MATHÉMATIQUES ET SCIENCES HUMAINES

H. ROUANET
D. LÉPINE

Note méthodologique. Statistiques de groupe, groupes d'observations

Mathématiques et sciences humaines, tome 41 (1972), p. 31-36
<http://www.numdam.org/item?id=MSH_1972__41__31_0>

Math. Sci. hum. (11e année, n° 41, 1973, p. 31-36)

NOTE MÉTHODOLOGIQUE

STATISTIQUES DE GROUPE, GROUPES D'OBSERVATIONS

par

H. ROUANET [1] et D. LÉPINE [2]

RÉSUMÉ

Cette note se situe dans le cadre d'un travail méthodologique en cours, sur la formalisation des procédures élémentaires d'analyse de données [3]. *Nous nous bornerons ci-dessous (§ 1) à rappeler les notions de base nécessaires à la compréhension de cette note.*

En sciences humaines, l'expression de "groupe d'observations" est souvent utilisée comme quasi-synonyme de "famille d'observations", mais avec une connotation plus restrictive : un groupe est une famille d'observations considérée comme "formant un tout" sur laquelle on doit a priori *travailler de façon séparée, quitte à se livrer ensuite à des "regroupements". Plus précisément encore, quand on parle de groupe, c'est qu'on traite les observations "de façon symétrique" : cette remarque va nous conduire à proposer une définition formelle de la notion de* groupe d'observations *compatible avec cet usage ; dans le même mode de pensée, nous définirons auparavant la notion de* statistique de groupe, *dont nous donnerons une caractérisation.*

SUMMARY

This note is part of a current program of methodological research on the formalization of elementary procedures in data processing [3]. *Here (§1) we will only recall the basic notions needed for the understanding of this note.*

In the human sciences, the term "group of observations" is frequently used as quasi-synonymous with "family of observations", but with a more restricted connotation : a group is a family of observations considered to "constitute a whole", on which one should a priori *work separately, with the possibility of subsequently proceeding with "regroupings". More precisely, one refers to groups when one treats observations "in a symmetrical manner" : this remark leads us to a formal definition of the notion of* group of observations *compatible with this usage ; in the same context, we begin by defining the notion of* group statistics, *which we also characterize.*

---

1. UER de Mathématiques, Logique Formelle et Informatique, Université René Descartes (Paris).

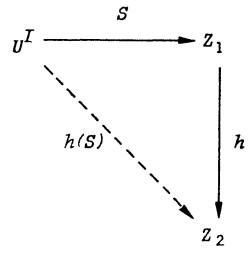
2. Laboratoire de Psychologie expérimentale et comparée, EPHE (3e Section) et Université René Descartes, Associé au CNRS (Paris).

3. *Cf.* en particulier référence [1] / *Cf.* in particular reference [1].

I. NOTIONS DE BASE

Étant donné un ensemble d'observations possibles, ou *espace d'observation* U, nous appellerons *protocole de support* I à valeurs dans U toute famille (x_i) d'éléments de U indexée par l'ensemble I, c'est-à-dire toute application de I dans U. Dans la suite on supposera que I est un ensemble *fini* d'indices. D'autre part, tous les protocoles que nous considérerons auront même espace d'observation : l'ensemble des protocoles possibles, ou *espace des protocoles*, sera donc U^I. A partir de la notion de protocole on définit celle de *sous-protocole* (*cf.* [1] pour une définition générale) ; dans ce texte nous n'utiliserons que les sous-protocoles qui sont des *restrictions* d'un protocole à une partie J du support I. Tout sous-protocole est un protocole.

Enfin, une *statistique* est définie comme une application [1] dont l'ensemble de départ est U^I. En particulier, nous appellerons $i^{\text{ième}}$ *statistique élémentaire*, ou $i^{\text{ième}}$ *observation*, la statistique-projection $X_i : U^I \to U$ qui à tout protocole $x : I \to U$ associe la valeur x_i. La famille (x_i), $i \in I$, des statistiques élémentaires est une statistique, qu'on identifiera avec la statistique identique (celle qui à tout protocole associe ce protocole lui-même). Z_1 et Z_2 étant deux ensembles, si S est une statistique à valeurs dans Z_1 et h une application de Z_1 vers Z_2, l'application $h \circ S$ est une statistique (à valeurs dans Z_2) : on notera également cette statistique $h(S)$ et on dira qu'elle est *fonction de la statistique* S. Toute statistique est fonction de la famille des statistiques élémentaires, soit plus brièvement : est fonction des observations.



