Test de permutation pour comparer des groupes indépendants : cas d'un nuage euclidien

Brigitte Le Roux [1] & Solène Bienaise [2] & Jean-Luc Durand [3]

[1] *MAP5, Université Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06 & CEVIPOF, Sciences Po Paris, Brigitte.LeRoux@mi.parisdescartes.fr*

[2] *bienaise.solene@hotmail.fr*

[3] *LEEC, Université Paris 13, F-93430, Villetaneuse, jean-luc.durand@univ-paris13.fr*

Résumé. Dans ce papier, nous présentons un test de permutation applicable à des nuages euclidiens qui sont, par exemple, construits par une méthode d'analyse géométrique des données (AGD). Les tests de permutation appartiennent à l'ensemble des méthodes de ré-échantillonnage. Ils ne sont pas basés sur un modèle aléatoire, mais sur des procédures de permutation formulées dans un cadre combinatoire. Les tests statistiques classiques s'appuient, la plupart du temps, sur des hypothèses invérifiables et sont donc souvent inapplicables aux données d'observation, en particulier en AGD. Les tests de permutation ne dépendent que des données observées et il n'est fait aucune hypothèse sur la distribution des données, c'est pourquoi, l'approche combinatoire est le plus en harmonie avec l'analyse inductive des données.

Les méthodes que nous proposons s'appliquent à des nuages euclidiens multidimensionnels et traitent de la comparaison des points moyens de plusieurs groupes d'observations (tests d'homogénéité pour des groupes indépendants). Nous présentons une application des méthodes à une enquête portant sur les « députés et la mondialisation ».

Mots-clés. Statistique computationnelle, tests de permutation, zone de compatibilité, analyse géométrique des données, distance de Mahalanobis.

Abstract. In this paper, we present a permutation test applicable to Euclidean clouds that, for instance, are constructed by a method of Geometric Data Analysis (GDA). The permutation tests belong to the set of resampling methods. They are not based on a random modelling but on permutation procedures that lie within the combinatorial inference framework. Classical statistical tests are mostly based on assumptions that are usually unverifiable and then most often not applicable to observational data, especially in GDA. Permutation tests only depend on observed data, no assumption is made on the data distribution. That is why the combinatorial approach is the most in harmony with inductive data analysis.

The methods presented here apply to multidimensional Euclidean clouds and deal with the comparison of the mean points of several groups of observations (homogeneity tests for independent groups). These methods are applied to a survey about the "members of the French Parliament and the globalisation".

Keywords. Computational statistic, Permutation tests, Compatibility zone, Geometric Data Analysis, Mahalanobis distance.

1 Homogénéité de plusieurs groupes indépendants

On considère un ensemble (fini) I de n individus et un facteur C tel que I soit emboîté dans C, d'où C groupes d'effectifs $(n_c)_{c\in C}$ formant une partition de I, que l'on note $I<C>$. On s'intéresse à l'homogénéité des C groupes [1] afin de répondre à la question suivante : *les* C *groupes sont-ils, ou non, hétérogènes ?* (Le Roux, 1998; Le Roux et Rouanet, 2004, 2010; Bienaise, 2013; Le Roux *et al.*, 2019).

Le test que nous présentons est multivarié, plus précisément, il s'applique à un nuage $(\mathrm{M}^i)_{i\in I}$ de n points d'un espace affine euclidien $\mathcal{U}$. Pour simplifier l'exposé, on supposera que l'espace est rapporté à un repère cartésien orthonormé ayant pour origine le point moyen du nuage et on restreindra l'espace au support affine du nuage. Le point moyen du nuage M^I est noté G et sa matrice de covariance $\mathbf{V}$. Aux C groupes est associée une partition du nuage M^I en C sous–nuages dont les points moyens sont notés $(\mathrm{G}^c)_{c\in C}$.

