MATHÉMATIQUES ET SCIENCES HUMAINES

BRIGITTE LE ROUX

Inférence combinatoire en analyse géométrique des données

Mathématiques et sciences humaines, tome 144 (1998), p. 5-14

<http://www.numdam.org/item?id=MSH_1998__144__5_0>

Math. Inf. Sci. hum., (36e année, n°144, *1998*, pp. 5-14)

INFÉRENCE COMBINATOIRE EN ANALYSE GÉOMÉTRIQUE DES DONNÉES

Brigitte LE ROUX[1]

RÉSUMÉ — *Dans cet article, on se propose de montrer comment, en analyse géométrique des données (analyse des correspondances, analyse en composantes principales ...) les statistiques descriptives utilisées comme aides à l'interprétation peuvent faire l'objet de procédures d'inférence combinatoire reposant sur des tests de permutation interprétés en termes de proportions d'échantillons plus extrêmes que les données, et qui prolongent directement la description statistique. Dans la première partie, on présente les tests de typicalité et d'homogénéité; dans la deuxième partie, on les applique aux variables principales de l'analyse des correspondances multiples, en prenant pour population l'ensemble des individus.*

SUMMARY — Combinatorial Inference in Geometric Data Analysis.

In this paper, we aim at showing how, in Geometric Data Analysis (Correspondence Analysis, Principal Component Analysis ...) descriptive statistics utilized as aids to interpretation can be used as combinatorial inference procedures based on permutation tests interpreted in terms of proportion of samples which are more extreme than the data. These procedures directly extend statistical description. In the first part, we will present typicality and homogeneity tests. In the second part, we will apply them to the principal variables provided by Multiple Correspondence Analysis, taking as the population the set of individuals.

INTRODUCTION

Par *Analyse géométrique des données* (AGD), nous entendons les méthodes multivariées dans lesquelles les données sont représentées sous forme de nuages de points dans des espaces multidimensionnels euclidiens, et où l'interprétation repose essentiellement sur les distances entre points : donc l'analyse des correspondances, mais aussi l'analyse en composantes principales (ACP), avec représentation du nuage des individus.

Les utilisateurs de l'AGD se limitent souvent aux procédures descriptives, malgré les nombreux travaux consacrés aux propriétés inférentielles de ces méthodes ; voir notamment les articles et chapitres pertinents de Lebart (1976), Lebart, Morineau, Warwick (1984), Greenacre (1984), de Leeuw & Van de Burg (1986), Daudin, Duby & Trécourt (1988), Saporta & Hatabian (1986), Gifi

[1]Centre de Recherche en Informatique de Paris 5 (CRIP5), laboratoire SBC, Université René Descartes, 45 Rue des Saints Pères, 75270 PARIS Cedex 06. E-mail : lerb@math-info.univ-paris5.fr

Je remercie M. Barbut de ses remarques sur une version antérieure de ce texte.

(1990), ter Braak (1992), Tenenhaus (1993), Alevizos & Morineau (1992,1993), Lebart, Morineau, Piron (1995). On peut penser que la réticence à utiliser les procédures inférentielles en AGD provient au moins en partie du fait que, pour les données traitées par les méthodes géométriques, les hypothèses probabilistes usuelles en inférence statistique — à savoir l'échantillonnage au hasard ou l'affectation au hasard des individus aux traitements (randomization) — sont rarement remplies, ce qui jette un doute sur la validité des conclusions inférentielles.

Des méthodes d'inférence statistique entièrement dépouillées d'hypothèses probabilistes, où les conclusions sont formulées en termes de *typicalité* d'un groupe d'observations ou d'*homogénéité* de plusieurs groupes d'observations, ont été proposées par Rouanet (1982), puis développées dans Rouanet et al. (1986), Rouanet, Bernard, Le Roux (1990), Rouanet et al. (1991,1998). Cette approche, appelée *inférence combinatoire*, repose sur des tests de permutation interprétés non pas en termes de probabilités mais en termes de proportions d'échantillons plus extrêmes que les données. L'inférence combinatoire, extension directe de la description statistique, apparaît donc comme privilégiée en AGD.

Le but de cet article est de montrer comment en AGD les statistiques descriptives utilisées comme aides à l'interprétation — voir en particulier Le Roux & Rouanet (1998) — peuvent faire l'objet de procédures d'inférence combinatoire prolongeant directement la description statistique.

L'organisation de l'article est la suivante. Dans la première partie, nous présentons d'abord la situation de base de l'inférence combinatoire (1.1), puis successivement, le test de typicalité pour comparer une moyenne à une valeur de référence (1.2), le test d'homogénéité pour comparer deux moyennes (1.3), le test d'homogénéité pour plusieurs moyennes (1.4), enfin le test de non-corrélation (1.5). Nous donnons ensuite des formules pour les méthodes approchées (1.6 & 1.7), puis nous établissons le lien entre inférence combinatoire et formulations probabilistes, avec le test du hasard (1.8). La deuxième partie est consacrée à l'Analyse des Correspondances Multiples (ACM). Nous rappelons d'abord les formules des statistiques descriptives du nuage des modalités (2.1), puis du nuage des individus (2.2). Nous donnons ensuite des formules pour les tests approchés (2.3).

