Université Paris–Dauphine
École doctorale de Dauphine

Médecine de Contrôle
des Industries Electriques et Gazières

Tests combinatoires en Analyse Géométrique des Données – Etude de l'absentéisme dans les Industries Electriques et Gazières de 1995 à 2011 à travers des données de cohorte

Thèse

Pour l'obtention du titre de

Docteur en Sciences – Mention Mathématiques Appliquées

Présentée par

Solène BIENAISE

Soutenue publiquement le 3 octobre 2013 devant le jury composé de

Gérard d'AUBIGNY *Rapporteur*	– Professeur, Université Pierre–Mendès–France, Grenoble
Mohamed NADIF *Rapporteur*	– Professeur, Université Paris Descartes
Avner BAR–HEN *Examinateur*	– Professeur, Université Paris Descartes
Mireille GETTLER–SUMMA *Examinatrice*	– Maître de Conférences, Université Paris Dauphine
Judith ROUSSEAU *Examinatrice*	– Professeur, Université Paris Dauphine
Pierre CAZES *Directeur*	– Professeur, Université Paris Dauphine
Brigitte LE ROUX *Co–directrice*	– Maître de Conférences (HDR), Université Paris Descartes
Catherine GODARD *Responsable scientifique*	– Médecin épidémiologiste, Service Général de Médecine de Contrôle des Industries Electriques et Gazières

L'Université n'entend donner aucune approbation ni improbation aux opinions émises dans les thèses : ces opinions doivent être considérées comme propres à leurs auteurs.

Remerciements

Je tiens ici à remercier toutes les personnes qui m'ont accompagnée tout au long de ces années de doctorat.

Tout d'abord, je remercie chaleureusement Brigitte Le Roux pour la grande aide et le soutien qu'elle m'a apportés durant cette thèse. Ses remarques et ses précieux conseils m'ont grandement orientée dans ce travail. Je garde un très bon souvenir de notre collaboration constructive et agréable, j'espère qu'elle continuera par la suite. Je voudrais également remercier Pierre Cazes pour sa disponibilité, ses conseils et ses relectures minutieuses. Merci aussi à Mireille Gettler–Summa pour son aide, pour ses encouragements dans les moments de doutes et pour son dynamisme. Je tiens également à remercier Catherine Godard qui m'a beaucoup soutenue pendant toute la durée de cette thèse. Merci pour sa disponibilité et pour son aide précieuse. Ce fût un plaisir de travailler aux côtés de ces personnes, elles m'ont beaucoup appris.

Je suis très honorée que Gérard d'Aubigny et Mohamed Nadif aient accepté d'être rapporteurs de ma thèse, je leur adresse un grand merci pour leurs remarques et leurs conseils. Merci aussi à Judith Rousseau et Avner Bar–Hen pour l'intérêt qu'ils ont porté à mon travail et pour leur participation à mon jury.

Je tiens à remercier Charles Goulfier et Bernadette Michelin pour l'accueil chaleureux qu'ils m'ont réservé au sein du SGMC. J'ai une pensée pour Charles Goulfier, je regrette sincèrement qu'il n'ait pas pu m'accompagner jusqu'au bout de ce doctorat.

Je remercie également Patrick Suppes, Trevor Hastie et Bradley Efron pour m'avoir reçu dans leurs locaux à l'université de Stanford. J'ai gardé à l'esprit tout de long de mon travail leurs commentaires et les conseils qu'ils m'ont prodigués durant nos entretiens.

Je voudrais aussi remercier tous mes collègues du SGMC, en particulier Annick, Didier, Monique et Sylvie, merci pour leur gentillesse, leur bienveillance et leur soutien. Merci également à tous les membres du CEREMADE pour les différents échanges que nous avons eus et à l'ensemble des thésards, en particulier ceux du bureau C614, pour l'ambiance de travail agréable. Merci aussi au personnel administratif de Dauphine pour son aide dans les procédures administratives parfois compliquées dans le cadre d'une convention CIFRE.

Je tiens à remercier mes nouveaux collaborateurs de Coheris SPAD, en particulier Benoît, Julien et Philippe pour leur appui et leur flexibilité au cours de mes premiers mois en poste qui ont coïncidé avec mes derniers mois de thèse.

Enfin, j'adresse un grand merci à mes proches pour leur soutien sans faille. Merci à mes parents Marie et Guy pour leur présence tout au long de mes études et pour m'avoir toujours soutenue et encouragée. Merci également à mes soeurs Anouchka et Lisa, à mon frère Alexis, à mes grands parents, à mes cousines et à toute ma famille. J'ai ici une pensée particulière pour mon grand–père Jean.

Je souhaite également remercier mes amis, en particulier Camille et Caroline, merci d'avoir écouté et compris mes inquiétudes, merci de votre présence durant toutes ces années. Merci aussi à tous les autres !

Un dernier mot pour Nicolas, merci pour son soutien, pour avoir partagé mes peurs et mes doutes et pour avoir supporté mon stress et mon indisponibilité, surtout pendant ces derniers mois.

Table des matières

Introduction générale **7**

1 Nuage euclidien **11**
1 Statistiques élémentaires 11
2 Décomposition inter–intra d'un nuage 13
2.1 Sous–nuage 14
2.2 Partition d'un nuage : nuages inter et intra 14
3 Directions principales d'un nuage 16
4 Distances de Mahalanobis – Normes 17
5 Ellipsoïdes d'inertie 20
6 Formulations numériques et matricielles 21

2 Test géométrique de typicalité **23**
1 Typicalité d'un point moyen par rapport à un point de référence 24
1.1 Test exact 24
1.1.1 Principe du test 24
1.1.2 Zone de compatibilité 31
1.1.3 Cas particulier d'un nuage unidimensionnel 32
1.2 Test approché 34
1.2.1 Zone approchée de compatibilité 35
1.2.2 Cas particulier d'un nuage unidimensionnel 36
1.3 Cas de deux groupes appariés. 37
2 Autres tests 38
2.1 Test de Hotelling d'écart nul 38
2.1.1 Modèle normal 38
2.1.2 Seuil observé du test 39
2.1.3 Zone de confiance 39
2.1.4 Cas particulier d'un nuage unidimensionnel : test de Student 40
2.2 Test du bootstrap 41
3 Applications : données Parkinson (cas de deux groupes appariés) 44
3.1 Test géométrique exact 48
3.2 Test géométrique approché 49

