REVUE DE STATISTIQUE APPLIQUÉE

PHILIPPE LAZAR

Partition d'un groupe hétérogène en sous-groupes homogènes

Revue de statistique appliquée, tome 14, n° 1 (1966), p. 39-43

<http://www.numdam.org/item?id=RSA_1966__14_1_39_0>

PARTITION D'UN GROUPE HÉTÉROGÈNE EN SOUS-GROUPES HOMOGÈNES

Philippe LAZAR*

Nombreuses sont les méthodes proposées dans la littérature pour mener à son terme la comparaison d'un groupe de k variables qu'un test global a révélées différentes ; LELLOUCH vient d'en donner un large aperçu synthétique en même temps qu'une classification [1]. On peut constater que, dans l'ensemble, ces méthodes sont relativement peu puissantes, parfois assez lourdes quant aux calculs, et surtout conduisent à des conclusions souvent difficiles à interpréter : le résultat final est en effet rarement simple, du type $(A = B) \neq (C = D = E) \neq (F = G)$. C'est que l'optique adoptée reste essentiellement démonstrative, et que, notamment, on veut continuer à limiter sévèrement le risque de 1ère espèce relatif à la comparaison de 2 ou plusieurs des k variables. Il nous a semblé que, dans certains problèmes, il pourrait être intéressant de se libérer partiellement de ces contraintes, en se plaçant dans une optique prospective, à la recherche d'idées nouvelles.

Lorsqu'on est amené en effet à pratiquer un test global d'homogénéité, par exemple une analyse de variance pour comparer k moyennes, c'est qu'il existe a priori entre ces moyennes une certaine symétrie. Si ce n'est pas le cas, on a bien souvent intérêt à adopter une stratégie autre qu'un test global. Il ne faut pas perdre de vue que l'objet essentiel du test global est avant tout de rompre cette symétrie apparente de l'ensemble des k variables, et que cette seule conclusion peut être en elle-même très riche.

Ainsi il peut être intéressant de savoir si par exemple un facteur comme la couleur, la taille, la forme de la cage peut jouer un rôle dans l'évolution des animaux soumis à une expérience : on placera pour cela des animaux dans des cages de couleurs, tailles, formes très diverses, et on fera un test global d'homogénéité. La conclusion essentielle sera acquise dès lors qu'on aura montré que les résultats sont ou ne sont pas globalement homogènes. Si l'on veut aller plus loin, et chercher par exemple quelles formes de cages donnent les résultats les plus différents, c'est presque d'un autre problème qu'il s'agit, et l'on conçoit qu'il puisse être parfaitement satisfaisant de se servir des résultats de la 1ère expérience non pour démontrer que telle forme conduit à des résultats différents de ceux que donne telle autre forme, mais seulement pour orienter les recherches futures, ce qui permet d'accepter, en connaissance de cause, des erreurs plus importantes. Dans cette perspective, nous avons cherché à nous raccrocher à l'un des modèles les plus simples à concevoir : classer objectivement et complètement les k valeurs hétérogènes en un nombre aussi petit que possible de sous groupes homogènes distincts.

---

* - Unité de Recherches Statistiques de l'Institut National de la Santé et de la Recherche Médicale.

PRINCIPE DE LA CLASSIFICATION

Le principe de cette partition repose sur la décomposition classique (exacte ou asymptotique) de tests globaux en sous-tests. Nous allons l'exposer dans le cas de k variables aléatoires m_i normales, de moyennes μ_i et de variances $\sigma_i^2 (i = 1 \ldots k)$.

Sous l'hypothèse nulle $[\mu_i = \mu \ (i = 1 \ldots k)]$, on sait qu'on peut estimer μ par

$$m = \frac{\Sigma w_i \ m_i}{\Sigma w_i}, \text{ de variance } \frac{1}{\Sigma w_i},$$

où

$$w_i = \frac{1}{\sigma_i^2},$$

et que

$$X = \Sigma w_i \ (m_i - m)^2 = \Sigma w_i \ m_i^2 - \frac{(\Sigma w_i \ m_i)^2}{\Sigma w_i}$$

est distribué selon une loi de χ^2 à (k - 1) d.l.

Sous n'importe quelle hypothèse alternative, cette expression est distribuée selon un χ^2 non centré. On peut préciser davantage : sous l'hypothèse H_j, par exemple, où les variables appartiennent à j sous-groupes distincts ($j \leq k$) on constate aisément que X peut se décomposer en 2 termes indépendants, l'un qui est un χ^2 non centré fondé sur les j valeurs distinctes, qu'on peut noter χ'^2, et l'autre un χ^2 à (k - j) d.l. qui exprime les fluctuations des variables autour de leurs moyennes respectives.

