Travaux Dirigés du Diplôme d'Etudes Approfondies de Didactique des Sciences

ANALYSE de DONNEES en DIDACTIQUE

par

Guy BROUSSEAU

Laboratoire Aquitain de Didactique des Sciences et des Techniques
Université BORDEAUX 1
et
Institut Universitaire de Formation des Maitres d'Aquitaine

1993

LADIST, Université BORDEAUX 1, 40 Rue Lamartine, 33 400 TALENCE

SOMMAIRE

1.Les Données

présentation

types de tableaux homogènes

réussites echecs
réponses nominales
réponses ordonnées
réponses à intervalles numériques
réponses numériques

matrices à traiter

les variables actives, les variables explicatives,
les sujets effectifs, les caractères explicatifs.

2. Les types d'études

typologies de questions, de situations

typologies d'élèves

recherches de dépendances diverses, de facteurs, etc.

construction d'un instrument d'évaluation,

représentation des connaissances,
recherche d'un questionnaire minimal pour une évaluation,
recherche d'une épreuve pour un examen, pour un concours

3. Les Méthodes

stratégies d'analyse suivant les types d'études: choix des traitements, résultats, décisions, pour de nouveaux traitements.

principales méthodes

l'analyse en composantes principales
l'analyse factorielle des correspondances, simple, multiple
l'analyse hierarchique,

le choix des variables ou des caractères explicatifs

l'analyse canonique,
l'analyse implicative,

4. La préparation de l'étude.

le plan d'expérience

les variables et les caractères retenus à priori pour l'analyse

étude de l'indépendance à priori des caractères et des variables:

transformations et traitements de la matrice à priori.

La mise au point du questionnaire suivant le type d'étude.

La présentation des informations nécessaires

sur le questionnaire
et sur le déroulement de l'enquête

5. Les divers traitements et leur mise en oeuvre

difficultés de manipulation des divers programmes

erreurs et corrections

6. L'étude formelle et la présentation des "résultats" des différents traitements.

Etude d'ensemble de chaque traitement

Etude de l'inertie, autocorrélation, facteur "G", axes significatifs
mise à l'épreuve globale des facteurs explicatifs avancés à priori
Signification des proximités et des voisinages de directions
Recherche des facteurs classiques,

Etudes suivant les plans principaux (AFC, ACP)

Recherches des facteurs
Identification des éléments pertinents nouveaux
Identification de discriminations,

Etude des niveaux des classifications hiérarchiques (C.A)

choix des distances et des méthodes d'agrégation
comparaison entre les différents traitements

Etude implicative

essais d'interprêtation en termes de dépendances stochastiques des résultats précédents
recherche des implications (CHIC)
Inférences, hypothèses et comparaisons particulières avec la matrice à priori.
Les facteurs, les interactions entre les facteurs explicatifs primitifs.

Les traitements secondaires:

modifications du corpus,
vérification des facteurs, nouvelles "matrices à priori"

Conclusions, importance relative des facteurs.

7. Présentation des résultats de l'analyse

Présentation standard et ses arguments

La formulation en termes de test d'hypothèses et de preuves

Les dangers de l'étude de l'étude d'un élément isolé

8. Les Conclusions de l'étude

apports de la contingence aux questions théoriques avancées

stabilité des explications, fiabilité des résultats,

Questions à reprendre,

critiques éventuelles sur la méthode

Questions nouvelles ouvertes grâce à l'étude

Analyse de Données en Didactique

1.Les données.

Présentation.

Les données recueillies se présentent généralement au moment du traitement sous la forme d'un ou de plusieurs tableaux de p colonnes et de n lignes.
Dans chaque cellule du tableau figurent des nombres ou des lettres.
En tête de colonnes figurent généralement un numéro de la colonne et un code
En tête de ligne figure soit un code soit simplement un numéro de ligne.
Ces lettres ne sont que des symboles élaborés selon des codages qui représentent les informations à traiter, et dont il faut absolument donner la clé au lecteur.
Nous allons considérer dans un premier temps que

- chaque ligne est associée à une observation, en général un sujet
- chaque colonne est associée à une variable ou à une question
- et que chaque cellule comprend la valeur de la variable-colonne attribuée au sujet-ligne.