1. — INFÉRENCE COMBINATOIRE

1.1. *Situation de base*

On se donne une *population de référence*, qui sera, dans la suite, une population numérique finie, de taille N, de moyenne μ et de variance σ^2. Dans le test de *typicalité*, on considère un *groupe d'observations*, de taille n et de moyenne m, on compare la moyenne m du groupe à la moyenne de la population de référence. Dans les tests d'*homogénéité*, on considère une *partition de la population* en K classes $k \in K$ (avec $\mathrm{K} \geq 2$)[2] d'effectifs $(n_k)_{k\in K}$ et de moyennes $(m^k)_{k\in K}$, et on compare entre elles certaines de ces moyennes (éventuellement toutes).

1.2 *Comparaison d'une moyenne à une valeur de référence : test de typicalité*

Etant donné un groupe d'observations de taille n et de moyenne m, le test combinatoire de comparaison de la moyenne m à la moyenne de référence μ, que l'on appellera *test de typicalité* du groupe d'observations vis-à-vis de la population selon la moyenne, est défini comme suit (cf. Rouanet & al, 1990, Chap. IV, p. 89-106).

On définit d'abord l'*espace des échantillons* comme l'ensemble $\mathcal{X}$ des $\binom{N}{n}$ sous-ensembles à n éléments de la population. A chaque échantillon $x \in \mathcal{X}$, on associe sa moyenne $M(x)$, d'où la

[2] On notera par la même lettre l'ensemble et son cardinal.

statistique Moyenne définie comme l'application M telle que :

$$\begin{array}{rccl} M: & \mathcal{X} & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & M(x) = m \end{array}$$

On définit ensuite la *distribution combinatoire* de la statistique M comme l'application de $M(\mathcal{X})$ dans l'intervalle $[0\,,\,1]$ qui à tout $m \in M(\mathcal{X})$ associe $P(M = m)$, proportion des échantillons pour lesquels M prend la valeur m.

$$\begin{array}{ccc} M(\mathcal{X}) & \longrightarrow & [0\,,\,1] \\ m & \longmapsto & P(M = m) \end{array}$$

Enfin, à partir de la distribution de M, on *situe le groupe d'observations étudié* par rapport aux échantillons. Lorsque $m > \mu$, on détermine $P(M \geq m) = p_{sup}$, proportion des échantillons dont la moyenne est supérieure ou égale à m. On prend la proportion p_{sup} comme *indice de typicalité* du groupe d'observations selon la moyenne : plus p_{sup} est petit, plus le groupe d'observations est atypique (à droite). Etant donné un seuil–repère α $(0 \leq \alpha < 0.5)$, si $p_{sup} \leq \alpha$, le groupe d'observations sera dit *atypique à droite* de la population au seuil α (unilatéral). Lorsque $m < \mu$, on raisonne de même en remplaçant p_{sup} par $P(M \leq m) = p_{inf}$.

Remarque. Le groupe d'observations dont on étudie la typicalité peut être un sous–ensemble (échantillon) de la population de référence.

1.3. *Comparaison de deux moyennes : test d'homogénéité*

Etant donnée une partition de la population de référence en K classes $k \in K$ $(\mathrm{K} \geq 2)$, on considère deux classes k et k' de la partition, d'effectifs n_k et $n_{k'}$, et de moyennes m^k et $m^{k'}$, et on s'intéresse à la différence (écart orienté) $m^k - m^{k'}$, noté $d^{kk'}$. Le test combinatoire de comparaison des moyennes m^k et $m^{k'}$, que l'on appellera *test d'homogénéité* des classes k et k', est alors défini comme suit (cf. Rouanet & al, 1990, chap. V, p. 115-130).

On définit d'abord l'*espace des protocoles* (couples d'échantillons) comme l'ensemble $\mathcal{X}_2$ des $N!/\left(n_k!n_{k'}!(N - n_k - n_{k'})!\right)$ couples de sous–ensembles disjoints de la population d'effectifs n_k et $n_{k'}$. A chaque couple $(x, x') \in \mathcal{X}_2$ on associe la différence d entre les moyennes $M(x)$ et $M(x')$, d'où la *statistique Différence* définie comme l'application D telle que :

$$\begin{array}{rccl} D: & \mathcal{X}_2 & \longrightarrow & \mathbb{R} \\ & (x, x') & \longmapsto & D(x, x') = M(x) - M(x') = d \end{array}$$