Principe du test. On construit le test d'homogénéité de la façon suivante :

1. On considère l'ensemble J des $n!/(\prod_{c\in C} n_c!)$ affectations possibles des n individus dans C groupes d'effectifs $(n_c)_{c\in C}$, c'est–à–dire l'ensemble des emboîtements possibles de type $(n_c)_{c\in C}$. On note $I<c>_j$ l'ensemble des individus appartenant au groupe c d'effectif n_c pour le j–ème emboîtement, $\mathrm{M}^{I<c>_j}$ le sous–nuage qui lui est associé et G^{cj} son point moyen (pondéré par n_c).

   Pour un j donné, les C nuages $(\mathrm{M}^{I<c>_j})_{c\in C}$ forment une partition du nuage M^I. La famille des J nuages partitionnés en C classes définit l'*espace des emboîtements*, on le note $\mathcal{J}$, avec $\mathcal{J} = (\mathrm{M}^{I<C>_j})_{j\in J}$.

2. On fait choix d'une *statistique de test*, notée V_{M} et appelée M–*Variance* :

$$\begin{array}{rccl} V_{\mathrm{M}}: & \mathcal{J} & \to & \mathbb{R}_{\geq 0} \\ & \mathrm{M}^{I<C>_j} & \mapsto & \frac{1}{2}\sum\limits_{c\in C'}\sum\limits_{c'\in C'} \frac{n_c n_{c'}}{n^2}|\mathrm{G}^{cj}\mathrm{G}^{c'j}|^2 = v_M[j] \end{array}$$

   où $|\mathrm{G}^{cj}\mathrm{G}^{c'j}|$ désigne la distance de Mahalanobis [2] entre les points G^{cj} et $\mathrm{G}^{c'j}$.

3. On détermine la proportion des emboîtements possibles dont la valeur de la statistique V_{M} est supérieure ou égale à la valeur observée, notée $v_{\mathrm{M}\,obs}$. Cette proportion définit le *seuil combinatoire du test* (p–value) : $p = p(V_{\mathrm{M}} \geq v_{\mathrm{M}\,obs})$.

4. On énonce la *conclusion du test* en termes d'hétérogénéité des groupes, pour la statistique de test M–*Variance*, à un seuil fixé α.

   Si $p \leq \alpha$, le résultat du test est significatif au seuil α. Pour la statistique M–*Variance*, on peut dire que les groupes sont hétérogènes au seuil α.

1. Sauf pour I, un ensemble et son cardinal sont notés avec la même lettre (majuscule italique).

2. Si $\mathbf{d}^{cj}$ et $\mathbf{d}^{c'j}$ désignent les vecteur–colonnes des coordonnées des points G^{cj} et $\mathrm{G}^{c'j}$, le carré de la distance de Mahalanobis entre les points G^{cj} et $\mathrm{G}^{c'j}$ est égal à $(\mathbf{d}^{cj} - \mathbf{d}^{c'j})'\mathbf{V}^{-1}(\mathbf{d}^{cj} - \mathbf{d}^{c'j})$.

Si $p > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique M–*Variance*, on ne peut pas dire que les groupes sont hétérogènes au seuil α.

Propriété 1.1. *Le nuage* G^{cJ} *admet pour point moyen le point moyen* G *du nuage* M^I.

Propriété 1.2. *La matrice de covariance du nuage* G^{cJ} *est égale à* $\frac{1}{n-1} \times \frac{n-n_c}{n_c} \times \mathbf{V}$.

En conséquence, la matrice de covariance du nuage $(\mathrm{G}^{cJ})_{c \in C}$ est proportionnelle à $\mathbf{V}$. C'est cette propriété qui nous a conduit à définir une statistique de test basée sur la distance de Mahalanobis entre points.

Propriété 1.3. *La moyenne de la statistique* M–*Variance est égale à* $L \times \frac{\mathrm{C}-1}{N-1}$, *L désignant la dimension du nuage (ou rang de la matrice* $\mathbf{V}$*).*

2 Cas de deux groupes indépendants

Dans le cas de deux groupes indépendants c_1 et c_2, le principe du test reste inchangé. Cependant, on fait choix d'une statistique de test plus simple, conduisant à un test équivalent (Pesarin et Salmaso, 2010), et permettant d'une part une *interprétation géométrique du seuil observé* et d'autre part la construction d'une *zone de compatibilité*.