Exemple: Un étudiant a recueilli les réponses d'un échantillon d'instituteurs à un questionnaire et a obtenu dans un premier temps le tableau ci après: "REPONSA". Nous allons suivre pas à pas les les étapes de son travail et relever les erreurs qu'il a commises.
Les étapes seront vraisemblablement nombreuses et justifieront la création de nombreux objets:
nouveaux fichiers, nouvelles variables, nouvelles valeurs ...
Si le chercheur ne veut pas perdre son temps et son travail, il doit noter soigneuseent au fur et à mesure les objets construits.

Le Codage.

Le choix des symboles

Leur choix est important.
Ils doivent:

- pouvoir être traités par l'ordinateur suivant les opérations prévues pour l'analyse: suivant que le symbole est fait de lettres ou de nombres effectuer une somme sera impossible ou possible.
- pouvoir entrer dans les normes des programmes (par exemple pour apparaître sur les graphes des plans principaux, dans STATITCF, les objets doivent être codés par des symboles de 3 lettres)
- pouvoir évoquer facilement l'objet qu'ils représentent pour l'analyste et le genre d'opérations mathématiques ou logiques qu'il est possible d'effectuer sur eux:

Le système utilisé pour produire les symboles peut peser très lourdement sur l'analyse.

Le codage des observations (sujets).

Il est possible de désigner les sujets par leur nom ou par un code évoquant leur nom, en particulier lorsque l'analyse est faite par l'enseignant qui peut faire bon usage de cette information.

En général en didactique l'intérêt se porte sur les comportement des sujets et sur les situations qui les provoquent. L'identité de ces sujets est sans intérêt. Le codage des sujets sera donc laissé aux soins de la machine qui se contentera de les numéroter automatiquement. Il est important que cette identification ne soit pas considérée comme une variable car ce n'est pas une information sur la contingence: L'analyste ne doit donc pas l'entrer lui même dans la matrice des données.

Il est clair que la première colonne "num" du tableau "REPONSA" doit être éliminée.

Le codage des variables

Au contraire il faut apporter le plus grand soin au codage des variables et faire figurer le tableau de définition et de codage des variables dans les annexes:

Pour chaque colonne indiquer le ***code*** de la variable, sa ***définition***, le ***code des valeurs*** et les ***valeurs*** que cette variable peut prendre et qui constituerons les comportements ou les réponses relevées:

Exemples: ***Code*** ***Définition de la variable***

AGE: Age du sujet

ANC: Ancienneté dans la profession

CLA: Classe dans laquelle le sujet enseigne:

	code des valeurs	***Valeurs de la variable***
	1	Ecole Maternelle
	2	Cours préparatoire
...	4	Cours élémentaire
	5	Cours moyen 1ère année
	6	CM2
	7	6 ème etc.

LIV: Livre utilisé avec les élèves

1	livre paru antérieurement à 1978
2	livre paru entre 1978 et 1985 etc

... 3XF: La variable n'a pas reçu de nom

Définition: Réponses à la questionà choix multiple suivante:

Au problème suivant: "Jean achète 3 petits pains à 4 F, combien doit il?" un élève écrit la réponse suivante: "il doit 3x4F = 12F" Comment jugez vous cette réponse?

Valeurs de la variable: "parfaite, acceptable, à rejeter absolument, je ne sais pas"

codage des réponses: 1, 2, 3, 4, 7

Le codage des valeurs des variables.

Les valeurs portées dans les cellules sont exactement l'une de celles annoncées dans le tableau. Toute modification ou complément doit se répercuter sur la définition de la variable.

Exemple: Nous trouvons dans la colonne 3 des valeurs étranges: 456 ou 56 qui ne figurent pas dans le codage. Manifestement on a voulu porter une information complémentaire du genre: " Le sujet enseigne (ou a enseigné) dans plusieurs niveaux (4 et 5 et 6) ; Ou bien il faut ajouter ces valeurs à la variable et rectifier les autres définitions: "le sujet n'enseigne qu'à tel niveau", mais alors les sujets enseignant au niveau 4 et 5 ne seront comptés ni avec les sujets enseignant au niveau 4, ni avec ceux enseignant au niveau 5. Ou bien il faut "éclater la variable" en variables binaires: enseignez vous en EM?, 1 pour oui, 0 pour non, enseignez vous en CP ?...etc.