On définit ensuite la *distribution combinatoire* de la statistique D comme l'application de $D(\mathcal{X}_2)$ dans $[0\,,\,1]$, qui à tout $d \in D(\mathcal{X}_2)$ associe la proportion des protocoles tels que $P(D = d)$.

$$\begin{array}{ccc} D(\mathcal{X}_2) & \longrightarrow & [0\,,\,1] \\ d & \longmapsto & P(D = d) \end{array}$$

Enfin, à partir de la distribution de D, on *situe le couple de classes* (k, k') par rapport aux protocoles. Si $m^k > m^{k'}$, on détermine $P(D \geq d^{kk'}) = p_{sup}$, proportion des couples dont la différence des moyennes d est supérieure ou égale à $d^{kk'}$. On prend la proportion p_{sup} comme *indice d'homogénéité* des classes k et k' selon leurs moyennes : plus p_{sup} est petit, plus les classes k et k' sont *hétérogènes*. Si $p_{sup} \leq \alpha$, on dit que les classes k et k' sont hétérogènes, la classe k étant de moyenne supérieure à celle de la classe k', au seuil α (unilatéral). Si $m^{k'} > m^k$, on raisonne de même à partir de $p_{inf} = P(D \leq d^{kk'})$.

Propriété. Considérons une classe k de la partition. La comparaison de la moyenne m^k à la moyenne de la classe complémentaire selon le test d'homogénéité est équivalente à la comparaison

de m^k à μ selon le test de typicalité. En effet, si on note P_1 les proportions d'échantillons (dans l'espace $\mathcal{X}$) et P_2 les proportions de protocoles (dans l'espace $\mathcal{X}_2$), on déduit de la relation $D = \frac{N}{N-n_k}(M-\mu)$: $P_1(M \geq m) = P_2(D \geq d)$, ce qui exprime l'équivalence des deux tests.

1.4. *Comparaison de plusieurs moyennes : test d'homogénéité de plusieurs classes*

Soit K' classes ($2 \leq K' \leq K$) d'effectifs $(n_k)_{k\in K'}$ et de moyennes $(m^k)_{k\in K'}$. Le test combinatoire de comparaison des moyennes $(m^k)_{k\in K'}$, que l'on appellera *test d'homogénéité des classes* $k \in K'$, est défini comme suit.

On définit l'*espace des protocoles* $\mathcal{X}_{K'}$ comme l'ensemble des $N!/\left((N-N')!\prod_{k\in K'} n_k!\right)$ K'-uplets d'échantillons $(x_k)_{k\in K'}$ de la population de taille N, deux à deux disjoints et d'effectifs $(n_k)_{k\in K'}$. A chaque protocole, on associe la variance inter-classes (variance des moyennes des classes pondérées par $(n_k)_{k\in K'}$), que l'on peut écrire comme moyenne pondérée des carrés des différences entre les moyennes des couples de classes :

$$v = \frac{1}{2}\sum_{k\in K'}\sum_{k'\in K'} \frac{n_k n_{k'}}{N'^2}(M(x_k) - M(x_{k'}))^2 \quad \text{avec} \quad N' = \sum_{k\in K'} n_k \tag{1}$$

On en déduit la *statistique Variance* définie comme l'application V telle que :

$$\begin{array}{rccc} V : & \mathcal{X}_{K'} & \longrightarrow & \mathbb{R} \\ & (x_k)_{k\in K'} & \longmapsto & V\left((x_k)_{k\in K'}\right) \end{array}$$

A partir de la distribution de V, on *situe les* K' *classes* par rapport aux protocoles de $\mathcal{X}_{K'}$. On détermine $P(V \geq v) = p$, et on prend la proportion p comme indice d'homogénéité (selon la variance inter-classes) : plus p est petit, plus les classes $k \in K'$ sont hétérogènes. Les K' classes sont dites hétérogènes au seuil α si $p < \alpha$.

Cas particuliers

- pour $K' = 2$, on a la relation $V = \dfrac{n_k\, n_{k'}}{(n_k + n_{k'})^2}\, D^2$.

  En supposant pour fixer les idées $d \geq 0$, on a $p = P(D \geq d) + P(D \leq -d)$, donc $p \geq p_{sup}$. Lorsque $n_k = n_{k'}$, la distribution de D est symétrique, et l'on a $P(D \leq -d) = P(D \geq d)$, d'où $p = 2p_{sup}$.

- pour $K' = K$, on a le test d'*homogénéité d'une partition* (homogénéité globale). L'espace des protocoles $\mathcal{X}_K$ s'obtient alors par permutation des observations entre les K classes, le test d'homogénéité avec son cadre d'interprétation combinatoire s'apparente au test de permutation classique de Pitman (1937) de comparaison de moyennes.

Les tests de typicalité et d'homogénéité sont applicables à des classes obtenues par regroupement de classes de la partition, en prenant pour moyenne du regroupement la moyenne pondérée des classes regroupées.