II. SOUS-PROTOCOLES INDUITS L'UN DE L'AUTRE

Deux sous-protocoles y' et y'' de même support J (où $J \subset I$) seront dits *induits l'un de l'autre* (par une permutation de leur support) s'il existe une permutation p de J telle que $y'' = y' \circ p$. Il est immédiat que la relation entre sous-protocoles de même support « être induits l'un de l'autre» est une relation d'équivalence sur l'espace de ces sous-protocoles, et également une relation d'équivalence sur l'espace des protocoles.

Si p est une permutation de J, on lui associe une permutation p^* de I définie par :

$$p^*(i) = p(i) \text{ pour } i \in J,$$
$$= \quad i \text{ pour } i \notin J.$$

Étant donné un protocole x de support I on appellera protocole induit de x par la permutation p de J (partie de I) le protocole $x' = x \circ p^*$.

III. DISTRIBUTION DES EFFECTIFS ET PERMUTATION DU SUPPORT : PROPRIÉTÉ FONDAMENTALE

A tout sous-protocole $y : J \to U$ on associe sa *distribution des effectifs*, c'est-à-dire la mesure sur U qui prend sur toute partie $\{u\}$ à un élément la valeur $n(u) = \delta(y^{-1}(u))$, où δ désigne la mesure sur I

1. Cette application est le plus généralement caractérisée par une règle opératoire permettant de « calculer » (au sens large) la valeur de la statistique pour le protocole considéré.

qui à toute partie de I associe son effectif ; la distribution des effectifs d'un sous-protocole est donc caractérisée par une application $n : U \to N$ (densité discrète de la distribution).

La distribution des effectifs d'un sous-protocole de support J peut être considérée comme la valeur prise par une statistique que, par extension de langage, on appellera également distribution des effectifs des sous-protocoles de support J. Cette statistique est caractérisée par une application de $U^I \to N^U$ que nous noterons E. La distribution des effectifs d'un sous-protocole y (restriction de x) ne dépend que de y ; par extension de notation nous la désignerons par $E(y)$.

Propriété

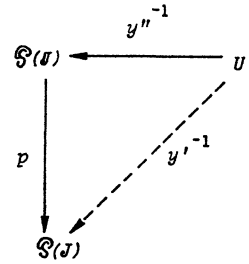
Pour que deux sous-protocoles de support J aient même distribution d'effectifs, il faut et il suffit qu'ils soient induits l'un de l'autre :

$$y'' = y' \cdot p \Leftrightarrow E(y') = E(y'').$$

Démonstration

1) Supposons qu'il existe une permutation p de J telle que $y'' = y' \circ p$, ou $y' = y'' \circ p^{-1}$. On en déduit, en désignant encore par p l'extension canonique de p à $\mathscr{P}(J)$ qui à tout $K \in \mathscr{P}(J)$ associe $p(K)$:

$$y'^{-1} = p \cdot y''^{-1}.$$



Pour tout $u \in U$: $y'^{-1}(u) = p(y''^{-1}(u))$

donc : $\delta(y'^{-1}(u)) = \delta(p(y''^{-1}(u)))$

$= \delta(y''^{-1}(u))$ car l'extension p est bijective.

L'égalité $\delta(y'^{-1}(u)) = \delta(y''^{-1}(u))$ exprime que y' et y'' ont même distribution d'effectifs.

2) Réciproquement, supposons que y' et y'' ont même distribution d'effectifs : pour tout $u \in U$, $\delta(y'^{-1}(u)) = \delta(y''^{-1}(u))$. On peut alors, pour tout u, construire une application bijective de l'ensemble $y''^{-1}(u)$ sur l'ensemble $y'^{-1}(u)$; en effectuant une telle construction pour chacune des classes de $J/_{y''} = \{ y''^{-1}(u) \mid u \in y''(J) \}$, on définit une permutation p de J.