Ce dernier procédé permet la transcription plus complête des informations mais il augmente le nombre de colonnes de façon assez artificielle puisque ces variablês sont presque toutes exclusives. Et comme le nombre de colonnes que peut accepter les microordinateurs que nous utilisons nous risquons de ne pas pouvoir effectuer les analyses pour conserver des renseignements d'intérêt mineur.

On va donc ne noter que la classe la plus haute dans laquelle le sujet enseigne.

On aurait pu aussi ajouter une seule variable: enseigne dans plusieurs classes, valeurs: "avec CM" ou "sans CM",

Y a-t-il d'autres erreurs semblables dans la matrice?

Erreurs. Corrections de la matrice

Un examen attentif de la matrice est indispensable. Il montre ici d'autres anomalies qui peuvent n'être que des erreurs de frappe.

Il est nécessaire d'éliminer les colonnes constantes: celles où tous les sujets présenteraient la même valeur: ces colonnes conduisent à des divisions par zero, certains programmes les traitent de façon spéciale mais d'autres s'interrompent brutalement sans rémission

Il est possible aussi d'éliminer alors de la matrice certaines lignes manifestement erronnées ou tout à fait étranges.

STATITCF permet la suppression d'observations, la suppression de colonnes s'effectue avec le sous/programme de "Transformation-fusion".

Les fichiers informatiques

Différents types de fichiers assiciés à une matrice

1. Chaque matrice est ainsi constituée seulement par les valeurs de chacune de ses n x p cellules. et elle constitue l'un des fichiers, le *fichier des données*.

Dans notre exemple ce fichier s'appelle REPONSAR

2.Elle est accompagnée d'un fichier, *le fichier des caractéristiques*, qui contient tous les renseignements nécessaires pour reconnaitre et manipuler cette matrice: titre de l'étude, nombre de lignes et de colonnes, les codes des variables.

Dans notre exemple il s'appelle REPONSA.

3. Si nous avions voulu identifier les observations à l'aide d'un code différent de la numérotation automatique, soit pour reconnaitre les sujets, soit parceque les liges représentent des questions, (après une transposition par exemple), ou pour toute autre raison il aurait fallu constituer un fichier spécial. Dans STATITCF ce fichier se serait appelé ici REPONSI.

Codage des fichiers

Avec ce logiciel l'analyste ne manipule que le fichier des caractéristiques, il ne doit jamais terminer le nom de ses fichier par un R sinon l'ordinateur le confondrait avec un fichier de données.

Chaque nom de fichier ne doit pas comporter plus de 8 lettres.

Il est important d'avoir un système de codage des fichiers qui permette de reconstituer sans peine son origine et ses caractères car il est fastidieux d'aller toujours consulter ses caractéristiques .

Les codages de l'exemple choisi ici ne sont pas particulièrement astucieux ni commodes.

Après avoir effectué les corrections et les suppressions annoncées ci dessus nous obtenons e fichier de caractéristiques "REPONSB" qui gère le fichier de données "REPONSBR",

Les différents types de variables.

Types selon leur structure mathématique

La matrice est elle homogène du point de vue de la structure mathématique des variables.

Quelles sont les variables qui sont numériques? celles qui sont ordinales,? celles qui sont nominales? Y a-t-il des dates?

Dans notre matrice deux variables seulement sont numériques: AGE et ANC, les autres sont nominales, la plupart non binaires. La matrice n'est pas homogène.

Nous verrons dans la suite de cet ouvrage comment on peut homogénéïser un ensemble de variables sans trop perdre d'informations (matrice éclatée) ou comment on peut agréger des données de natures différentes.

Les variables actives et les variables explicatives ou supplémentaires

Dans ce questionnaires,

- certaines questions portent sur l'objet d'étude: le traitement didactique de la commutativité de la multiplication: c'est le cas des variables des colonnes 11 à 25.