3.3 Test de Hotelling 50
3.4 Test du bootstrap 50
3.5 Tests unidimensionnels 51

3 Test ensembliste de typicalité 53
1 Typicalité d'un groupe d'observations par rapport à une population de référence 53
1.1 Test exact 54
1.1.1 Principe du test 54
1.1.2 Zone de compatibilité 61
1.1.3 Cas particulier d'un nuage unidimensionnel 63
1.2 Test approché 64
1.2.1 Zone approchée de compatibilité 65
1.2.2 Cas particulier d'un nuage unidimensionnel : test Z approché 66
2 Typicalité d'un groupe d'observations par rapport à une distribution de référence gaussienne : test Z 67
2.1 Seuil observé du test 68
2.2 Zone de compatibilité 68
2.3 Cas particulier d'un nuage unidimensionnel : test Z unidimensionnel . 69
3 Applications 70
3.1 Exemple des races canines 70
3.1.1 Test exact 73
3.1.2 Test approché 74
3.2 Données Parkinson 75
3.2.1 Test exact 77
3.2.2 Test approché 79
3.2.3 Tests unidimensionnels. 80

4 Test d'homogénéité 83
1 Homogénéité de plusieurs groupes indépendants 84
1.1 Test exact 85
1.1.1 Principe du test 85
1.1.2 Cas particulier d'un nuage unidimensionnel 93
1.2 Test approché 93
1.2.1 Principe du test 93
1.2.2 Cas particulier d'un nuage unidimensionnel 94
2 Homogénéité de deux groupes indépendants 94
2.1 Test Exact 95
2.1.1 Principe du test 95
2.1.2 Zone de compatibilité 100
2.1.3 Cas particulier d'un nuage unidimensionnel 107
2.2 Test approché 109

2.2.1 Principe du test . . . 109
2.2.2 Zone approchée de compatibilité . . . 109
2.2.3 Cas particulier d'un nuage unidimensionnel . . . 110
3 Autres tests . . . 111
3.1 Test de Hotelling pour la comparaison de deux groupes indépendants 111
3.2 Tests basés sur le modèle normal pour la comparaison de plus de deux groupes . . . 112
3.3 Test du bootstrap . . . 113
4 Applications : exemple des races canines . . . 114
4.1 Test d'homogénéité exact . . . 116
4.1.1 Comparaison globale . . . 116
4.1.2 Comparaisons associées à des contrastes orthogonaux . . . 117
4.2 Test d'homogénéité approché . . . 119
4.2.1 Comparaison globale . . . 119
4.2.2 Comparaisons associées à des contrastes orthogonaux . . . 120
4.3 Test du bootstrap . . . 121

5 L'absentéisme dans les IEG de 1995 à 2011 : étude statistique de la cohorte EPIEG 123

Introduction . . . 123
1 La cohorte EPIEG . . . 124
1.1 Déploiement vertical et concept de personnes/temps . . . 125
1.2 Population étudiée . . . 125
1.3 Evolution de la structure de la population . . . 127
2 Les épisodes d'arrêt . . . 131
3 Evolution des principaux indices d'absence . . . 133
4 Analyses des correspondances . . . 142
4.1 Longue Maladie . . . 142
4.1.1 Le tableau de données . . . 142
4.1.2 Analyse des correspondances (année 2010) . . . 143
4.1.3 Evolutions des liaisons . . . 152
4.2 Maladie Courte durée . . . 156
4.2.1 Le tableau de données . . . 156
4.2.2 Analyse des correspondances (année 2010) . . . 156
4.2.3 Evolutions des liaisons . . . 162
Conclusion . . . 163

Conclusion générale 165

Annexes **168**

A Les bases de données du SGMC **171**

B Analyse Géométrique des Données, compléments **177**
1 Analyse géométrique des données 177
1.1 Principales méthodes 177
1.2 Recherche des axes principaux : représentation géométrique 178
2 Analyses sur tableaux à trois entrées 180
2.1 Approche « Modèles » 181
2.1.1 Décomposition PARAFAC 181
2.1.2 Décomposition de TUCKER 183
2.2 Autres approches 184

C Mise en oeuvre informatique **185**
1 Notices d'utilisation 185
1.1 Test géométrique de typicalité 185
1.2 Test ensembliste de typicalité 189
1.3 Test d'homogénéité 192
2 Étapes de développement (test géométrique de typicalité) 196

Communications **201**

Références **203**

Introduction générale

Cette thèse a été effectuée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE), mise en place entre le Centre de Recherche en Mathématiques de la Décision (CEREMADE) et le Service Général de Médecine de Contrôle (SGMC) des Industries Electriques et Gazières (IEG). Elle comporte deux parties :

- La première, développée dans le cadre universitaire, traite de problèmes d'inférence combinatoire en Analyse Géométrique des Données (AGD) (chapitres 1 à 4).
- La seconde, effectuée au sein du SGMC, présente une analyse statistique de l'absentéisme de la population salariale des IEG de 1995 à 2011, avec construction de la base de données utilisée. Les méthodes d'AGD y sont notamment employées (chapitre 5).

L'ensemble des travaux théoriques de cette thèse (1ère partie) s'inscrivent dans le cadre de l'Analyse Géométrique des Données [1]. L'AGD est l'approche de la statistique multivariée due à J.–P. Benzécri et développée initialement autour de l'Analyse des Correspondances et de la classification euclidienne.

