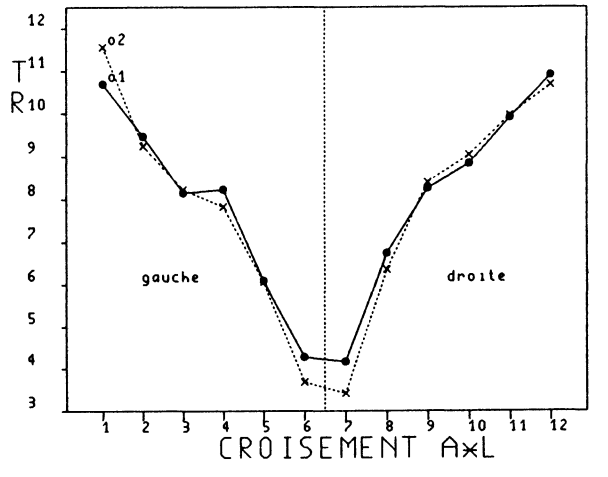
Une des hypothèses de l'expérience “Horloge” stipule qu'on attend une croissance linéaire du temps de réponse **TR** selon l'angle **A**.

On peut tout d'abord représenter la variable **TR** en fonction de **A** au niveau global : c'est ce qu'on a fait au graphique 9 obtenu par la demande “**Graph A->A,TR**”, habillé de la façon habituelle (**join A ; labtyp 2**) et dans lequel on a demandé, en plus, la droite du meilleur ajustement linéaire (“**regyx**”). A ce premier niveau, l'ajustement paraît assez bon.

Graph A->A,TR (gr.9)



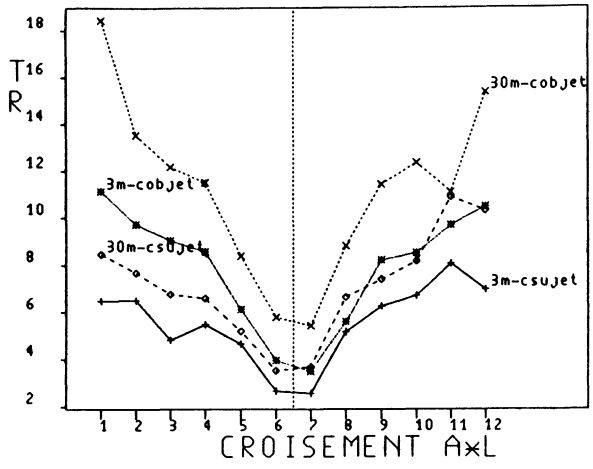
Graph AL*O->AL,TR (gr.10)



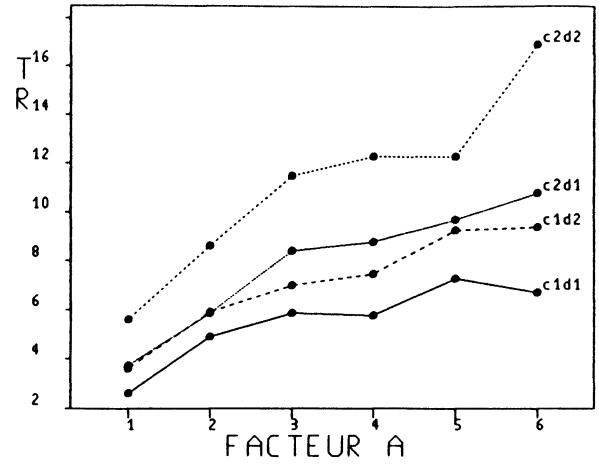
Mais ce premier graphique est obtenu par moyennage sur tous les autres facteurs, et deux profils fortement non-linéaires peuvent se moyenner en une droite parfaite. D'où la nécessité d'étudier, aussi, l'effet de **A** à des niveaux d'analyse plus locaux, ce qui revient à examiner les éventuelles *interactions* de **A** avec les autres facteurs du plan. Les interactions avec les facteurs secondaires **L** et **O** apparaissent faibles au vu du graphique 10

"Graph AL*O->AL,TR" [21]. Les graphiques 11 et 12 correspondent à deux autres niveaux d'analyse locale, respectivement "Graph AL*C*D->AL,TR" et "Graph A*C*D->A,TR".