Ces variables sont dites ***variables principales*** ou ***variables actives***. Elles constituent une matrice dite matrice principale ou ***matrice à expliquer***

- d'autrs questions renseignent sur les caractéristiques professionnelles ou personnelles des sujets interrogés. Le chercheur espère trouver l'explication des différences de comportements à l'endroit de la commutativité dans ces caractéristiques des sujets. Ces variables seront dites ***supplémentaires*** ou ***explicative***s. Elles constituent la matrice des variables explicatives.

Dans l'AFC ou l'ACP seule la matrice principale sert à déterminer les axes d'inertie. Après cette détermination, les variables explicatives sont projetées dans l'espace à expliquer et leur proximité avec certains objets, ou leur liaison avec certains axes indiquent dans quelle mesure telle ou telle caractéristique des sujets est liée à leur comportement.

Les observations supplémentaires: la matrice des caractères explicatifs des variables

Types d'explications

Il n'y a aucune raison de se cantonner en didactique à l'examen des caractéristiques des sujets, des élèves en l'occurence. Il est peut être tout aussi intéressant de s'interroger sur le rôle des variables elles mêmes, c'est à dire des questions, ou des situations. Le véritable objet d'étude est l'ensemble des questions, ou des conditions des comportements: ils constituent l'échantillon des variables à décrire. Les élèves sont les révélateurs des structures de cet ensemble de questions ou de conditions. Ce point de vue est inhabituel parce que nous avons l'habitude de considérer que la structuration des questions n'est pas une affaire de statistiques mais de logique, de connaissance, de culture et qu'elle doit se faire à priori. Il est exact que les questions semblent différer entre elles beaucoup plus que les sujets entre eux. (Est-ce bien exact?)

Exemples de caractères variables

Par exemple on peut imaginer que, pour une question sur la commutativité de la multiplication, le fait de repésenter les objets à compter en tableau, disposé d'une façon ou d'une autre peut jouer un rôle dans les choix des réponses.

On constitue alors une variable définie comme suit sur l'ensemble des QUESTIONS:
RLC: La question comprend elle le dessin d'un rectangle montrant les objets liés en colonne?

Un des objets de la didactique est bien de confronter cette structure à priori des conditions et des questions avec la contingence des comportements humains et de déterminer le rôle réciproque des deux structures: celle des sujets représentée par la matrice supplémentauire et celle des situations dans l'évolution des comportements.

Nous pouvons envisager un ensemble de caractères pertinents pour opposer les situations présentées aux sujets:
OTI: Le dessin et l'écriture proposée aux élèves se correspondent suivant les instructions ministérielles de 1972.
NTI: Il y a un dessin mais l'écriture contredit les instructions de 1972
UTI: 1 l'écriture de la commutativité est utile au calcul
2 l'écriture de la commutativité est utile au raisonnement de l'élève
3 Déclarer la commutativité ne sert qu'à justifier l'oubli d'une règle
4 Déclarer la commutativité ne sert à rien.
etc.

Matrice à priori

Le chercheur relève donc les caractères des questions ou des variables dont il pense qu'ils pourraient avoir un rôle de caractère explicatif: la ***matrice des caractères explicatifs des variables*** ou ***matrice à priori.*** Elle se présentera comme des observations supplémentaires, ou comme des ***sujets fictifs supplémentaires***.

Le problème se résoudra de la même manière si l'on veut comparer les observations ou les sujets à des PATRONS de réponses, qui représentent des comportements considérés comme typiques. La position des patrons (ou des individus supplémentaires) permettra d'identifier la densité des sujets dans leur voisinage (donc leur caractère représentatif) et leur position par rapport aux facteurs et aux éléments d'explication retenus.

La matrice à priori comprend p colonnes comme la matrice principale. Cela veut dire que les caractères des questions doivent être définis sur TOUTES les questions. Au besoin il est possible de définir une variable à plusieurs valeurs: caractère non défini, caractère défini telle valeur, caractère défini telle autre valeur etc. Elle présente autant de lignes que de caractères envisagés comme explication des comportements.
Comme la matrice principale la matrice des observations supplémentaires à son fichier de données et son fichier des caractéristiques.
Il est très important de bien remarquer que cette matrice des caractères des questions représente l'opinion de l'observateur et non pas la contingence: elle ne servira pas à déterminer les axes de la matrice principale, comme les variables supplémentaires.