Statistique de test. Si $d_{\mathrm{M}}[j]$ désigne la distance de Mahalanobis entre les points $\mathrm{G}^{c_1 j}$ et $\mathrm{G}^{c_2 j}$ de l'emboîtement j, on a $v_{\mathrm{M}}[j] = \frac{n_{c_1} n_{c_2}}{n_{c_1}+n_{c_2}} d^2_{\mathrm{M}}[j]$. Par conséquent, la statistique D^2_{M} définie ci-après conduit à un test équivalent.

$$\begin{array}{rccl} D^2_{\mathrm{M}} : & \mathcal{J} & \to & \mathbb{R}_{\geq 0} \\ & \mathrm{M}^{I<C>_j} & \mapsto & |\mathrm{G}^{c_1 j}\mathrm{G}^{c_2 j}|^2 = d^2_{\mathrm{M}}[j] \end{array}$$

Le seuil combinatoire du test est la proportion $p(D^2_{\mathrm{M}} \geq |\mathrm{G}^{c_1}\mathrm{G}^{c_2}|^2)$.

Nuage des points–écarts. Il est construit comme suit. Au bipoint $(\mathrm{G}^{c_1 j}, \mathrm{G}^{c_2 j})_{j \in J}$, on associe d'abord le vecteur–écart $\overrightarrow{d^j} = \mathrm{G}^{c_2 j} - \mathrm{G}^{c_1 j}$ puis le point–écart $\mathrm{D}^j = \mathrm{O} + \overrightarrow{d^j}$ (le point O représente l'écart nul), d'où le nuage D^J des points–écarts. On note $\mathrm{D}_{obs} = \mathrm{O} + \overrightarrow{d_{obs}}$ le point associé à l'écart observé $\overrightarrow{d_{obs}} = \mathrm{G}^{c_2} - \mathrm{G}^{c_1}$.

Propriété 2.1. *La matrice de covariance du nuage des points–écarts* D^J, *notée* $\mathbf{V}_{\mathrm{D}}$, *est telle que* $\mathbf{V}_{\mathrm{D}} = \frac{n}{n-1} \times (\frac{1}{n_{c_1}} + \frac{1}{n_{c_2}}) \times \mathbf{V}$, *où* $\mathbf{V}$ *est la matrice de covariance du nuage* M^I.

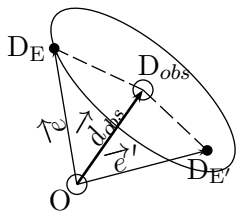
Par conséquent, les ellipsoïdes principaux (ou d'inertie) du nuage des points–écarts D^J se déduisent de ceux du nuage M^I.

Interprétation géométrique du seuil observé. La valeur prise par la statistique de test pour l'emboîtement j est égale à $|\mathrm{OD}^j|^2$ et donc le seuil combinatoire s'écrit $p = p(D^2_{\mathrm{M}} \geq |\mathrm{OD}^2_{obs}|)$. Par conséquent, le seuil observé s'interprète comme la proportion des points–écarts situés sur ou à l'extérieur de l'ellipsoïde principal du nuage M^I de centre O et passant par le point D_{obs}.

Zone de compatibilité. Le test revient à comparer l'écart observé entre les deux groupes à l'écart nul. Le but de la région de compatibilité est de définir un ensemble d'écarts qui sont *compatibles* avec l'écart observé. Pour ce faire, on procède en 3 étapes.