Par exemple si les valeurs m_1 et m_2 sont extraites de populations de moyenne μ_A; que m_3, m_4 et m_5 sont extraites de populations de moyenne μ_B, on a

$$X = \sum_{i=1}^{5} w_i \ m_i^2 - \frac{(\Sigma w_i \ m_i)^2}{\Sigma w_i}$$

d'où

$$X = \left[w_1 \ m_1^2 + w_2 \ m_2^2 - \frac{(w_1 m_1 + w_2 m_2)^2}{w_1 + w_2} \right]$$

$$+ \left[w_3 m_3^2 + w_4 m_4^2 + w_5 m_5^2 - \frac{(w_3 m_3 + w_4 m_4 + w_5 m_5)^2}{w_3 + w_4 + w_5} \right]$$

$$+ \left[w_A \ m_A^2 + w_B \ m_B^2 - \frac{(w_A m_A + w_B m_B)^2}{w_A + w_B} \right]$$

puisque

$$m_A = \frac{w_1 m_1 + w_2 m_2}{w_1 + w_2}$$

de variance

$$\frac{1}{w_A}$$

avec

$$w_A = w_1 + w_2$$

et que

$$m_B = \frac{w_3 m_3 + w_4 m_4 + w_5 m_5}{w_3 + w_4 + w_5}$$

avec

$$w_B = w_3 + w_4 + w_5$$

Le 3ème terme entre crochets est un χ^2 non centré, la somme des 2 premiers termes un χ^2 à 3 dl.

REGLE DE PARTITION

1) - on fait le test global. S'il est non significatif on s'arrête. S'il est significatif on continue.

2) - on essaie toutes les coupures en 2 sous-groupes. A chaque césure correspond une décomposition de X en 2 termes, l'un correspondant à l'hétérogénéité entre les sous-groupes, l'autre à l'homogénéité à l'intérieur des sous-groupes. La césure qui maximise l'un minimise l'autre puisque leur somme est constante. C'est celle-là qu'on choisit comme la "meilleure" césure en 2 sous-groupes. Reste à s'assurer qu'elle est satisfaisante, c'est-à-dire que le test d'homogénéité est non-significatif. On ne sert pour cela de la table de χ^2 à (k - 2) d.l. Si le test est non significatif, on s'arrête et on adopte cette césure. (La significativité de l'autre terme de la décomposition donne une indication sur la valeur réelle de cette partition). Si par contre le test d'homogénéité est significatif on continue.

3) - on essaie de novo toutes les coupures en 3 sous-groupes. A chaque césure correspond encore une décomposition de X en 2 termes additifs, l'un qui est le test d'hétérogénéité, l'autre la somme des tests d'homogénéité entre les 3 sous-groupes. La césure qui maximise l'un minimise l'autre etc....

DISPOSITION PRATIQUE

On classe les valeurs observées par ordre croissant. On peut éliminer d'emblée de nombreux essais manifestement inutiles. Les résultats des césures en 2 sous-groupes donnent des indications sur la marche à suivre pour 3 sous-groupes etc ... En pratique les calculs sont jamais très longs .... s'il n'y a pas trop de sous-groupes distincts.

CARACTERISTIQUES DE LA METHODE

On peut montrer aisément que la méthode proposée limite à αj la probabilité de trouver plus de sous-groupes qu'il n'en existe réellement, si on choisit le seuil de significativité αj pour le test de partition en j sous-groupes, j étant le nombre réel de sous-groupes. En effet supposons par exemple que H_3 soit vraie (3 sous-groupes). De deux choses l'une : ou bien le manque de puissance fait arrêter la décomposition aux étapes de 1 groupe homogène ou de 2 sous-groupes, ou bien on aborde l'étape 3 sous-groupes.

Dès lors, puisqu'on fait (au moins théoriquement) toutes les décompositions en 3 sous-groupes, on fait en particulier la "bonne", c'est-à-dire celle qui permet d'écrire

$$X = \chi'^2 + \chi^2_{[k-3]}$$

On teste l'homogénéité par le $\chi^2_{[k-3]}$ qui a α_3 chances d'être significatif.

- s'il est significatif, il reste possible qu'une autre césure conduise à un test non-significatif

- s'il est non significatif, il est possible qu'une autre césure conduise à une valeur encore plus faible, mais de toute façon on s'arrêtera à 3 sous-groupes.

Donc la probabilité de continuer à décomposer est limitée par α_3.