Valeurs des caractères des variables

Il n'y a pas de raison qu'un caractère d'une variable ait la même structure mathématique que la variable elle même, donc qu'elle présente les mêmes valeurs et qu'elles soient codées de la même

manière, ce qui est naturellement réalisé avec un patron de réponse. La position des caractères explicatifs dans les plans principaux dépendra des formes de codage.

Il sera prudent de coder les cellules de la matrice à priori, avec des symboles appartenant à la liste de ceux qui sont utilisés pour le matrice princuipale et il pourra être utile de faire divers essais, d'autant que l'on va juger la valeur des moyens d'explication déployés à priori par le chercheur à l'amplitude des liaisons de ces caractères des variables avec les axes d'inertie.

Matrices primitives, matrices dérivées ou secondaires.

Il est possible de construire de nouvelles variables en combinant les valeurs des variables déjà obtenues pour mieux faire apparaître certaines propriétés. Les variables ainsi obtenues sont appelées variables dérivées ou secondaires. Ces transformations n'apportent aucune information nouvelle, au contraire en général elles en perdent. C'est la matrice des données initiales ou matrice primitive qui fait l'objet des analyses dont il est question dans ce cours.

2. Les Types d'études

Les raisons de rassembler des informations peuvent être diverses

1A. Structurer l'ensemble des élèves selon un modèle

Matrices « réussite échec »

Les questions posées sont déjà structurées de façon à établir les conributions à un rangement ou à un classement des élèves.

Il s'agit par exemple de ranger les élèves par ordre de "mérite" pour un concours ou d'identifier ceux qui ont telle ou telle conception d'une notion mathématique.

L'exemple le plus simple est celui d'un concours ou d'un examen basé sur une suite d'épreuves: une seule variable à deux valeurs est attachée à chaque épreuve: la réussite ou l'échec. Toutes ces variables sont supposées être des indices d'une même variable: la qualité des connaissances de l'élève qui peut être représentée par la somme (ou par toute autre fonction) des réussites à chacune des épreuves.

La construction et l'étude de ces variables dérivées peut se faire assez facilement soit avec STATITCF, soit avec MULTIPLAN. La plus simple de ces fonctions, la somme est très utilisée. La distribution des sommes est dite distribution marginale élèves. Erreur à éviter: Evoquer èses caractéristiques (moyenne, écart type, modalité...) sans que le mode de construction soit vraiment précisé.

Matrices « type d'erreurs »

Il est possible détailler les comportements des élèves en relevant d'autres types de variables comme les sortes d'erreurs observées. Il faut alors que le modèle d'agrégation, c'est à dire la formule qui permet de construire la structuration cherchée tiennent compte de la signification des valeurs retenues contribution (pour ne pas ajouter des nombres de réussites avec des numéros d'erreurs par exemple).

Un procédé fréquemment employé consiste à déterminer à priori un patron de comportements (un pattern) et à compter les élèves qui présentent ce profil. L'analyse de données peut permettre d'assouplir ce projet rigide (souvent très peu d'élèves présentent le profil déclaré) et de montrer quels sont les élèves qui possèdent des profils "assez voisins" du profil cherché.

En général on ne mélange pas les variables "réussite-echecs" et les variables types d'erreurs dans ce genre d'études pour éviter de compliquer l'analyse.

1B. Rechercher une (ou des) structuration(s) de l'ensemble des élèves.

II s'agit à l'inverse du cas précédent de rechercher et d'établir des patrons adaptés à la contingence. Plusieurs cas se présentent:

Construction de barêmes

Il s'agit de construire une fonction définie sur l'ensemble des valeurs des variables pour ranger ou discriminer les sujets.

Par exemple il faut fabriquer un barême pour un examen ou pour un concours de façon à obtenir certaines propriétés.

Pour un concours on veut avoir le moins possible d'ex aeqos. Pour un examen on veut avoir le moins possible de candidats dans la zone de discrimination entre les reçus et les ajournés etc.