Soit alors $i \in J$. Si $i \in y''^{-1}(u)$, de par la construction de p : $p(i) \in y'^{-1}(u)$, donc $y'(p(i)) = u$. Or : $u = y''(i)$. Donc : $y' \circ p = y''$.

Il résulte de cette propriété que les deux relations d'équivalence entre sous-protocoles « être induits l'un de l'autre » et « avoir même distribution d'effectifs » coïncident.

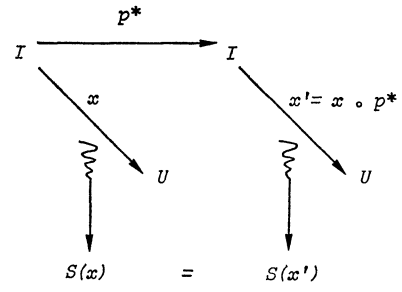
IV. STATISTIQUE DE GROUPE POUR UN SOUS-PROTOCOLE

Définition

Nous dirons qu'une statistique est une statistique de groupe pour les sous-protocoles de support J si elle est invariante par le groupe des permutations de ce support.

Si y est un sous-protocole de x, y' le sous-protocole induit de y par une permutation p de J, et x' le protocole induit de x par cette permutation p, on peut traduire la définition par la relation :

S statistique de groupe pour $J \Leftrightarrow (y' = y \circ p \Rightarrow S(x) = S(x'))$.



(S statistique de groupe pour (J)) $\Leftrightarrow S(x) = S(x')$

Formulation équivalente

Une statistique est une statistique de groupe pour les sous-protocoles de support J si elle est une fonction symétrique des observations de ce sous-protocole.

En général, si J est strictement inclus dans I, une statistique de groupe pour les sous-protocoles de support J n'est *pas* une statistique de groupe pour les protocoles de support I.

Exemples

Pour les sous-protocoles de support J :

— Les statistiques élémentaires (observations) X_i ne sont pas des statistiques de groupe ;

— Si $V \subset U$: $y^{-1}(V)$ n'est pas une statistique de groupe ; en revanche $\delta(y^{-1}(V))$ est une statistique de groupe ;

— Les statistiques : distribution des effectifs et distribution des fréquences sont des statistiques de groupe.

Pour un sous-protocole à valeurs dans **R**, les statistiques usuelles sont des statistiques de groupe: aussi bien les statistiques linéaires (moyenne, variance, etc.), dont la définition fait ressortir immédiatement qu'elles sont des fonctions symétriques des observations, que les statistiques non-linéaires : médiane, quantiles, $\max_{i \in J} X_i$, $\max_{i \in J} X_i - \min_{i \in J} X_i$, etc.

Si x est un protocole numérique, le protocole dérivé constitué par les moyennes de plusieurs sous-protocoles disjoints est une statistique de groupe pour chacun de ces sous-protocoles, mais n'est pas une statistique de groupe pour le protocole complet.

V. STATISTIQUE DE GROUPE, CARACTÉRISATION

Il résulte de la propriété fondamentale du § 3 la propriété caractéristique suivante des statistiques de groupe :

Pour qu'une statistique soit une statistique de groupe pour les sous-protocoles de support J, il faut et il suffit qu'elle soit fonction de la distribution des effectifs de ces sous-protocoles.

(En d'autres termes, une statistique est de groupe pour un sous-protocole si et seulement si sa valeur peut être calculée à partir de la distribution des effectifs de ce sous-protocole.)

Démonstration

1) Par définition :

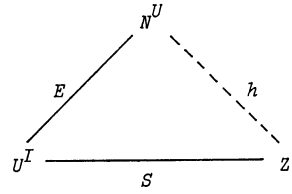
S statistique de groupe pour $(J) \Leftrightarrow (y'' = y' \circ p \Rightarrow S(x'') = S(x'))$ (§ 4).