Homogénéité partielle et inférence spécifique

Si $K' < K$ (et $N' < N$) (homogénéité partielle), on pourra prendre comme population de référence la population constituée par la réunion des K' classes considérées ; ce faisant nous dirons qu'on procède à une *inférence spécifique.*

1.5. *Test d'indépendance (non-corrélation)*

Un test apparenté au test d'homogénéité est le test (combinatoire) d'indépendance. Supposons que, sur l'ensemble définissant la population de taille N, on dispose d'une autre variable

numérique y, et qu'on s'intéresse à l'écart à l'indépendance — par exemple, au coefficient de corrélation linéaire r — entre cette variable et la variable de référence x. On définit l'espace des $N!$ protocoles en considérant tous les appariements possibles entre les N valeurs de x et les N valeurs de y, d'où la distribution combinatoire de la statistique R (coefficient de corrélation), à partir de laquelle on situe la valeur observée r, donc en considérant $P(R > r)$ (si $r > 0$). Ici encore, le test combinatoire s'apparente au test classique de permutation du coefficient de corrélation entre variables numériques. On démontre facilement que le test d'homogénéité d'une partition en deux classes est équivalent au test de non-corrélation obtenu en prenant comme coefficient de corrélation le coefficient point-bisérial associé à la dichotomie.

1.6. *Caractéristiques des distributions d'échantillonnage*

Les résultats qui suivent ou sont classiques ou découlent de la formule $\mathrm{Cov}(m^k, m^{k'}) = -\frac{\sigma^2}{N-1}$ (pour $k \neq k'$), formule fondamentale en théorie de l'échantillonnage dans une population finie ; voir par exemple Kendall & Stuart (1976, Vol. 3, p. 173, Vol 2, p. 492), ou Freedman & al. (1991, p. A22).

- *Typicalité*

  $\mathrm{Moy}\, M = \mu$; $\mathrm{Var}\, M = \dfrac{\sigma^2}{n} \times \dfrac{N-n}{N-1}$

- *Homogénéité de deux classes*

  $\mathrm{Moy}\, D = 0$; $\mathrm{Var}\, D = \sigma^2 \times \dfrac{N}{N-1} \times \left(\dfrac{1}{n_k} + \dfrac{1}{n_{k'}}\right)$

  En particulier, si les deux classes forment une *partition*, on a, en posant $f_k = n_k/n$:

  $\mathrm{Var}\, D = \dfrac{\sigma^2}{N-1} \times \dfrac{1}{f_k(1-f_k)}$

- *Homogénéité de plusieurs classes*

  $\mathrm{Moy}\, V = (\mathrm{K}' - 1) \times \dfrac{\sigma^2}{N-1} \times \dfrac{N}{N'}$ en posant $N' = \sum_{k \in K'} n_k$.

  En particulier, pour une partition de la population ($\mathrm{K}' = \mathrm{K}$), on a :

  $$\mathrm{Moy}\, V = (\mathrm{K} - 1)\frac{\sigma^2}{N-1}$$

  D'où en définissant $\eta^2 = V/\sigma^2$ (variance inter / variance totale) : $\mathrm{Moy}\, \eta^2 = \frac{\mathrm{K}-1}{N-1}$

- *Non-corrélation*

  $\mathrm{Moy}\, R = 0$ et $\mathrm{Var}\, R = \dfrac{1}{N-1}$

1.7. *Méthodes approchées*

Typicalité

En ajustant la distribution combinatoire de M par une *distribution normale* de même moyenne et de même variance, et en prenant comme statistique de test l'écart réduit T, avec[3] :

$$T = \frac{M-\mu}{\sqrt{\mathrm{Var}\, M}} = \frac{M-\mu}{\sigma}\sqrt{n\,\frac{N-1}{N-n}} \qquad (2)$$

[3] Voir Lebart & al (1995), valeurs-tests, p. 181-183.

alors, la distribution approchée de T est une distribution normale réduite. Si t désigne la valeur prise par T pour $M = m$, le seuil observé est approché par $p(T \geq t)$ (lorsque $m > \mu$), ou $p(T \leq t)$ (lorsque $m < \mu$).

Homogénéité de deux classes

En ajustant la distribution combinatoire de D par une distribution normale de même moyenne et de même variance, et en prenant comme statistique de test l'écart réduit T défini par :

$$T = \frac{D}{\sqrt{\text{Var } D}} = \frac{D}{\sigma} \times \sqrt{\frac{N-1}{N}} \frac{1}{\sqrt{\frac{1}{n_k} + \frac{1}{n_{k'}}}} \tag{3}$$

alors, la distribution approchée de T est une distribution normale réduite. Si t désigne la valeur prise par T pour $D = d$, le seuil observé $P(D \geq d)$ est approché par $p(T \geq t)$ (lorsque $d > 0$)) ou par $p(T \leq t)$ (lorsque $d < 0$).