Classification automatique, recherche de conceptions

On veut déterminer des groupes d'élèves qui se ressemblent suffisamment entre eux par leurs réussites ou par le genre d'erreurs qu'ils commettent par exemple pour leur faire effectuer des exercices appropriés. Par ce moyen on peut tenter d'identifier les ***conceptions*** qui ont cours chez les élèves et qui présideraient à de tels regroupements.

Constitution d'échantillons

Ou encore on veut extraire de la population

- des ***échantillons*** plus ***homogènes*** pour les étudier
- ou au contraire des échantillons moins nombreux mais bien ***représentatifs*** de l'ensemble de la population des élèves.

Par exemple au lieu d'utiliser la matrice de "réussite et échecs" seulement suivant un modèle déclaré pour fabriquer un ordre total des élèves on peut l'analyser pour regrouper les élèves qui ont réussi ou échoué à peu près aux mêmes questions et rechercher ainsi une structuration lattente ou des patrons plus pertinents pour décrire l'ensemble des élèves. Les rapprochements seront fondés sur des distances formelles. (cf: Stratégies de l'analyse statistiques, chapitre 2, ou le cours de Statistiques pour les professeurs d'écoles).

Il convient alors à l'inverse de ce que nous avons fait au paragraphe précédent d'éliminer les traces de jugement à priori sur la signification des variables recueillies.

Matrice « éclatée »

Un bon moyen consiste à remplacer la matrice initiale par une matrice "éclatée":

- C'est un matrice booléenne dans les cellules ne figurent que des 0 ou des 1
- Chaque valeur i de chaque variable X a été associée à deux colonnes:

Va: i(X) = 1 et Vb: i(X) = 0

Pour un sujet donné si i(X) = 1 c'est à dire s'il présente la valeur i pour la variable X,

l'observateur porte 1 dans la colonne Va, et 0 dans la colonne Vb

Dans le cas contraire l'observateur porte 0 dans la colonne Va et 1 dans la colonne Vb.

De cette manière toute fonction définie ***sur les variables*** (en particulier la somme des valeurs) est constante.

Les proximités ne dépendent plus que des similarités formelles et pas de leur signification.

2A. Classer des questions en fonction des élèves

On peut considérer comme objet d'étude aussi bien l'ensemble des questions que l'ensemble des élèves et par conséquent avoir pour objectif des études symétriques avec les même types de matrices:

Classer les épreuves suivant le succés des élèves conduit à l'étude de la somme des effectifs pour chaque colonne, puis à l'étude de la distribution marginale des variables ou ds questions.

Il n'y a guère d'autres fonctions d'agrégation que la somme car on considère que tous les sujets sont équivalents à priori, mais l'hypothèse ne doit pas être exclue: On peut rechercher les questions caractéristiques de tels types d'élèves, si on possède une telle classification à priori

2B. construire une typologie de questions ou de situations

Identifier des dépendances entre les questions ert les facteurs qui les regroupent

Il s'agit de dire dans quelle mesure les questions se ressemblent par le fait que les mêmes élèves les traitent de façon identique. Comment conviendrait il de les regrouper de ce point de vue.

Nous possédons des moyens de dire si deux variables covarient, ou sont corrélées, si elles ne sont pas indépendantes l'une de l'autre, ou si l'une détermine l'autre, ou si l'un implique l'autre. Les moyens de faire de telles déclarations dépendent de la structure mathématique des variables considérées (cf. Les stratégies de l'analyse statistique).

L'analyse de données permet d'identifier des facteurs qui représentent des corrélations entre un certain nombre de variables, et non plus seulement entre deux. Il s'agit d'éviter d'expliquer de telles liaisons de façon isolée en ignorant les facteurs qui peuvent agir en même temps sur deux variables sans que l'une influence directement la valeur de l'autre. (Cf exemple de la corrélation entre le taux de natalité et" le nombre de nids de cigogne dans les commues d'Alsace).

Dans la mesure ou la matrice représente bien l'ensemble des variables envisagées comme pertinentes, elle permettra de rechercher les facteurs généraux et de rejeter les conclusions trop particulières.