Graph AL*C*D->AL,TR (gr.11)



Graph A*C*D->A,TR (gr.12)



L'inspection de l'ensemble de ces quatre graphiques témoigne d'une "tendance" linéaire à chacun des niveaux d'analyse, identique quant à sa direction (augmentation de a1 à a6, comme attendu dans l'hypothèse), mais quelque peu variable quant à sa grandeur, surtout selon le facteur composé D*C : la pente de cette composante linéaire de l'effet de A augmente avec la difficulté de la tâche (d1c1, puis d1c2 et d2c1, et enfin d2c2). Ceci se vérifie aisément numériquement en calculant la pente de la droite du meilleur ajustement linéaire, au niveau global ("Lin A"), et au niveau local ("Table D*C Lin A"). Les résultats en sont donnés dans les tableaux ci-après. On constate que cette pente est plus de deux foix supérieure en d2c2 qu'en d1c1.

Lin A -> TR

1.308

Table D*C Lin A -> TR

	c1	c2
d1	0.788	1.354
d2	1.129	1.959

Puisqu'à chacun des niveaux envisagés, l'effet de A va toujours dans le sens d'une augmentation du TR avec l'angle, on s'intéressera désormais ici à l'étude de A au seul niveau global.

Le nouveau mot-clé à droite, "Lin" qui vient d'être introduit, donne la pente de la droite d'ajustement linéaire (celle qu'on visualise par la commande "**regyx**"). Ce mot-clé correspond, en fait, à l'application d'un *contraste* particulier aux six moyennes de A : le *contraste linéaire*, qui vaut dans cet exemple [-5/35,-3/35,-1/35,1/35,3/35,5/35] [22]. L'effet du facteur A, *i.e.* A(Z), a 5 dll, et le contraste "Lin A" ne représente qu'un seul ddl parmi ceux-ci. L'effet résiduel de A(Z), par rapport à "Lin A", comporte les 4 dll manquants, et mesure la non-linéarité. La composante linéaire "Lin A" ne dit rien sur ce résidu (les deux comparaisons associées sont orthogonales), et ainsi la composante linéaire peut être importante, que la composante résiduelle le soit ou non.

Autrement dit, l'hypothèse de recherche initiale apparaît en fait comme une question double : *(i)* l'effet Lin A est notable, *et (ii)* l'effet résiduel est négligeable. Le tableau d'analyse de la variance qui suit permet de se prononcer sur ces deux effets, au niveau descriptif (avec les indices de grandeur, Rsdcor, et d'importance de l'effet, Ec), et au

[21] Le facteur technique AL est confondu avec A*L et permet d'obtenir, au centre, les petits angles, et quand on s'en écarte, les plus grands, avec à gauche, le côté 11, et à droite, le côté 12.

[22] Intuitivement, ce contraste revient à opposer les modalités a6, a5 et a4, d'une part, à a1, a2, a3, d'autre part, mais en donnant des poids plus forts aux modalités extrêmes qu'aux modalités centrales.

niveau inductif avec, à la fois, le test F et les conclusions bayésiennes standard (cf. chapitre VI.). Ce qui figure dans ce tableau ne peut être obtenu avec EyeLID seul ; on a utilisé conjointement EyeLID et PAC pour le remplir [23].

Source de variation	Rss	Rdf	Rms	Rsdcor	ec_{obs}	F_{obs}	$Prob(?) = 0.95$
A	755.991	5	151.598	2.513	1.629	63.66	
S(D*0).A	218.126	100	2.381	1.543			
Z Lin A	718.062	1	718.062	1.308	2.057	101.52	$ec_{par} > 1.705$
S(D*0) Lin A	141.465	20	7.073	0.636			
résidu	39.928	4	9.982	0.645	0.587	8.26	$ec_{par} > 0.449$
adjointe du résid	96.661	80	1.208	1.099			

En conclusion de cette analyse, on peut dire que :