La distribution approchée de $T^2 = \frac{D^2}{\sigma^2} \frac{N-1}{N} \frac{1}{\frac{1}{n_k} + \frac{1}{n_{k'}}} = \frac{V}{\sigma^2} \times \frac{N-1}{N} \times (n_k + n_{k'})$ est celle d'un χ^2 à 1 d.l. En particulier, si les deux classes forment une *partition*, on a :

$$T = \frac{D}{\sigma} \sqrt{N-1} \sqrt{f_k(1-f_k)} = \sqrt{N-1}\, r_{\cdot bis} \tag{4}$$

où $r_{\cdot bis} = \frac{D}{\sigma} \sqrt{f_k(1-f_k)}$ est la corrélation point-bisériale entre la variable étudiée et la variable indicatrice de la classe k.

Homogénéité de plusieurs classes

Si on ajuste la distribution de V par celle d'un χ^2 calibré à $\mathtt{K}' - 1$ degrés de liberté, de sorte que Moy V soit égale à la moyenne de ce χ^2 calibré, on pourra prendre comme statistique de test la statistique T^2 définie par :

$$T^2 = \frac{V}{\frac{\sigma^2}{N-1} \frac{N}{N'}} = \frac{N-1}{N} N' \times \frac{V}{\sigma^2} = (N-1)\frac{\text{Ctai}}{\sigma^2} \tag{5}$$

où $\text{Ctai} = V\frac{N'}{N}$ désigne la contribution intra des points moyens des classes à la variance σ^2. La distribution de T^2 est celle d'un χ^2 à $(\mathtt{K}' - 1)$ d.l. Le seuil observé $p = P(V \geq v)$ est alors approché par $p\left(\chi^2 > (N-1)\frac{v}{\sigma^2}\frac{N'}{N}\right)$.

Homogénéité globale. Si on considère toutes les classes de la partition ($\mathtt{K}' = \mathtt{K}$ et $N' = N$), alors : $(N-1)V/\sigma^2 = (N-1)\eta^2$ est distribué χ^2 à $\mathtt{K} - 1$ d.l.

Homogénéité partielle et inférence spécifique. Si $\mathtt{K}' < \mathtt{K}$ (et $N' < N$) (homogénéité partielle), le test d'homogénéité spécifique consistera à remplacer dans les formules précédentes, la variance σ^2 par la variance de la réunion des classes considérées.

Test de non-corrélation

Si on ajuste la distribution du coefficient de corrélation par une distribution normale de moyenne 0 et de variance $1/(N-1)$, la distribution approchée de $R\sqrt{N-1}$ est une distribution normale réduite.

1.8. *Interprétation probabiliste de l'inférence combinatoire : tests du hasard*

Les tests combinatoires de typicalité, d'homogénéité et de non corrélation peuvent s'interpréter comme des *tests du hasard.*

Ainsi pour le test de typicalité, l'hypothèse du hasard est que le groupe d'observations est assimilable à un échantillon au hasard de taille n de la population. Sous l'hypothèse du hasard, le seuil observé (indice de typicalité) $P(M \leq m)$ s'interprète comme la probabilité d'extraire

de la population un échantillon au moins aussi extrême que le groupe d'observations étudié. Si cette probabilité est petite — groupe d'observations atypique, résultat significatif — on rejette l'hypothèse du hasard : on conclut que le groupe d'observations étudié n'est pas assimilable à un échantillon au hasard ; on est alors fondé, pour l'interprétation des données, à prendre en compte l'écart de la moyenne du groupe à la moyenne de référence. Si cette probabilité n'est pas petite — groupe d'observations non atypique, résultat non-significatif — on ne peut pas rejeter l'hypothèse du hasard : on ne peut pas exclure que le groupe d'observations soit assimilable à un échantillon au hasard ; pour l'interprétation, il n'y a pas lieu de prendre en compte l'écart de la moyenne du groupe à la moyenne de référence.

De même pour les tests d'homogénéité, l'hypothèse du hasard est que le protocole observé est assimilable à un protocole au hasard de K'-uplets deux à deux disjoints et d'effectifs $(n_k)_{k\in K'}$. Sous l'hypothèse du hasard, le seuil observé (indice d'homogénéité) $P(D \leq d)$ s'interprète comme la probabilité d'extraire de la population un protocole au moins aussi extrême que le protocole étudié. Si cette probabilité est petite — protocole hétérogène, résultat significatif — on rejette l'hypothèse du hasard : on conclut que le protocole étudié n'est pas assimilable à un protocole au hasard ; on est alors fondé, pour l'interprétation des données, à prendre en compte les écarts entre les moyennes des classes. Si cette probabilité n'est pas petite — protocole non hétérogène, résultat non-significatif — on ne peut pas rejeter l'hypothèse du hasard : on ne peut pas exclure que le protocole soit assimilable à un protocole au hasard ; pour l'interprétation, il n'y a pas lieu de prendre en compte les écarts entre les moyennes des classes.