L'ensemble des méthodes d'AGD [2] repose sur différentes approches :

- *Approche géométrique :* les données sont représentées par des nuages de points à valeurs dans des espaces euclidiens et toute l'interprétation est basée sur ces nuages de points. Les données ne sont donc pas confinées à leur description numérique et sont considérées comme des *objets géométriques* à part entière.
- *Approche formelle :* les procédures et démonstrations sont guidées par des structures mathématiques. Le cadre théorique mathématique sous–jacent aux procédures d'AGD est le cadre de l'*algèbre linéaire* (concepts d'espaces vectoriels, d'applications linéaires, de produits scalaires, *etc.*), une distinction claire est faite entre vecteurs (éléments de l'espace vectoriel), points (éléments de l'espace géométrique) et ensemble de nombres (*i.e.* coordonnées). *A contrario*, les courants classiques en statistique multivariée utilisent presque exclusivement des opérations matricielles, souvent pas assez puissantes pour prendre en compte toutes les structures pertinentes (voir la critique du « calcul matriciel » dans Benzécri & al., 1973, p.58 [7] et dans Le Roux & Rouanet, 2004, p.9 et p.449 [50]).
- *Approche inductive :* dans la philosophie de l'AGD, la phase descriptive est la première phase indispensable à toute analyse de données. Les courants classiques en statistique multivariée stipulent que les données sont obtenues par un processus d'échantillonnage et négligent le plus souvent cette phase descriptive.

L'inférence statistique, en particulier les tests de signification, sont souvent absents des travaux utilisant l'AGD, et ce malgré un certain nombre de recherches consacrées aux propriétés inférentielles de ces méthodes (*cf.* par exemple Lebart, 1975 [52] et 1976 [53] ; Dau-

1. « Geometric Data Analysis » selon la suggestion de P. Suppes (*cf.* Le Roux & Rouanet, 2004, p.vii) [50].

2. Descriptif des principales méthodes d'AGD en annexe (p.177).

din & al., 1988 [22] ; Saporta & Hatabian, 1986 [79] ; Lebart & al., 1995 [54]). En réalité, la mise en oeuvre des méthodes d'AGD n'exclut en rien leur *visée inductive*, c'est–à–dire le fait que les conclusions souhaitées aillent au delà des conclusions descriptives. D'après B. Le Roux (1998 [48]) : « 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 (randomisation) — sont rarement remplies, ce qui jette un doute sur la validité des conclusions inférentielles ». Ces problèmes peuvent être surmontés si l'on comprend que l'algorithme sous–jacent à toute procédure inférentielle doit être dissocié du cadre d'interprétation. Trois cadres d'interprétation et de justification ont ainsi été définis par Rouanet & al. (1986 [73] ; 1990 [72] ; 1998 [71]), ils diffèrent selon le mode d'intervention des probabilités.

- Cadre *fréquentiste* : c'est le plus répandu à l'heure actuelle, il repose sur des hypothèses fortes portant sur le processus d'obtention des données : échantillonnage au hasard, randomisation, observations indépendamment et identiquement distribuées, distributions gaussiennes (ou normales), *etc.* Les probabilités interviennent ici sous forme de l'aléatoire et sont regardées comme des fréquences idéalisées.
- Cadre *combinatoire* : il permet de se libérer des contraintes du cadre fréquentiste. A la base de toute procédure inductive, on trouve toujours la construction d'un *ensemble de référence* par rapport auquel on situe les données. En inférence combinatoire, la construction de cet ensemble relève de techniques « combinatoires » (la plupart du temps permutationnelles), libres de toutes hypothèses concernant les distributions. Les formulations probabilistes classiques sont remplacées par des formulations en termes de proportions des éléments de l'ensemble de référence plus extrêmes que les données, selon une certaine statistique d'intérêt. L'inférence combinatoire est le prolongement naturel de la statistique descriptive ; elle apparaît comme le premier degré de l'inférence, en ce sens qu'elle fournit le cadre minimal pour évaluer le *potentiel inductif des données*. Ces méthodes ont l'avantage de s'appliquer à des situations où des modèles probabilistes seraient non–valides, non–pertinents ou dont les hypothèses seraient invérifiées ou invérifiables. Tous les tests développés dans cette thèse s'inscrivent et s'interprètent dans le cadre combinatoire.
- Cadre *bayésien* : il constitue un élargissement du cadre fréquentiste dans lequel les probabilités sont vues comme des *degrés de confiance accordés aux hypothèses*, conditionnellement aux données. L'inférence bayésienne, qui ferait sortir du cadre des « statistiques sans probabilités » dans lequel s'inscrit cette thèse, n'est pas abordée. Pour plus de détails concernant l'inférence bayésienne, voir Rouanet (1996 [70]) et Rouanet & al. (1976 [76] ; 1998 [71]).

Selon le cadre choisi, le même algorithme peut ainsi faire l'objet d'interprétations tout–à–fait différentes. Par exemple, lorsqu'on compare la moyenne $\bar{x}$ d'un groupe d'observations, de taille n et de variance v, à une moyenne de référence x_0 selon le test de Student :

- Dans le cadre fréquentiste : on suppose que le groupe d'observations est un échantillon indépendamment et identiquement distribué d'une population parente gaussienne de moyenne μ et de variance σ^2. L'hypothèse à tester est « $\mathcal{H}_0 : \mu = x_0$ ». Sous l'hypothèse $\mathcal{H}_0$, la distribution d'échantillonnage de la statistique *Rapport de Student* (t–ratio) suit une loi de Student à $n-1$ degrés de liberté :

$$t = \sqrt{n-1} \times \frac{\bar{x} - \mu}{\sqrt{v}} \sim t_{n-1}$$

  Le seuil observé unilatéral est alors défini comme étant la *probabilité*, sous $\mathcal{H}_0$, que la

statistique de test t soit plus extrême, du côté des données, que la statistique observée $t_{obs} = \sqrt{n-1} \times \frac{\bar{x}-x_0}{\sqrt{v}}$.

– Dans le cadre combinatoire : la statistique de test (t–ratio) est calculée sur des sous–ensembles de taille n extraits d'une population gaussienne, d'où sa distribution d'échantillonnage. Le seuil observé (combinatoire) du test est alors la *proportion* des sous–ensembles dont le t–ratio dépasse le t–ratio observé. En d'autres termes, au lieu de supposer que les données sont un échantillon aléatoire d'une distribution gaussienne, on situe les données par rapport à une distribution gaussienne de moyenne donnée.