- La pente de la composante linéaire de A est *(i)* descriptivement importante ($ec_{obs} = 2.06 > 0.60$), *(ii)* significativement différente de *0* au seuil unilatéral 0.001/2 ($F_{obs} = 101.52$ à $[1,20]$ ddl, $p_{obs} < 10^{-6}$), mais, qui plus est, *(iii)* notable : $Prob(ec_{par} > 1.705) = 0.95$. La première partie de l'hypothèse est donc vérifiée : la composante linéaire de l'effet de A est notable.
- Par contre, l'effet résiduel *(i)* n'est pas faible au niveau descriptif ($ec_{obs} = 0.587$) ; on ne pourra donc certainement pas le déclarer négligeable au niveau inductif (et ce quel que soit le résultat du test, cf. chapitre VI.) ; *(ii)* le test F indique l'existence de cette composante non-linéaire, test significatif au seuil unilatéral 0.001/2 ($F_{obs} = 8.26$ à $[4,80]$ ddl, $p_{obs} = 0.62\ \ 10^{-5}$) ; *(iii)* les méthodes bayésiennes indiquent que cette composante est *non-négligeable* à la garantie 0.95 : $Prob(ec_{par} > 0.449) = 0.95$. La seconde partie de l'hypothèse est infirmée : la composante non-linéaire de l'effet de A n'est pas négligeable.

5.3. Analyse d'un effet d'interaction multivarié

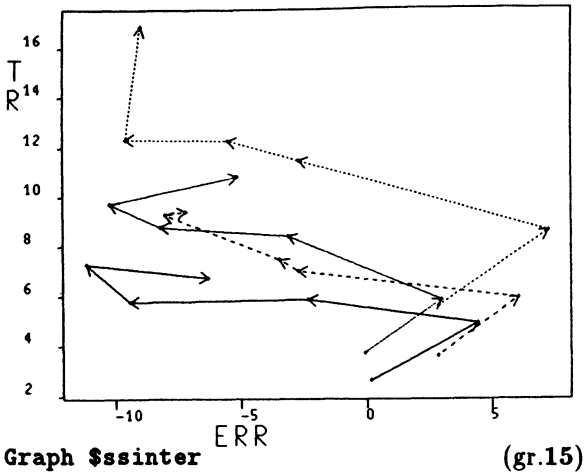
Ce dernier exemple servira à illustrer quelques possibilités supplémentaires du logiciel EyeLID. On cherche à examiner ici l'effet d'interaction entre le facteur A et le facteur composé D*C, *i.e.* "A.D*C" (déjà commenté au §5.2. pour la seule variable RT) en considérant ici l'*effet multivarié* correspondant : "A.D*C->ERR,RT".

Le graphique 13, du protocole "A*D*C->ERR,RT,A", constitue un *diagramme d'interaction* ; comme on voit, le protocole comporte plus de deux variables et seules les deux premières ont, pour l'instant, été représentées. On a effectué l'habillage habituel (join A ; jointyp D*C), avec en supplément la commande "arr" qui fait apparaître une flèche à l'extrémité de chaque lien. Pour chaque modalité de D*C, on a donc un profil orienté, qui va de a1 jusqu'à a6. A l'aide de la seule commande "varxy 3 1", on change les variables représentées (la 3ème variable, A, est affectée aux x ; la 1ère, ERR, aux y), et on obtient le graphique d'interaction univarié 14. Dans chacun de ces graphiques la présence d'interaction s'exprime par le non-parallèlisme des profils.

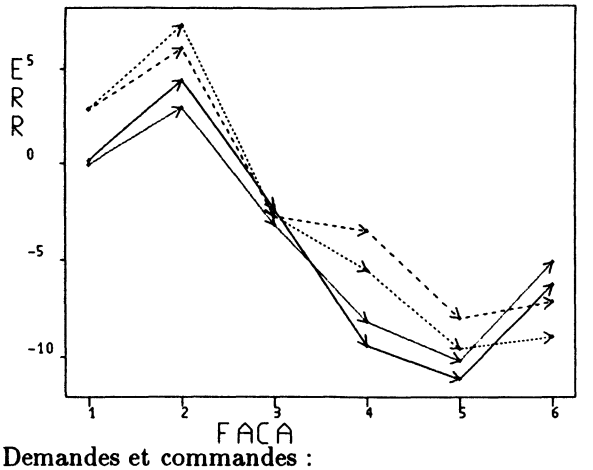
Le graphique 15 est le *protocole sans interaction* associé au précédent (à nouveau représenté dans le plan "ERR/TR"). Il s'obtient par *soustraction canonique* (opérateur "Sub") du protocole d'interaction "A.D*C" au protocole "A*D*C". Les demandes d'analyses et commandes permettant de l'obtenir sont indiquées ci-dessous.