Le test du hasard, prolongement direct du test combinatoire, apparaît ainsi comme le premier stade de l'inférence probabiliste.

2. ANALYSE DES CORRESPONDANCES MULTIPLES

Nous reprenons les notations introduites dans Rouanet & Le Roux, 1993, p. 252. On note Q l'ensemble des questions et $\mathtt{Q}$ son cardinal ; on note K_q l'ensemble des modalités de la question $q \in Q$, et $K = \cup_{q\in Q} K_q$ l'ensemble de toutes les modalités. On désigne par I l'ensemble des individus et N son cardinal. On note n_k le nombre d'individus ayant choisi la modalité k et $f_k = n_k/N$ sa fréquence.

L'ACM conduit à définir deux nuages : le nuage des modalités et le nuage des individus.

2.1. *Nuage des modalités : statistiques descriptives*

Le *nuage des modalités* est constitué de points notés $(\mathrm{M}^k)_{k\in K}$, de poids $(p_k = f_k/\mathtt{Q})_{k\in K}$, de coordonnées principales sur un axe $(y^k)_{k\in K}$. On rappelle que :

1. la distance du point M^k au point moyen G du nuage est $\mathrm{GM}^k = \sqrt{\frac{1}{f_k} - 1}$;
2. la distance entre deux modalités (k, k') d'une même question est $\mathrm{M}^k\mathrm{M}^{k'} = \sqrt{\frac{1}{f_k} + \frac{1}{f_{k'}}}$;
3. l'angle θ_k $(0 \leq \theta_k \leq \pi)$ entre $\overrightarrow{\mathrm{GM}}^k$ et l'axe principal est défini par $\cos\theta_k = \frac{y^k}{\mathrm{GM}^k} = \sqrt{\frac{f_k}{1-f_k}}\, y^k$, d'où la qualité de représentation de k : $\cos^2\theta_k = \frac{f_k}{1-f_k}(y^k)^2$;
4. la contribution absolue de la modalité k, notée Cta_k, est égale à $\mathrm{Cta}_k = p_k(y^k)^2 = \frac{f_k}{Q}(y^k)^2$;
5. la contribution absolue de la question q, notée Cta_q, est, par définition, égale à la somme des contributions de ses modalités : $\mathrm{Cta}_q = \sum_{k\in K_q} \mathrm{Cta}_k$;

6. l'angle $\theta_{kk'}$ $(0 \leq \theta_{kk'} \leq \pi)$ entre $\overrightarrow{\mathrm{M}^k\mathrm{M}^{k'}}$ et l'axe, pour deux modalités k et k' d'une même question q (dipôle (k,k')), est défini par :
$$\cos\theta_{kk'} = \frac{y^{k'} - y^k}{\mathrm{M}^k\mathrm{M}^{k'}} = \frac{y^{k'} - y^k}{\sqrt{\frac{1}{f_k} + \frac{1}{f_{k'}}}},$$
d'où la qualité de représentation du dipôle : $\cos^2\theta_{kk'} = \dfrac{(y^k - y^{k'})^2}{\frac{1}{f_k} + \frac{1}{f_{k'}}}$;

7. la variance du dipôle (k,k') est $v_{kk'} = \frac{n_k n_{k'}}{(n_k+n_{k'})^2}(y^k - y^{k'})^2$;

8. la contribution absolue intra du dipôle (k,k'), notée $\mathrm{Ctai}_{kk'}$, est :
$$\mathrm{Ctai}_{kk'} = \frac{f_k + f_{k'}}{\mathtt{Q}} v_{kk'} = \frac{1}{\mathtt{Q}} \cos^2\theta_{kk'}$$

Propriétés

1. Pour une modalité k et le regroupement, noté $\widetilde{k}$, des modalités de la question q autres que k, les points M^k, G et $\mathrm{M}^{\widetilde{k}}$ sont alignés (avec $f_k\overrightarrow{\mathrm{GM}^k} = -(1-f_k)\overrightarrow{GM^{\widetilde{k}}}$). On a (en projection sur l'axe) :
$$\cos^2\theta_k = \cos^2\theta_{\widetilde{k}} = \cos^2\theta_{k\widetilde{k}} = v_{k\widetilde{k}} \quad \text{et} \quad \mathrm{Ctai}_{k\widetilde{k}} = \tfrac{1}{\mathtt{Q}}\cos^2\theta_k$$
Le coefficient $\cos\theta_k$ est égal à la corrélation entre la variable sur I (ensemble des individus) indicatrice de k et la variable principale sur I.