Dépendances isolées

Il y a une autre raison pour mettre en garde contre la recherche directe de liaisons particulières entre des variables isolées:

L'informatique moderne permet de traiter des ensembles très vastes de variables. Dans un modeste Micro PCAT Statitcf traite jusqu'à 60 variables. La matrice de corrélation comprend donc 1770 cellules. En supposant que ces 60 variables soient stochastiquement indépendantes, si le seuil de signification choisi est de O,O5 on peut estimer que 5 pour cent de ces 1770 valeur excéderont PAR HASARD la valeur seuil, c'est à dire environ 85. Aller chercher à postériori les cellules présentant ces

grandes valeurs est bien naturel, mais il est impossible de maintenir ici le pari qui est à la base du test d'hypothèse:

Lorsqu'on détermine A L'AVANCE 2 questions que l'on veut croiser et qu'on les trouve liées par un coefficient exceptionnel on examine l'alternative bien connue:

Soit les variables SONT indépendantes et dans cette hypothèse il y a moins de 5 chances sur cent d'obtenir un coefficient aussi élevé, soit les variables NE SONT PAS indépendantes (le risque dans ce cas n'est pas quantifiable);

et on conclut en rejetant l'hypothèse de l'indépendance, sur la base du risque encourru. Parcequ'il y avait effectivement un pari

AVANT de tirer à pile ou face il y a une chance sur deux d'obtenir chaque issue. Mais APRES une suite de tirages il est parfaitement erroné de prétendre que le fait d'obtenir pile au troisième essai nous apprend quelque chose sur les conditions du troisième essai, sous pretexte qu'on obtient pile alors qu'on avait seulement une chance sur deux d'obtenir pile à ce moment là!

La méthode consistera à prévoir à l'avance les cellules où des liaisons sont attendues et à tester l'hypothèse que les coefficients dans ces cellules sont particulièrement élevés, ou que le nombre des coefficients significatifs est significativement élevé dans ces cellules par rapport aux autres.

Construire un instrument d'évaluation résumé

Il s'agit de choisir un résumé de quelques questions qui représentent bien un ensemble plus vaste de questions.

Par exemple on a posé à un petit nombre d'élèves un ensemble de questions assez diversifié pour représenter les différents aspects d'une connaissance. Il n'est pas possible ou il serai trop cher de poser un tel questionnaire à un échantillon représentatif d'élèves. Mais il est possible de choisir quelques questions telles que les réponses obtenues à celles-ci permettraient de deviner sans grande incertitude quels auraient été les comportements des élèves sur l'ensemble des autres.

C'est le mode canonique d'élaboration d'un questionnaire. Il n'est pas toujours utilisé, ce qui rend les interprêtations parfois tendancieuses

L'analyse de données permet assez facilement de construire de tels résumés.

3 Regrouper les comportements suivants les questions et les élèves

3. Les méthodes

Le choix dépend de la structure à priori des données

Nous allons conjuguer plusieurs méthodes, mais toutes ne s'appliquent pas(à toutes les sortes de données.

Le choix de la structure à priori des données a son importance, nous l'avons vu en fonction du but de l'étude. Il est important aussi pour déterminer le genre de distances utilisables et pour les methodes à utiliser.

Dans un cas assez général, un ensemble de données peut se présenter sous forme d'un tableau dont les n lignes sont des objets observés et les p colonnes représentent des variables dont les valeurs sont les observations faites à propos de ces objets. Chaque case du tableau porte par exemple un nombre qui peut représenter une mesure, un rang, un effectif, une présence ou une absence... Le tableau est dit table de contingence si les valeurs portées sont des effectifs.

On peut alors considérer ce tableau de façons très différentes, par exemple:

- a) comme un simple ensemble amorphe de nxp points pris dans R,
- b) comme un point unique de R^{n+p} (l'espace sur R de dimension n+p),
- c) comme p vecteurs de R^n représentant les p variables,
- d) ou comme n vecteurs de R^p représentant les n observations,
- e) comme une application de R^{n-1} dans R
- f) ou comme une application vectorielle de R^k dans R^{p-k}

etc...

En appliquant notre premier principe (cf Stratégies de l'analyse statistiques) à la modélisation de ce tableau suivant ces différentes structures à priori on engendre des types d'analyses statistiques différentes, comme par exemple ceux qui suivent:

L'Analyse en Composantes Principales. (ACP)

Dans le cas général du point d) les sujets sont des points dans l'espace R^p, dont les axes représentent chacun un des caractères observés. La méthode est la même que ci dessus. Elle consiste à rechercher les plans principaux déterminés par les axes d'inertie et le centre de centre de gravité du nuage pour le représenter par les meilleures projections.