1. On prend comme nuage de référence le nuage intra–C dont l'écart entre les deux groupes d'intérêt est nul et dont la matrice de covariance est la matrice de covariance intra–groupes. Il est noté $\mathrm{R}^{I<C>}$.
2. Étant donné un vecteur $\overrightarrow{e}$, on construit un nuage $\mathrm{E}^{I<C>}$ de point moyen G obtenu par translation des deux sous–nuages $\mathrm{R}^{I<c_1>}$ et $\mathrm{R}^{I<c_2>}$ de sorte que l'écart entre leurs points moyens soit égal à $\overrightarrow{d_{obs}} - \overrightarrow{e}$.
3. On applique le test au nuage $\mathrm{E}^{I<C>}$: le vecteur $\overrightarrow{e}$ est dit compatible avec $\overrightarrow{d_{obs}}$ au seuil α si, pour ce nuage, le seuil combinatoire du test est supérieur à α ($p > \alpha$).

La zone de compatibilité est définie comme l'ensemble des vecteurs $\overrightarrow{e} = \overrightarrow{\mathrm{OD}_E}$ compatibles avec $\overrightarrow{d_{obs}}$. On démontre que la zone ajustée de compatibilité est définie par les points $\mathrm{D_E}$ situés sur ou à l'intérieur d'une ellipse principale du nuage de référence (définie par la matrice de covariance intra) centrée sur le point D_{obs} (voir figure ci–contre).



3 Application : les députés et la mondialisation

Le 29 Mai 2005 a eu lieu un referendum sur la constitution européenne. Un an après, un groupe de chercheurs de Sciences Po Paris[3] a décidé d'étudier les représentations de la mondialisation chez les députés. Dans ce but, ils ont développé un questionnaire qui fut envoyé à tous les députés. Les répondants furent au nombre de 163. Parmi eux, 92 appartenaient au groupe parlementaire UMP, 48 au groupe PS, 15 au groupe UDF, 5 au groupe communiste et 3 étaient sans affiliation. On a effectué une ACM du questionnaire à partir de 12 questions et interprété deux axes.

Le *premier axe* oppose une attitude favorable à une attitude défavorable à la mondialisation. Le *deuxième axe* oppose la protection de l'agriculture à la préservation de l'emploi et de l'environnement.

Le nuage des 163 répondants dans le plan 1-2 est représenté sur la figure 1.

3. Cette étude, qui fut dirigée par Zaki Laïdi, Cynthia Fleury et Brigitte Le Roux, constituait une partie du mémoire de master d'un groupe de 12 étudiants.

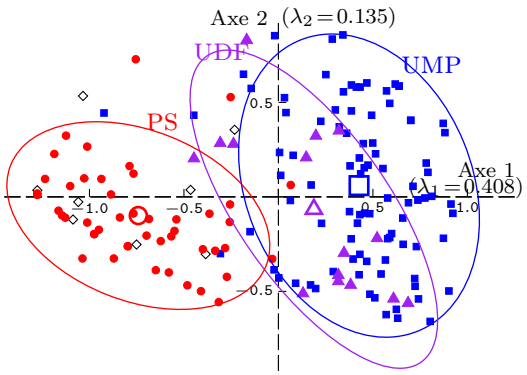


FIGURE 1 – Nuage des députés : 92 UMP (■), 15 UDF (▲), 48 PS (●) et 8 autres (◇). Sous–nuages des 3 principaux groupes avec leurs points moyens et leurs ellipses de concentration (Cramér, 1947 ; Le Roux et Rouanet, 2004).

Pour étudier l'homogénéité des groupes UMP et PS, on procède à une comparaison spécifique des deux groupes en restreignant les données aux deux groupes. Dans le plan 1-2, $\eta^2 = 0.57$, valeur qui est notable.
Pour prolonger cette conclusion descriptive, on effectue le test d'homogénéité pour deux groupes indépendants. L'histogramme de la distribution de la statistique de test est représenté sur la figure 2. La valeur observée de la statistique de test est égale à 3.38, donc le seuil combinatoire est très petit ($p \ll 0.001$, résultat très significatif).