2. A un sous–ensemble K' $(K' \subseteq K_q)$ de modalités d'une question q, d'effectif $N' = \sum_{k\in K'} n_k$, on associe les K' coordonnées $(y^k)_{k\in K'}$, dont on notera $v_{K'}$ la variance ; alors la contribution intra des modalités $(k)_{k\in K'}$, notée $\mathrm{Ctai}_{K'}$, est $\mathrm{Ctai}_{K'} = \dfrac{1}{\mathtt{Q}}\dfrac{N'}{N}v_{K'}$.

   Cas particulier $K' = K_q$

   La contribution intra des K_q modalités de la question q est égale à Cta_q et, si v_q désigne la variance de la question sur l'axe ($v_q = \mathrm{Var}(y^k)_{k\in K_q} = \sum_{k\in K_q} f_k(y^k)^2$), on a : $\mathrm{Cta}_q = \frac{1}{\mathtt{Q}}v_q$.

2.2 *Nuage des individus : statistiques descriptives*

Le *nuage des individus* est constitué de N points $(\mathrm{M}^i)_{i\in I}$, de poids $(p_i = 1/N)_{i\in I}$, de coordonnées principales sur un axe $(y^i)_{i\in I}$. La variable principale a pour moyenne $\mu = 0$ et pour variance la valeur propre, que l'on notera ici σ^2.

Toute question q de modalités $k \in K_q$ définit une partition des N individus en K_q classes : le sous–nuage des individus ayant choisi la modalité k a pour point moyen le point moyen modalité $\overline{\mathrm{M}}^k$, de poids $f_k = n_k/N$ et de coordonnée principale $\overline{y}^k$.

Propriétés

1. La coordonnée principale du point moyen modalité $\overline{\mathrm{M}}^k$ est $\overline{y}^k = \sigma y^k$.

2. Soit $K' \subseteq K_q$. La variance des K' points moyens modalités $(\overline{\mathrm{M}}^k)_{k\in K'}$ est égale à $\sigma^2 v_{K'}$ et la contribution intra est égale à $\mathtt{Q}^2\sigma^2\,\mathrm{Ctai}_{K'}$.

3. Pour la partition du nuage des individus par une question q $(K' = K_q)$, on définit le rapport de corrélation η_q^2, égal au rapport variance inter/ variance totale, également appelé coefficient de discrimination de la question q, et l'on a, pour chaque axe, les relations :
$$\eta_q^2 = v_q = \mathtt{Q}\,\mathrm{Cta}_q$$

2.3. *Tests combinatoires en ACM*

Chaque problème de test est posé dans le nuage des individus et étudié pour chaque axe principal. Les statistiques de test seront exprimées en fonction des statistiques descriptives du nuage des modalités. On applique les formules du paragraphe 1.7 en prenant comme population l'ensemble I des N individus et pour variable une variable principale sur I.

Typicalité d'une classe

Pour étudier la typicalité de la classe k selon la moyenne, on prendra comme statistique de test T (équation (2)), qui a pour valeur observée[4] $t_k = \frac{\overline{y}^k}{\sigma}\sqrt{n_k \frac{N-1}{N-n_k}}$, soit :

$$t_k = \sqrt{N-1}\, y^k \sqrt{\frac{f_k}{1-f_k}} = \sqrt{N-1}\cos\theta_k \tag{6}$$

D'où $t_k^2 = (N-1)\cos^2\theta_k = (N-1)v_{k\widetilde{k}} = (N-1) \times \mathtt{Q} \times \mathrm{Ctai}_{k\widetilde{k}} = \dfrac{N-1}{1-f_k} \times \mathtt{Q} \times \mathrm{Cta}_k$.

On situe t_k par rapport à la distribution normale réduite, d'où le seuil observé approché.

Homogénéité de deux classes

Pour étudier l'homogénéité des deux classes k et k' (d'une même question) selon la différence des moyennes, on prendra comme statistique de test T (équation (3)), qui a pour valeur observée $t_{kk'} = \frac{\overline{y}^{k'} - \overline{y}^k}{\sigma}\sqrt{\frac{N-1}{N}} \times \frac{1}{\sqrt{\frac{1}{n_k}+\frac{1}{n_{k'}}}}$, soit :

$$t_{kk'} = \sqrt{N-1}\,\frac{y^{k'} - y^k}{\sqrt{\frac{1}{f_k}+\frac{1}{f_{k'}}}} = \sqrt{N-1}\cos\theta_{kk'} \tag{7}$$

D'où $t_{kk'}^2 = (N-1)\cos^2\theta_{kk'} = (N-1) \times (f_k + f_{k'})v_{kk'} = (N-1) \times \mathtt{Q} \times \mathrm{Ctai}_{kk'}$.

On situe $t_{kk'}$ par rapport à la distribution normale réduite, d'où le seuil observé approché.