Elle permet en outre de préciser dans quelle mesure les variables sont corrélées ou liées entre elles.

Dans l'interprêtation statistique les axes principaux seront nommés "facteurs principaux".

Il n'est pas exclu en théorie d'utiliser l'A.C.P. pour étudier le point de vue c), mais les corrélations entre sujets sont plus délicates à interprêter.

NB L'analyse en composante principale ne PEUT être effectuée que sur des tableaux composés de variables QUANTITATIVES. Cela ne signifie pas que le programme ne fonctionnera pas sur un tableau qualitatif dont les valeurs sont codées par des nombres (le programme ignore ce que représentent ces nombres) mais que les corrélations ne pourront pas être interprêtées. Il ne faudra pas tenter de le faire. Les liaisons qui apparaîtraient dans ces conditions seraient des artéfacts.

Exemple par conséquent, l'analyse qui figure ci après du tableau CONINST permettra d'illustrer ce propos

L'analyse factorielle des correspondances.(AFC)

Elle suit le même principe mais la distance utilisée est celle du X^2 (légèrement différente de $d_4(a, E)$ donnée plus haut) ce qui permet de mieux traiter le cas d'une matrice d'incidence (variables booléennes).

Cette distance est la suivante:

$$D_8(V_i, V_j) = \sum_k \frac{S}{S_k} \cdot \left(\left|\frac{S_{ki}}{S_i} - \frac{S_{kj}}{S_j}\right|\right) \qquad S = \sum_k S_k$$

où Sk est la somme de la ligne k (valeurs correspondant au sujet k), Si et Sj les sommes des colonnes correspondant aux variables V_i et V_j, S_{ki} et S_{ki} les valeurs des variables Vi et Vj observées pour le sujet k et S la somme des valeurs du tableau

$$S = \sum_k S_k$$ (voir exercice 6).

De plus les sujets et les variables peuvent être placés dans un même espace. Il est possible alors d'interprêter les proximités des sujets entre eux, celle des variables entre elles, et celle des sujets avec les variables.

Elle traite ainsi en même temps le point de vue c) et d).

Nous avons vue qu'une variable nominale (qualitative) non booléenne peut toujours être "éclatée" et être l'objet d'une analyse factorielle des correspondances simples.

Cependant il est plus commode d'utiliser une analyse factorielle des correspondances multiples.

L’analyse hiérarchique (classification automatique)

L’analyse canonique (AC)

L'analyse canonique prend le point de vue f). Elle permet de considérer un certain nombre k de variables, comme explicatives des autres variables (causes par exemple). L'espace R^n est décomposé en l'espace engendré par ces k variables et leur espace supplémentaire.

Chacune des p-k autres variables se projette sur ces deux sous espaces orthogonaux. Celles dont la composante dans l'espace explicatif est nulle sont non expliquées (ou indépendantes). Celles dont la composante dans l'espace orthogonal à l'espace explicatif est nulle sont totalement déterminées (c'est à dire explicables, calculables, déterminées par les k "causes"). Les autres sont partiellement dépendante des k premières.

Cas particulier: l'analyse de régression linéaire, où k = n-1 et où est pris le point de vue e).

Nous en avons vu un exemple au chapitre précédent et nous reviendrons concrêtement sur cette méthode au chapitre 6.

L’analyse implicative.

En fait nous remplacerons l'analyse canonique par une analyse basée non plus sur une distance symétrique, mais sur des dépendances dissymétriques.

Autres méthodes

Le premier principe s'applique à des cas plus "modestes" de modélisation, en particulier à toutes celles qui font appel à des présentations graphiques (au sens général de BERTIN). L'usage raffiné de divers types de diagrammes pour la représentation des données et de leurs relations permet parfois de créer des effets inattendus. La difficulté de contrôler le code de la représentation et de conserver une idée de l'importance relative des relations mises en évidence est l'écueil principal de ces méthodes empiriques et suggestives.

*