2) $y'' = y' \circ p \Leftrightarrow E(y') = E(y'')$ (propriété § 3).

3) D'où par substitution dans (1) :

S statistique de groupe pour $(J) \Leftrightarrow (E(y') = E(y'') \Rightarrow S(x') = S(x''))$.

La relation $E(y') = E(y'') \Rightarrow S(x') = S(x'')$ exprime que si $S : U^I \to Z$, il existe $h : N^U \to Z$ tel que $S = h \circ E$: la statistique S est fonction de la statistique E.



De la propriété précédente, il résulte que si au lieu de se donner un sous-protocole (application), on se donne seulement la distribution des effectifs correspondante (mesure sur U), on ne pourra calculer que des statistiques de groupe pour ce sous-protocole.

VI. GROUPE D'OBSERVATIONS

1) *Vis-à-vis d'un ensemble de statistiques :*

Si un ensemble de statistiques est constitué de statistiques de groupe *pour un sous-protocole*, on dira que celui-ci est un groupe d'observations (relativement à l'ensemble des statistiques considérées).

2) *Vis-à-vis d'un modèle probabiliste satisfaisant à la condition de symétrie :*

On dira qu'un modèle probabiliste de l'espace des protocoles U^I satisfait à la condition de symétrie pour les sous-protocoles de support J si cette probabilisation est une fonction symétrique des observations de ces sous-protocoles. La fonction caractérisant la distribution des sous-protocoles (densité discrète, ou fonction de répartition, ou densité de probabilité selon le cas) est alors une fonction symétrique des x_i $(i \in J)$: les sous-protocoles équivalents de support J (au sens du § 3) ont alors la même probabilité (lorsque U est fini) ou la même densité de probabilité (lorsque par exemple, $U = \mathbf{R}$ ou $\mathbf{R}^k$) [1].

On vérifie que la condition de symétrie est équivalente à la condition suivante : *la distribution des sous-protocoles conditionnellement à une distribution d'effectifs est uniforme :* en d'autres termes, si la distribution d'effectifs est caractérisée par l'ensemble des valeurs $x^1, x^2, \ldots, x^k$ observées respectivement $n_1, n_2, \ldots, n_k$ fois (avec $\sum_{j=1}^{k} n_j = n$) chacun des $\binom{n}{n_1, n_2, \ldots, n_k}$ sous-protocoles possibles ayant cette distribution a la probabilité conditionnelle $\dfrac{1}{\binom{n}{n_1, n_2, \ldots, n_k}}$.

1. La propriété pour un modèle de satisfaire ou non à la condition de symétrie est souvent liée aux options sur le caractère aléatoire ou systématique des facteurs associés aux variables contrôlées (distinction classique en analyse de la variance).

On peut exprimer cette propriété en disant que la distribution des effectifs est *la statistique exhaustive minimale pour tous les modèles satisfaisant à la condition de symétrie.*

Un cas fréquent de modèles satisfaisant à la condition de symétrie est le modèle d'échantillonnage au hasard, dans lequel le sous-protocole est considéré comme la réalisation d'une famille de variables aléatoires indépendantes équidistribuées ; le schéma d'urne correspondant est celui des tirages non-exhaustifs. On remarque d'ailleurs que le modèle correspondant au schéma d'urne des tirages exhaustifs satisfait également à la condition de symétrie. Un contre-exemple usuel est celui d'un sous-protocole considéré comme la réalisation d'un processus, par exemple markovien.

Vis-à-vis d'un modèle probabiliste satisfaisant à la condition de symétrie pour les sous-protocoles de support J, on dira que le protocole observé de support J est un *groupe d'observations*, ce qui conduira à ne calculer, dans le cadre de ce modèle, que des statistiques de groupe pour ces sous-protocoles.

Remarque

Si on interprète le support J d'un sous-protocole comme un ensemble de numéros, une caractéristisation opérationnelle de la notion de groupe d'observations est que le numérotage des observations de ce sous-protocole est arbitraire.

BIBLIOGRAPHIE

[1] ROUANET, H. et LÉPINE, D., "Notions fondamentales d'analyse des données : Protocoles", à paraître.