Cette approche est générale, toute procédure fréquentiste peut donner lieu à une procédure combinatoire. La probabilité pour un échantillon aléatoire de satisfaire une *propriété d'intérêt* est alors convertie en proportion des échantillons qui satisfont cette propriété. Clairement, les procédures combinatoires dépendent de la taille n, elles constituent donc des procédures inductives à part entière.

A partir d'un nuage de points, toutes les statistiques descriptives usuelles peuvent être calculées (point moyen, variance, contributions, *etc.*). Le but de cette thèse est de montrer comment, dans le cadre de l'AGD et du multidimensionnel, les statistiques descriptives peuvent faire l'objet de procédures inductives combinatoires prolongeant les conclusions descriptives. Nous nous intéressons en particulier aux problèmes suivants :

– Comparaison d'un point moyen à un point de référence [3] (ou d'une moyenne à une moyenne de référence en unidimensionnel), problème que nous résolvons en mettant en place le *test géométrique de typicalité* (chapitre 2).
– Comparaison d'un groupe d'observations à une population de référence : *test ensembliste de typicalité* (chapitre 3).
– Comparaison de plusieurs groupes indépendants : *test d'homogénéité* (chapitre 4).

Ces *tests combinatoires multidimensionnels* s'inscrivent dans la ligne des travaux de H. Rouanet, de B. Le Roux et de plusieurs de leurs collaborateurs sur l'inférence combinatoire. Les tests de typicalité et d'homogénéité (pour la comparaison de deux groupes) sont présentés dans le cas unidimensionnel dans Rouanet & al. (1990, chapitres IV et V [72]) et dans Rouanet & al. (1998, chapitre 4 [71]). Dans Le Roux (1998 [48]), le test d'homogénéité est étendu à la comparaison de plusieurs groupes indépendants (toujours dans le cas unidimensionnel). Dans Le Roux & Rouanet (2004 [50]), on trouve une ébauche du test géométrique (p.326) et dans Le Roux et Rouanet (2010, chapitre 5 [51]), on trouve certains des éléments constitutifs des tests de typicalité et d'homogénéité (pour la comparaison de deux groupes) appliqués à des nuages projetés sur un plan principal. L'ensemble de ces travaux sont repris en profondeur dans cette thèse et prolongés systématiquement au cas multidimensionnel, avec comparaison de plusieurs groupes pour l'homogénéité. Des zones de compatibilité, qui n'ont jamais été étudiées dans ce cadre auparavant, sont conçues pour chaque test [4] ; leurs procédures de construction sont clairement définies.

Les procédures combinatoires fournissent des réponses appropriées aux problèmes de typicalité et d'homogénéité, en particulier lorsque la sémantique de l'aléatoire n'a pas lieu d'intervenir (ce qui est le cas dans un grand nombre de situations). En effet, à partir du seuil observé, nous concluons en termes « combinatoires », c'est–à–dire en termes d'atypicalité d'un point moyen par rapport à un point de référence (test géométrique de typicalité), d'atypicalité d'un groupe d'observations par rapport à une population de référence (test

3. La comparaison inductive peut bien sûr porter sur des objets plus compliqués, tels que la médiane ou la variance par exemple.

4. Nous constatons qu'elles s'apparentent dans la plupart des cas aux zones de confiance obtenues de façon traditionnelle sous le modèle normal.

ensembliste de typicalité) ou d'hétérogénéité de plusieurs groupes (test d'homogénéité), jamais en termes « probabilistes » où l'hypothèse d'échantillonnage aléatoire doit être remplie. Cette conception combinatoire correspond exactement à la « nonstochastic interpretation » du seuil de signification du test de Freedman & Lane (1983 [33]).

Le test géométrique de typicalité et le test d'homogénéité appartiennent tous les deux à la famille des *tests de permutation*. Initiés par Fisher (1935 [31]) et par Pitman (1937 [65]), les tests de permutation sont des tests « non paramétriques » classiques (*cf.* Kendall & Stuart, 1973 [45]). Ils constituent la principale alternative au modèle normal, leur mise en oeuvre est en effet libre de toute hypothèse concernant les distributions. En contrepartie, la technique des tests de permutation requiert des moyens de calcul importants, c'est pourquoi elle a été écartée des procédures inférentielles durant plusieurs décennies. Cette restriction n'est plus valide aujourd'hui où de puissants ordinateurs permettent d'effectuer des calculs très volumineux. De nombreuses approches inférentielles permutationnelles ont d'ailleurs été développées depuis les années 1990 (*cf.* par exemple Crowley, 1992 [21] ; Edgington, 1995 [26] et Manly, 1997 [58]). Cependant, les tests de permutation sont souvent confinés à des situations impliquant la randomisation, c'est–à–dire l'affectation au hasard des individus dans les groupes. Cette considération restreint fortement le domaine d'applicabilité des tests de permutation et n'est pas utilisée dans cette thèse.

Les tests de permutation sont très utilisés en statistique multivariée où les hypothèses sur les distributions sont souvent difficiles à remplir. De nombreux travaux existent dans le contexte de la régression multiple (*cf.* par exemple Oja, 1987 [62] ; Anderson & Legendre, 1999 [2] ; Shadrokh (thèse), 2006 [81]), mais aussi dans le contexte de la comparaison de plusieurs groupes (homogénéité) (*cf.* par exemple Edgington, 1995 [26] et Anderson, 2001 [1]). Edgington parle en particulier de « modèle géométrique » pour la comparaison de plusieurs groupes ; cependant la statistique de test (qui s'apparente à la statistique de Fisher) et la distance utilisée pour la calculer (distance géométrique usuelle) diffèrent de celles que nous proposons dans cette thèse.

Dans la première partie de cette thèse, nous rappelons d'abord les différentes notions inhérentes aux nuages euclidiens (chapitre 1). Puis, nous répondons aux problèmes de typicalité et d'homogénéité pré–cités en construisant des tests combinatoires (chapitres 2 à 4). Les procédures sont exposées pour des nuages de points multidimensionnels et appliquées, sans perte de généralité, à des nuages à deux dimensions. Nous comparons également les résultats avec ceux obtenus par d'autres procédures inférentielles basées d'une part, de façon traditionnelle, sur le modèle normal et d'autre part, sur des procédures de rééchantillonnage de type bootstrap.