[23] Les lignes "résidu" et "adjointe du résidu" peuvent se déduire des précédentes en utilisant l'additivité des Rss et Rdf, ou s'obtenir dans VAR3 par "A-LIN A", ou encore dans PAC par "-LIN A". Les écarts-calibrés s'obtiennent comme rapport de deux Rsdcor (ou directement par PAC). Les F et énoncés bayésiens s'obtiennent par PAC.

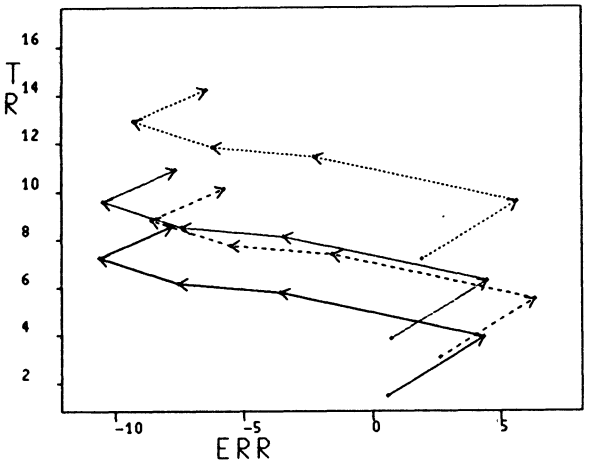
Graph A*D*C->ERR,TR,A (gr.13)



Graph A*D*C->ERR,TR,A (gr.14)



Graph $ssinter (gr.15)

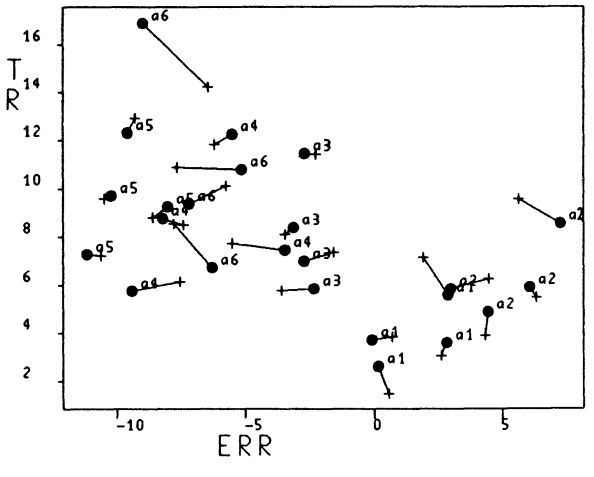


Demandes et commandes :

```
Store A.D*C->ERR,TR,A
  [$inter]
Store A*D*C->ERR,TR,A Sub $inter
  [$ssinter]
Graph $ssinter
  join A ; jointyp D*C ;
  arr ;
```

Enfin, pour étudier une interaction portant, comme ici, sur des facteurs ayant des nombres élevés de modalités, il peut être extrêmement utile de localiser les modalités particulières qui sont "responsables" de l'interaction globale. Ceci se fait en superposant le protocole initial du graphique 13 au protocole sans interaction du graphique 15, puis en les joignant canoniquement, comme dans le graphique 16. Chaque vecteur ainsi obtenu exprime l'effet d'interaction bivarié pour chacune des modalités de "A*D*C" [24].

Graph A*D*C->ERR,TR,A + $ssinter (gr.16)



Demandes et commandes :

```
Graph A*D*C->ERR
  mksiz 3 ; mktyp 13 ;
  nlab ; lab A ;
  recall "$ssinter" ;
  gsup ; mksiz 3 ; mktyp 1 ;
  pjoin ;
```

[24] On prendra garde de ne pas trop hâtivement interpréter les longueurs de ces vecteurs (dans le plan), car les deux variables sont ici "sans commune mesure", et les rapports d'échelles par conséquent arbitraires.

Math. Inf. Sci. hum. (31e année, n° 126, *1994*, p. 97-98)

BIBLIOGRAPHIE GÉNÉRALE

ABDI H., *Introduction au traitement statistique des données expérimentales*, Grenoble, Presses Universitaires de Grenoble, 1987.