La probabilité de trouver trop de sous-groupes est donc raisonnablement limitée. Naturellement la constitution de ces sous-groupes ne sera pas toujours exacte, mais on conçoit que la probabilité d'une composition tout à fait aberrante soit faible. Cette question mériterait d'être précisée, mais est difficile. Quoi qu'il en soit les groupes ainsi constitués ont des chances raisonnables de refléter au moins partiellement la vérité.

Deux avantages non négligeables en contrepartie : toutes les variables sont classées de façon unique, et il n'est pas nécessaire de supposer l'égalité des variances σ_i^2, ce qui est bien commode dans les problèmes d'observation où on n'est justement pas maître de la précision de chaque variable.

EXEMPLE : APPLICATION AU PROBLEME DU SEX-RATIO

Il est bien connu que la fréquence des cancers n'est pas la même parmi les hommes et les femmes pour les organes communs aux deux sexes. On constate que le sex-ratio (rapport du nombre de cas masculins au nombre de cas féminins) est beaucoup plus élevé pour des organes comme par exemple la langue, l'œsophage, les bronches, pour lesquels on peut invoquer une étiologie "tabac" ou "alcool" - c'est-à-dire plus spécifiquement masculine, que pour des organes comme l'intestin, les reins, etc...

Il a paru intéressant de tester l'homogénéité du sex-ratio entre les sous-localisations d'un même organe, de façon à rechercher s'il pouvait exister des conditions étiologiques différentes selon ces sous-localisations. La méthode décrite ci-dessus a été appliquée à un ensemble de 65.000 cas de cancers [2], et a permis de dégager un certain nombre d'hypothèses étiologiques devant faire l'objet de vérifications plus directes (enquêtes). Il faut souligner, à propos de cet exemple, les raisons qui peuvent inciter à adopter un test global suivi d'une recherche de sous-groupes dans un esprit prospectif, et notamment :

- l'importance du test global qui permet de choisir entre l'hypothèse d'une étiologie unique et celle d'étiologies multiples.

- la difficulté de faire une classification a priori des sous-localisations [par exemple pour l'œsophage, c'est le tiers médian qui se détache du reste, alors que pour l'estomac c'est la partie haute (cardia)].

- la faible importance relative de mal classer une ou deux sous-localisations lorsqu'on est averti que cette erreur est possible - c'est-à-dire à condition de ne pas prendre pour argent comptant la classification établie.

- l'intérêt de ne pas être limité par une clause d'égalité des variances, les effectifs étant très différente dans les diverses sous-localisations.

EXTENSION A L'ANALYSE DE VARIANCE

On peut montrer très facilement que la fluctuation entre traitements peut se décomposer, à chaque étape, en 2 termes additifs de façon tout à fait analogue à ce qui précède, l'un relatif à l'hétérogénéité entre les j sous-groupes, l'autre à leur homogénéité interne. La méthode précédente pourra donc être facilement étendue à ce cas. Elle ne requiert pas l'égalité des effectifs de chaque groupe.

EXTENSION AUX TESTS D'HOMOGENEITE DES TABLEAUX DE CONTINGENCE PAR UN χ^2

Ici la décomposition du χ^2 n'est qu'asymptotique, mais en pratique suffisamment exacte pour qu'on puisse procéder de façon tout à fait analogue à ce qui précède.

CONCLUSION

La méthode proposée ne prétend pas apporter une solution démonstrative après qu'un test global ait permis d'établir l'hétérogénéité d'ensemble d'un groupe de k valeurs. Elle part du principe que si on a fait un test global c'est qu'a priori les variables jouaient un rôle symétrique, et que l'essentiel de la démonstration consistait dès lors à rompre cette apparente symétrie par des moyens suffisamment économiques (test global). Une fois l'homogénéité mise en cause, on peut adopter une attitude plus libre, qui, pour être pleinement efficace, doit déboucher sur des idées relativement simples destinées à introduire d'autres observations. La classification objective en un nombre minimal de sous-groupes qui a été exposée ici est une des façons de faire qui obéisse à cet impératif de clarté ; en contre-partie elle peut introduire des erreurs de classement, et il importe de s'en servir dans l'esprit prospectif sur lequel on a insisté ici, c'est-à-dire avec précautions.

[1] - J. LELLOUCH - Quelques Aspects du Problème des Comparaisons multiples. Revue de statistique appliquée (même numéro).

[2] - R. FLAMANT, O. LASSERRE, Ph. LAZAR, J. LEGUERINAIS, P. DENOIX & D. SCHWARTZ - Différences in Sex-Ratio According to Cancer Site and Possible Relationship With Use of Tobacco and Alcohol. Review of 65.000 Cases. - Journal of the National Cancer Institute - Vol. 32, n° 6 - June 1964.