La seconde partie de cette thèse (chapitre 5) est consacrée à l'exposé d'une partie des travaux effectués au sein du SGMC (dans le cadre de la convention CIFRE mise en place avec le CEREMADE). L'étude présentée traite de l'absentéisme de la population salariale des IEG de 1995 à 2011 [5]. Un premier travail très important de construction d'une cohorte épidémiologique a été effectué, il est également décrit ici. Le contexte est plus largement détaillé en introduction de cette partie (p.123).

5. D'autres études ont été effectuées mais ne peuvent pas être présentées ici en raison de la confidentialité des données.

Chapitre 1

Nuage euclidien

Les méthodes d'inférence statistique développées ci–après ont toutes le même objet d'étude : un *nuage euclidien*. C'est pourquoi nous nous proposons dans ce chapitre d'en redéfinir les principales caractéristiques. Plusieurs notions propres aux nuages euclidiens sont également rappelées.

On dispose d'un nuage euclidien lorsque les observations sont des points d'un espace euclidien. En analyse de données, on étudie souvent des nuages que l'on rapporte à leurs axes principaux (comme par exemple ceux issus d'une analyse des correspondances, d'une analyse des correspondances multiples ou d'une analyse en composantes principales), les colonnes du tableau initial peuvent donc être soit des variables numériques, soit des variables catégorisées. Cependant, on peut également construire des nuages à partir d'un tableau brut de données numériques.

Dans ce chapitre, nous présentons d'abord les différentes notions concernant les nuages euclidiens : statistiques élémentaires, décomposition inter–intra, directions principales, normes et ellipsoïdes d'inertie. Puis, en faisant choix d'un repère cartésien, nous en donnons les formulations numériques et matricielles. L'ensemble de ces notions sont nécessaires pour appréhender les méthodes d'inférence statistique développées dans cette thèse.

1 Statistiques élémentaires

Soit $\mathcal{U}$ un espace affine euclidien de dimension K et $\mathcal{V}$ l'espace vectoriel directeur de $\mathcal{U}$. $\mathcal{U}$ est muni du produit scalaire noté $\langle .|. \rangle$ et de la norme euclidienne associée, ou norme géométrique, notée $|| \ . \ ||$ $(= \sqrt{\langle .|. \rangle})$.

Notations. Les points sont notés par des lettres majuscules : M, P, A, *etc.* Les vecteurs géométriques, éléments de l'espace vectoriel $\mathcal{V}$ sous–jacent à $\mathcal{U}$, sont fléchés : $\overrightarrow{\alpha}$, $\overrightarrow{u}$, *etc.* Le vecteur associé au bipoint (M, P) est noté $\overrightarrow{\mathrm{MP}}$. La distance (géométrique) entre deux points P et M, notée PM, est telle que : $\mathrm{PM} = ||\overrightarrow{\mathrm{PM}}||$.

Définition 1.1 (Nuage Euclidien)**.** *Un nuage euclidien pondéré, en bref nuage, est une application de l'ensemble (fini) I à valeurs dans un espace affine euclidien $\mathcal{U}$. Les observations sont des points, notés* M^i *:*

$$\begin{array}{rccll} \mathrm{M}^I : & I & \to & \mathcal{U} & \textit{(espace affine euclidien)} \\ & i & \mapsto & \mathrm{M}^i & \textit{(point)} \end{array}$$

Chaque point est muni du poids $(n_i)_{i\in I}$*, on note* n_I *la mesure effectif :*

$$\begin{array}{rccl} n_I : & I & \to & \mathbb{R} \\ & i & \mapsto & n_i \ (>0) \end{array}$$

Notons $n = \sum_{i\in I} n_i$ le poids total du nuage. Lorsque pour tout $i \in I$, $n_i = 1$, le nuage est dit *élémentaire*.

Notons $\mathcal{M}$ le support affine du nuage, c'est–à–dire le plus petit sous–espace affine euclidien contenant le nuage, et $\mathcal{L}$ le sous–espace vectoriel directeur de $\mathcal{M}$. Notons L la dimension de $\mathcal{M}$ (ou de $\mathcal{L}$), c'est la *dimension du nuage*, avec par conséquent $L \leq K$.

Dans ce chapitre, nous étudions un nuage euclidien, selon la démarche de la géométrie, indépendamment du choix d'un repère. Cependant, la dernière section de ce chapitre est consacrée aux formulations numériques et matricielles obtenues en faisant choix d'un repère cartésien.

L'exposé est illustré par le nuage plan de huit points représenté sur la figure ci–après.

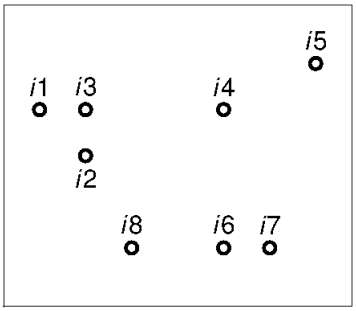


FIGURE 1.1 – Nuage euclidien.