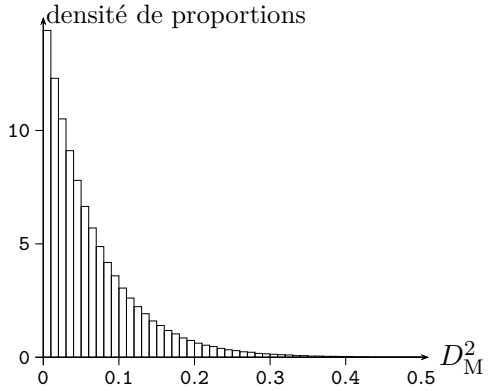


FIGURE 2 – Distribution de D^2_M.

On peut assurément conclure que, dans le plan 1-2, *les deux groupes sont hétérogènes : l'écart entre les deux points moyens n'est pas dû au hasard.*

La figure 3 montre la zone de compatibilité au seuil 5%.

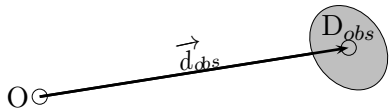


FIGURE 3 – Ellipse ajustée de compatibilité au seuil 5% avec l'écart UMP-PS (translation de la κ-ellipse principale du nuage de référence de vecteur $\vec{d}_{obs}$), avec $\kappa = 0.44$.

Les écarts compatibles avec l'écart observé ($\vec{d}_{obs}$) sont représentés par des vecteurs d'origine O et dont les extrémités appartiennent à l'intérieur de l'ellipse. On voit que tous les écarts compatibles avec l'écart observé sont importants, ce qui renforce la conclusion d'hétérogénéité des deux groupes.

Conclusion

Les tests présentés ici s'inscrivent dans la ligne des tests de permutation développés par Fisher et Pitman (1937), ils en sont une généralisation multivariée. L'applicabilité des tests de permutation est souvent confinée aux situations impliquant la « randomisation »(Edgington, 2007). Cependant dans beaucoup de situations, la randomisation ou l'échantillonnage au hasard n'est pas assuré (voir Freedman et Lane, 1983) ; une telle conception restreint donc sévèrement le domaine d'application des tests de permutation. Dans les tests développés ici, l'échantillonnage au hasard n'est pas une condition requise, on détermine la proportion des emboîtements possibles plus extrêmes que les données et non la probabilité, sous l'hypothèse d'échantillonnage au hasard, d'obtenir un emboîtement plus extrême que l'emboîtement observé (Rouanet *et al.*, 1990; Le Roux *et al.*, 2019).

De nos jours, les tests combinatoires exacts, utilisant la méthode exhaustive lorsque cela est possible ou la méthode de Monte Carlo dès que la taille des données le nécessite, sont praticables. Pour leur mise en œuvre, nous avons écrit des programmes en langage R qui sont disponibles auprès des auteurs.

Références

Bienaise, S. (2013). *Méthodes d'inférence combinatoire sur un nuage euclidien/Etude statistique de la cohorte EPIEG*. Thèse de doctorat, Université Paris Dauphine, ceremade.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton : Princeton University Press.

Edgington, E. (2007). *Randomization Tests*. London : Chapman & Hall/CRC, fourth édition.

Freedman, D. et Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Le Roux, B. (1998). Inférence combinatoire en analyse géométrique des données. *Mathématiques et sciences humaines*, 144:5–14.

Le Roux, B., Bienaise, S. et Durand, J.-L. (2019). *Combinatorial Inference in Geometric Data Analysis*. London : Chapman & Hall/CRC.

Le Roux, B. et Rouanet, H. (2004). *Geometric Data Analysis : From Correspondence Analysis to Structured Data Analysis (Foreword by P. Suppes)*. Dordrecht : Kluwer.

Le Roux, B. et Rouanet, H. (2010). *Multiple Correspondence Analysis, 163*. QASS. Thousand Oaks (CA) : Sage Publications.

Pesarin, F. et Salmaso, L. (2010). *Permutation Tests for Complex Data : Theory, Applications and Software*. Chichester : Wiley.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society*, 4:119–130.

Rouanet, H., Bernard, J.-M. et Le Roux, B. (1990). *Analyse inductive des données*. Paris : Dunod.