AMORIM M.-A., STUCCHI N., "Viewer and object-centered mental explorations of an imagined environment are not equivalent", soumis pour publication (1994).

ATP avec BALDY R., BERNARD J.-M., DUNN G., EVERITT B.S., HAND D.J., LE ROUX B., ROUANET H., SCHILTZ M.-A., TAYLOR C., *Comparative study of statistical methods applied to social science data*, ATP franco-britannique, contrat CNRS 955191 et ESRC I 01 23 0026, 1988.

BERNARD J.-M., "Méthodes d'inférence bayésiennes sur des fréquences", *Informatique et Sciences Humaines*, 68-69 (1986), pp. 89-133.

BERNARD J.-M., "Inférence bayésienne et prédictive sur les fréquences", dans ROUANET *et al.*, 1991, pp. 121-153.

BERNARD J.-M., "Bayesian interpretation of frequentist procedures for a Bernoulli process", soumis pour publication, 1993.

BERNARD J.-M., LE ROUX B., ROUANET H., SCHILTZ M.-A., "L'analyse des données multidimensionnelles par le langage d'interrogation des données LID", *Bulletin de Méthodologie Sociologique*, 23 (1989), pp. 3-46.

BERNARD J.-M., BALDY R., ROUANET H., "The language for interrogating data - LID", in *Data Analysis and Informatics V*, Ed. E. Diday, Elsevier Science Publishers B. V. (North Holland), 1988.

BERNARD J.-M., ROUANET H., BALDY R., *EyeLID-2, Version 2.03, Manuel de référence & Guide de l'utilisateur*, édité par INDIA S.A., 22 rue de Douay, 75009 Paris, 1993.

BOX G. E. P., TIAO G. C., *Bayesian Inference in Statistical Analysis*, Reading, Addison-Wesley, 1973.

CHOKRON S., IMBERT M., "Egocentric reference and asymmetric perception of space", *Neuropsychologia*, 31 No. 3 (1993), pp. 267-275.

CORROYER D., BERT M.-C., "De l'ère des tables à l'ère informatique : Faire de l'inférence sans Khi^2, sans T et sans F ?", *L'Année Psychologique*, 90 (1990), pp. 381-401.

CORROYER D., ROUANET H., "Sur l'importance des effets et ses indicateurs", à paraître dans *L'Année Psychologique* (1994).

DUQUENNE V., "Un programme de description de données", *Cahiers de Psychologie*, 19 (1976), pp. 109-118. DUQUENNE V., "Représentation optimale d'un plan quasi-complet", colloque IRIA, *Analyse des données et informatique*, (1977), pp. 297-302.

DUQUENNE V., "What can lattices do for experimental design ?", *Mathematical Social Science*, 11 (1986), pp. 243-281.

DUQUENNE V., MONJARDET B., "Relations binaires entre partitions", *Mathématiques et Sciences humaines*, 80 (1982), pp. 5-37.

EHRLICH M.-F., "Problèmes pédagogiques - L'enseignement des notions élémentaires de méthodologie expérimentale : la planification des expériences", *Math. Sci. hum.*, 50 (1975), pp. 39-50.

FAYE B., BERNARD J.-M., "Exploration de données par l'analyse post-factorielle : utilisation d'un logiciel d'interrogation de données structurées", à paraître dans les actes de "Ecopathologie et gestion de la santé animale", Clermont-Ferrand, dans *Veterinary Research*, Octobre 1993.

GUIGUES J.-L., "Angles de deux comparaisons inter-groupes ; applications à l'interaction", Note interne, Groupe Mathématiques et Psychologie, Université Paris V, Sorbonne, Paris, 1981.

HAYS W. L., *Statistics*, New York, Holt Rinehar and Winston, 1981 (3è ed.).

HOC J.-M., *L'analyse planifiée des données en psychologie*, Paris, Presses Universitaires de France, Collection "Le Psychologue", 1983.

JEFFREYS H., *Theory of Probability*, 3rd ed., Oxford, Clarendon Press, 1961

LEBEAUX M.-O., LEPINE D., ROUANET H., *Notice d'utilisation du programme* VAR3, Groupe Mathématiques et Psychologie, Université Paris V, Paris, Sorbonne, 1976.