Définition 1.2 (Point Moyen)**.** *Soit* P *un point quelconque de* $\mathcal{U}$*, le point moyen du nuage, noté* G*, est défini par*

$$\overrightarrow{\mathrm{PG}} = \sum_{i\in I} f_i \overrightarrow{\mathrm{PM}^i} \quad avec \quad f_i = n_i/n$$

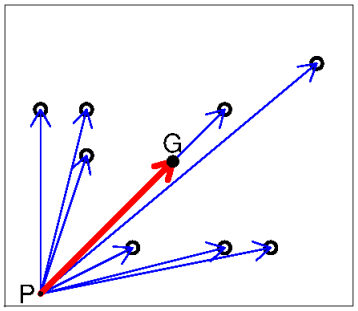
Le point G ne dépendant pas de P, on écrit :

$$\mathrm{G} = \sum_{i\in I} f_i \mathrm{M}^i$$

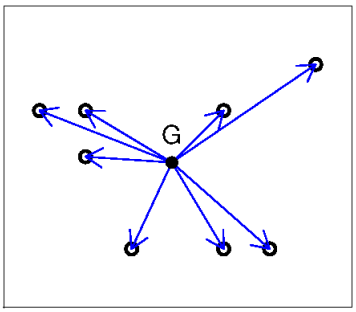
En posant P = G dans la définition précédente, on obtient la *caractérisation barycentrique du point moyen* :

$$\sum_{i\in I} f_i \overrightarrow{\mathrm{GM}^i} = \overrightarrow{0}$$

.



$$\overrightarrow{\mathrm{PG}} = \sum_{i\in I} f_i \overrightarrow{\mathrm{PM}^i}$$



$$\sum_{i\in I} f_i \overrightarrow{\mathrm{GM}^i} = \overrightarrow{0}$$

Définition 1.3 (Variance)**.** *La variance du nuage est la moyenne pondérée des carrés des distances des points* M^i *au point moyen* G*.*

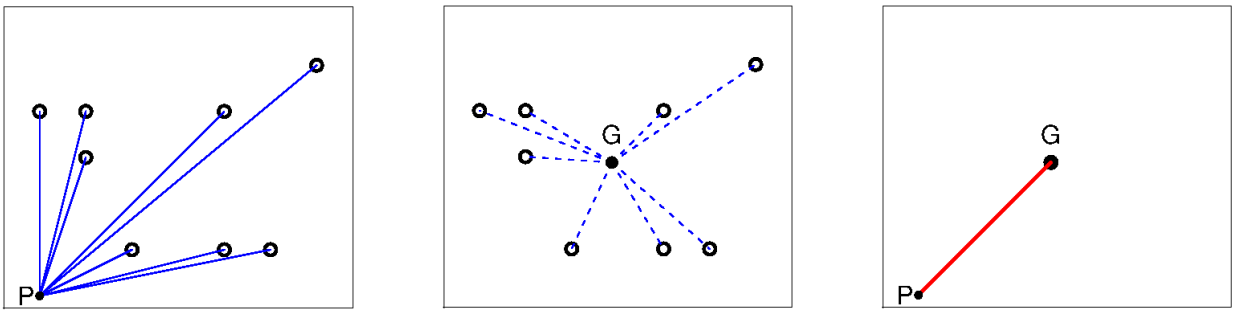
$$\mathrm{Var}\,\mathrm{M}^I = \sum_{i\in I} f_i (\mathrm{GM}^i)^2$$

(*cf.* Le Roux & Rouanet, 2004, p.78 [50])

Remarque : Nous privilégions ici le terme « variance » réservé aux moyennes de carrés, plutôt que le terme « inertie » utilisé pour les sommes de carrés.

Théorème 1.1 (Premier théorème de Huyghens)**.** *La moyenne des carrés des distances des points d'un nuage à un point de référence* P *est la somme de la variance du nuage et du carré de la distance du point moyen au point de référence* P*.*

$$\forall \mathrm{P} \in \mathcal{U},\ \sum_{i\in I} f_i (\mathrm{PM}^i)^2 = \mathrm{Var}\,\mathrm{M}^I + (\mathrm{PG})^2$$



$$\sum_{i\in I} f_i (\mathrm{PM}^i)^2 \quad = \quad \mathrm{Var}\,\mathrm{M}^I \quad + \quad (\mathrm{PG})^2$$

2 Décomposition inter–intra d'un nuage

Les notions présentées dans cette section sont particulièrement utiles pour appréhender les paragraphes concernant la caractérisation de la zone de compatibilité du test combinatoire d'homogénéité (chapitre 4, p.100).

2.1 Sous–nuage

Soit $I<c>$ une partie de I, le sous–ensemble des points $(\mathrm{M}^i, n_i)_{i\in I<c>}$ définit le sous–nuage $\mathrm{M}^{I<c>}$, son poids est égal à $n_c = \sum_{i\in I<c>} n_i$.

– Le *point moyen* du sous–nuage $\mathrm{M}^{I<c>}$ est :

$$\mathrm{G}^c = \sum_{i\in I<c>} n_i \mathrm{M}^i / n_c$$

il est muni du poids n_c (poids du sous–nuage).

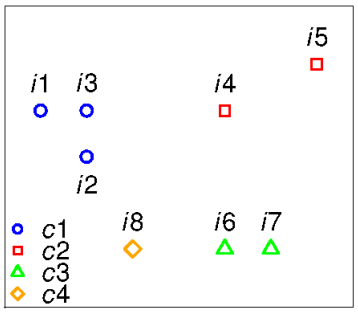
– La variance du sous–nuage $\mathrm{M}^{I<c>}$, notée $\mathrm{Var}\,\mathrm{M}^{I<c>}$, est telle que :

$$\sum_{i\in I<c>} \frac{n_i}{n_c} (\mathrm{G}^c\mathrm{M}^i)^2$$

2.2 Partition d'un nuage : nuages inter et intra

Soit une partition de I en C groupes (ou classes). L'ensemble indexant les groupes est noté $C = \{1, \ldots, c, \ldots, \mathrm{C}\}$ [1]. Le groupe $(I < c >)_{c\in C}$ de I est d'effectif $n_c = \sum_{i\in I<c>} n_i$. A ce groupe est associé le sous–nuage $\mathrm{M}^{I<c>} = (\mathrm{M}^i)_{i\in I<c>}$ admettant le point G^c pour point moyen.

Nous illustrons ce paragraphe en considérant la partition en $\mathrm{C} = 4$ groupes suivante : $I<c1>= \{i1, i2, i3\}$, $I<c2>= \{i4, i5\}$, $I<c3>= \{i6, i7\}$ et $I<c4>= \{i8\}$.