Propriété. Le test de typicalité de la classe k est équivalent au test d'homogénéité des classes complémentaires k et $\widetilde{k}$ de la question q : $t_k = t_{k\widetilde{k}}$

Homogénéité de plusieurs classes

Pour étudier l'homogénéité de K' classes d'une même question q ($K' \subseteq K_q$), on prendra comme statistique de test T^2 (équation 5 p.6), qui a pour valeur observée $t_{K'}^2$, avec :

$$t_{K'}^2 = (N-1)\frac{N'}{N}v_{K'} = (N-1) \times \mathtt{Q} \times \mathrm{Ctai}_{K'} \tag{8}$$

On situe $t_{K'}^2$ par rapport à la distribution du χ^2 à $K'-1$ d.l. ; d'où le seuil observé approché.

Cas particulier : homogénéité de la question q

$$T^2 = (N-1) \times \mathtt{Q} \times \mathrm{Cta}_q = (N-1)\eta_q^2 = (N-1)v_q \tag{9}$$

CONCLUSION

Le cas de l'ACM est exemplaire de la propriété valable pour toute inférence combinatoire ; les statistiques de test s'expriment en fonction des statistiques descriptives utilisées comme aides à l'interprétation : contributions des points et des écarts, qualités de représentation, coefficient d'homogénéité (ou rapport de corrélation) η^2, etc.

[4] Voir Lebart & al. (1995), valeurs-tests, p. 123-124.

BIBLIOGRAPHIE

ALEVIZOS P. & MORINEAU A., Tests et valeurs–tests : application à l'étude des mastics utilisés dans la fabrication des vitraux, *Revue de Statistique Appliquée*, 40(4), p. 27- 43, 1992 & 41(1), p. 8-22, 1993.

DAUDIN J-J., DUBY C. & TRÉCOURT P., Stability of principal components studied by the bootstrap method, *Statistics*, 19, p.241-258, 1988.

FREEDMAN D., PISANI R., PURVES R. & ADHIKARI A., *Statistics*, 2nd edition, New York, W.W.Norton & co, 1991.

DE LEEUW J. & VAN DER BURG E., The Permutational Limit Distribution of Generalized Canonical Correlations, in *Data analysis and Informatics*, IV, Diday & al. (Eds), 1986.

GIFI A., *Non linear Multivariate Analysis*, Wiley, Chichester, 1990.

GREENACRE M., *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984.

KENDALL M. & STUART A., *The Advanced Theory of Statistics*, Vol 2 & 3, 3ème édition, London, Griffin, Vol 2, 1973, Vol 3, 1976.

LEBART L., The significance of eigenvalues issued from correspondence analysis, *Proceedings in Comp. Statist., COMPSTAT*, Physica verlag, Vienna,p. 38-45, 1976.

LEBART, L., MORINEAU A. & WARWICK M., *Multivariate Descriptive Statistical Analysis*, New-York, Wiley, 1984.

LEBART L., MORINEAU A. & PIRON M., *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 1995.

PITMAN E.J.P., Significance tests which may be applied to samples to any populations, *J. Roy. Stat. Soc. Suppl., 4, 119-130*, 1937.

LE ROUX B. & ROUANET H., Interpreting Axes in Multiple Correspondence Analysis : Method of the Contributions of Points and Deviations, in *Visualization of Categorical Data*, Blasius J. & Greenacre M. (Eds), New York, Academic Press, 1998.

ROUANET H., Mesures, proportions et probabilités : sur l'emploi des formulations ensemblistes en inférence statistique, *Math. Sci. hum.*, n° 80, p.83-89, 1982.

ROUANET H., BERNARD J-M. & LECOUTRE B., Non probabilistic statistical inference : a set–theoretic approach, *American Statistician*, 1986.

ROUANET H., BERNARD, J-M. & LE ROUX B., *Analyse inductive des données*, Paris, Dunod, 1990.

ROUANET H., LECOUTRE M-P., BERT M-C., LECOUTRE B., & BERNARD J-M., *L'inférence statistique dans la démarche du chercheur*, Peter Lang, Berne, 1991.

ROUANET H., BERNARD J-M., BERT M-C., LECOUTRE M-P., LECOUTRE B. & LE ROUX B., *New ways in Statistical Methodology : From Significance Tests to Bayesian Inference*, Peter Lang, Berne, 1998.

ROUANET H.& LE ROUX B., *Analyse des données multidimensionnelles*, Paris, Dunod, 1993.

SAPORTA G. & HATABIAN, Régions de confiance en analyse factorielle, in *Data analysis and Informatics*, IV, Diday & al. (Eds), 1986.

TENENHAUS M., LEROUX Y., GUIMART C. & GONZALEZ P-L., Modèle linéaire généralisé et analyse des correspondances, *Revue de Statistique Appliquée*, 41(2), p. 59-86, 1993.

TER BRAAK C.J.F., Permutation versus Bootstrap significance Tests in *Multiple Regression and* ANOVA, Jöckel & al (Eds), Berlin, Springer, 1992.