LECOUTRE B., *L'analyse bayésienne des comparaisons*, Lille, Presses Universitaires de Lille, 1984.

LECOUTRE B., "Ouvrage sur l'analyse des comparaisons", Texte provisoire non publié, Groupe Mathématiques et Psychologie, Université Paris V, Sorbonne, Paris, Mai 1991.

LECOUTRE B., POITEVINEAU J., *PAC : Programme d'Analyse des Comparaisons*, Paris, Diffusé par : CISIA, 1, avenue Herbillon, 94160, Saint Mandé, 1992.

LÉPINE D., "Facteurs et plans, I : Structure de finesse", *Math. Sci. hum.*, 57 (1977a), pp. 5–26.

LÉPINE D., "Facteurs et plans, II : Plans quasi-complets", *Math. Sci. hum.*, 58 (1977b), pp. 5–24.

LE ROUX B., ROUANET H., "L'analyse statistique des protocoles multidimensionnels : Analyse des comparaisons (Nuage pondéré sur le croisement de deux facteurs)", *Pub. Inst. Stat. Univ.*, 28 fasc. 1 et 2 (1983), pp. 47–70.

LE ROUX B., ROUANET H., "L'analyse multidimensionnelle des données structurées", *Math. Sci. hum.*, 85 (1984a), pp. 5–18.

LE ROUX B., ROUANET H., "Le plan S<A2*B2> : dérivations, effets et comparaisons ; le paradoxe du renversement des effets", Note interne non publiée, Groupe Mathématiques et Psychologie, Université Paris V, Sorbonne, Paris, Février 1984b.

POITEVINEAU J., LECOUTRE B., "PIF : un programme d'inférence fiducio-bayésienne", *Informatique et Sciences Humaines*, 68–69 (1986), pp. 77–78.

REUCHLIN M., *Introduction à la recherche en psychologie*, Paris, Nathan, Collection : Université – Psychologie, 1992.

ROUANET H., "Some aspects of Bayesian multivariate analysis", Communication à la "Multivariate section of the Royal Statistical Society", 1985.

ROUANET H., "Asserting the Importance or the Negligibility of Effects with Bayesian Methods", soumis pour publication (1994).

ROUANET H., BERNARD J.-M., LECOUTRE B., "Nonprobabilistic statistical inference : a set-theoretic approach", *The American Statistician*, 40 (1986), pp. 60–65.

ROUANET H., BERNARD J.-M., LE ROUX B., *Statistiques en sciences humaines : Analyse inductive des données*, Paris, Dunod, 1990.

ROUANET H., LECOUTRE B., "Specific inference in ANOVA ; From significance tests to Bayesian procedures", *Brit. J. Math. Stat. Psyc.*, 36 (1983), pp. 252–268.

ROUANET H., LECOUTRE M.-P., BERT M.-C., LECOUTRE B., BERNARD J.-M., *L'inférence statistique dans la démarche du chercheur*, Berne, Peter Lang, Publications Universitaires Européennes, Série VI : Psychologie, 1991.

ROUANET H., LÉPINE D., "Comparison between treatments in a repeated-measurement design : ANOVA and multivariate methods", *Brit. J. Math. Stat. Psyc.*, 23 (1970), pp. 147–163.

ROUANET H., LÉPINE D., "Note méthodologique ; Statistiques de groupe, groupes d'observations", *Mathématiques et Sciences Humaines*, 41 (1973), pp. 31–36.

ROUANET H., LÉPINE D., "Structures linéaires et analyse des comparaisons", *Mathématiques et Sciences humaines*, 56 (1976), pp. 5–46.

ROUANET H., LÉPINE D., "Introduction à l'analyse des comparaisons pour le traitement des données expérimentales", *Informatique et Sciences humaines*, 33–34 (1977).

ROUANET H., LÉPINE D., EHRLICH M.-F., MARQUER P., PLAS R., "Introduction aux procédures élémentaires d'analyse descriptive des données", *Bulletin de Psychologie*, (1975–1976), pp. 212–221.

ROUANET H., LE ROUX B., *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.

ROUANET H., LE ROUX B., BERT M.-C., *Statistiques en sciences humaines : Procédures naturelles*, Paris, Dunod, 1987.

SCHEFFÉ H., *The analysis of variance*, New York, Wiley, 1959.