Définition 2.1 (Nuage inter–C)**.** *Le nuage inter–C, noté (M^C, n_C), est constitué des* C *points moyens G^c, chaque point moyen étant muni du poids n_c.*

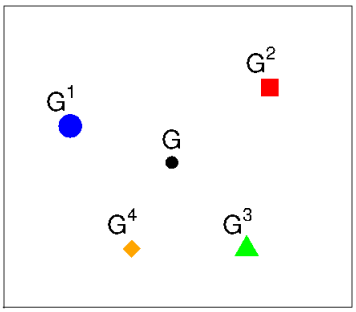


FIGURE 1.2 – Nuage inter–C.

1. La plupart du temps dans ce chapitre, l'ensemble et son cardinal sont identifiés par la même lettre, l'ensemble en italique et son cardinal en lettre droite.

Par construction, le nuage inter–C admet le point G pour point moyen. Sa variance, notée $\mathrm{Var}\,\mathrm{M}^C$, est appelée *variance inter* et est telle que :

$$\mathrm{Var}\,\mathrm{M}^C = \sum_{c\in C} f_c(\mathrm{GG}^c)^2 \quad \text{où} \quad f_c = \frac{n_c}{n}$$

Définition 2.2 (Nuage intra–c)**.** *Le sous–nuage intra-c, noté* $\mathrm{M}^{I(c)}$ *est obtenu par translation de vecteur* $\overrightarrow{\mathrm{G}^c\mathrm{G}}$ *du sous–nuage* $\mathrm{M}^{I<c>}$ *:*

$$\forall i \in I<c>,\ \mathrm{M}^{i(c)} = \mathrm{M}^{i<c>} + \overrightarrow{\mathrm{G}^c\mathrm{G}} = \mathrm{G} + \overrightarrow{\mathrm{G}^c\mathrm{M}}^{i<c>}$$

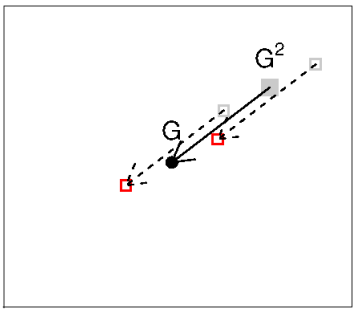


FIGURE 1.3 – Nuage intra–c2.

Par construction, le point G est le point moyen du sous–nuage intra–c, sa variance est égale à la variance du sous–nuage $\mathrm{M}^{I<c>}$.

Définition 2.3 (Nuage intra–C)**.** *Le nuage intra–C, noté* $\mathrm{M}^{I(C)}$*, est la réunion des* C *nuages intra–c ; il est pondéré par* n_I*.*

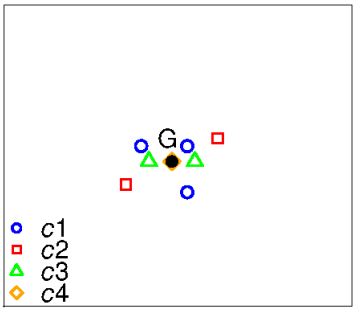


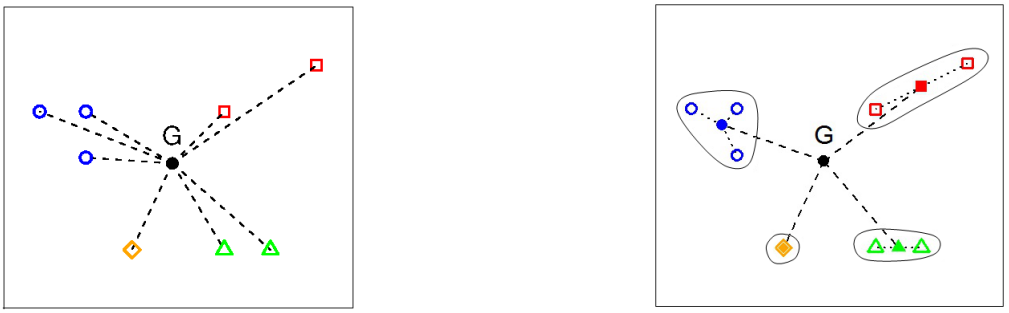
FIGURE 1.4 – Nuage intra–C.

Par construction, le nuage intra–C admet le point G pour point moyen. Sa variance, notée $\mathrm{Var}\,\mathrm{M}^{I(C)}$, est égale à la moyenne pondérée des variances des C sous–nuages $(\mathrm{M}^{I<c>})_{c\in C}$, c'est–à–dire à la *variance intra* :

$$\mathrm{Var}\,\mathrm{M}^{I(C)} = \sum_{c\in C} f_c\,\mathrm{Var}\,\mathrm{M}^{I<c>}$$

Décomposition inter–intra de la variance. La variance du nuage M^I est la somme de la variance inter–C et de la variance intra–C :

$$\underset{\text{Variance totale}}{\operatorname{Var}\mathrm{M}^I} = \underset{\text{Variance inter}}{\operatorname{Var}\mathrm{M}^C} + \underset{\text{Variance intra}}{\operatorname{Var}\mathrm{M}^{I(C)}}$$



Variance totale = Variance inter + Variance intra

3 Directions principales d'un nuage

Définition 3.1 (Application Som_{P})**.** *Soit* P *un point de* $\mathcal{U}$*, l'application* Som_{P} *est l'endomorphisme de* $\mathcal{V}$ *qui à tout vecteur* $\overrightarrow{\alpha}$ *de* $\mathcal{V}$ *fait correspondre le vecteur* $\sum_{i\in I} f_i\langle\overrightarrow{\mathrm{PM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{PM}^i}$ *:*

$$\begin{array}{rccc} Som_{\mathrm{P}}: & \mathcal{V} & \to & \mathcal{V} \\ & \overrightarrow{\alpha} & \mapsto & \sum_{i\in I} f_i\langle\overrightarrow{\mathrm{PM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{PM}^i} \end{array}$$

(*cf.* Le Roux & Rouanet, 2004, p.86–88 [50])

Cas particulier. Dans le cas particulier où P est le point moyen G du nuage, cet endomorphisme est simplement noté *Som*[2] :

Définition 3.2 (Endomorphisme de covariance *Som*)**.** *L'endomorphisme de covariance Som est l'endomorphisme de* $\mathcal{V}$ *qui à tout vecteur* $\overrightarrow{\alpha}$ *de* $\mathcal{V}$ *fait correspondre le vecteur* $\sum_{i\in I} f_i\langle\overrightarrow{\mathrm{GM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{GM}^i}$ *:*

$$\begin{array}{rccc} Som: & \mathcal{V} & \to & \mathcal{V} \\ & \overrightarrow{\alpha} & \mapsto & \sum_{i\in I} f_i\langle\overrightarrow{\mathrm{GM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{GM}^i} \end{array}$$

Propriété 3.1. *Som_{P} est un endomorphisme symétrique défini ou semi défini positif.*

Démonstration. $\forall\overrightarrow{\alpha},\overrightarrow{\beta}\in\mathcal{V}: \langle Som_{\mathrm{P}}(\overrightarrow{\alpha})|\overrightarrow{\beta}\rangle = \sum_{i\in I} f_i\langle\overrightarrow{\mathrm{PM}^i}|\overrightarrow{\alpha}\rangle\langle\overrightarrow{\mathrm{PM}^i}|\overrightarrow{\beta}\rangle = \langle\overrightarrow{\alpha}|Som_{\mathrm{P}}(\overrightarrow{\beta})\rangle$;

$\forall\overrightarrow{\alpha}\in\mathcal{V}\ :\ \langle Som_{\mathrm{P}}(\overrightarrow{\alpha})|\overrightarrow{\alpha}\rangle = \sum_{i\in I} f_i\langle\overrightarrow{\mathrm{PM}^i}|\overrightarrow{\alpha}\rangle^2 \geq 0.$

c.q.f.d.

Propriété 3.2.

$$\forall\overrightarrow{\alpha}\in\mathcal{V},\ Som_{\mathrm{P}}(\overrightarrow{\alpha}) = Som(\overrightarrow{\alpha}) + \langle\overrightarrow{\mathrm{PG}}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{PG}}$$

2. Défini et appelé par Benzécri « Application *Som* » (*cf.* Benzécri, 1980, p.148–150 [5]).

Démonstration. $Som_{\mathrm{P}}(\overrightarrow{\alpha}) = \sum_{i\in I} f_i \langle \overrightarrow{\mathrm{PM}}^i | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{PM}}^i$, en posant $\overrightarrow{\mathrm{PM}}^i = \overrightarrow{\mathrm{PG}} + \overrightarrow{\mathrm{GM}}^i$ (relation de Chasles), on obtient :

$$\begin{aligned} Som_{\mathrm{P}}(\overrightarrow{\alpha}) &= \sum_{i\in I} f_i (\langle \overrightarrow{\mathrm{PG}} + \overrightarrow{\mathrm{GM}}^i | \overrightarrow{\alpha}\rangle)(\overrightarrow{\mathrm{PG}} + \overrightarrow{\mathrm{GM}}^i) \\ &= \sum_{i\in I} f_i \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{PG}} + \sum_{i\in I} f_i \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{GM}}^i + \sum_{i\in I} f_i \langle \overrightarrow{\mathrm{GM}}^i | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{PG}} + \sum_{i\in I} f_i \langle \overrightarrow{\mathrm{GM}}^i | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{GM}}^i \end{aligned}$$

En vertu de la caractérisation barycentrique du point moyen, dans l'expression ci–dessus, les deuxième et troisième termes sont nuls. D'où :

$$Som_{\mathrm{P}}(\overrightarrow{\alpha}) = Som(\overrightarrow{\alpha}) + \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{PG}}$$

c.q.f.d.

Propriété 3.3. *La variance du nuage dans la direction $\overrightarrow{\alpha}$ est égale à :*

$$\frac{\langle Som(\overrightarrow{\alpha}) | \overrightarrow{\alpha}\rangle}{\| \overrightarrow{\alpha} \|^2}$$

Théorème 3.1 (Directions principales)**.** *Les vecteurs directeurs des droites principales du nuage, notés $(\overrightarrow{\alpha_\ell})_{\ell=1,\dots,L}$, sont vecteurs propres de l'endomorphisme Som :*

$$\forall \ell = 1, \dots L,\ Som(\overrightarrow{\alpha_\ell}) = \lambda_\ell \overrightarrow{\alpha_\ell}$$

Propriété 3.4. *$\forall\ \ell' \neq \ell$, les vecteurs propres $\overrightarrow{\alpha_\ell}$ et $\overrightarrow{\alpha_{\ell'}}$ sont orthogonaux.*

Démonstration. Découle directement de la propriété 3.1.

Lemme 3.1. *Si $\overrightarrow{\mathrm{PG}} = a\overrightarrow{\alpha_\ell}$ ($a \in \mathbb{R}$), alors Som_{P} a les mêmes vecteurs propres que Som associés aux mêmes valeurs propres, sauf la ℓ–ème qui vaut $\lambda_\ell + a^2 ||\overrightarrow{\alpha_\ell}||^2$.*

Démonstration. $\forall\ \ell' \neq \ell :\ Som_{\mathrm{P}}(\overrightarrow{\alpha_{\ell'}}) = Som(\overrightarrow{\alpha_{\ell'}}) + \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha_{\ell'}}\rangle \overrightarrow{\mathrm{PG}}$
$= Som(\overrightarrow{\alpha_{\ell'}}) + a^2 \langle \overrightarrow{\alpha_\ell} | \overrightarrow{\alpha_{\ell'}}\rangle \overrightarrow{\alpha_\ell}$
$\overrightarrow{\alpha_\ell}$ et $\overrightarrow{\alpha_{\ell'}}$ sont orthogonaux, donc $Som_{\mathrm{P}}(\overrightarrow{\alpha_{\ell'}}) = Som(\overrightarrow{\alpha_{\ell'}})$
$= \lambda_{\ell'} \overrightarrow{\alpha_{\ell'}}$

$$\begin{aligned} \forall\ \ell' = \ell :\ Som_{\mathrm{P}}(\overrightarrow{\alpha_\ell}) &= Som(\overrightarrow{\alpha_\ell}) + \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha_\ell}\rangle \overrightarrow{\mathrm{PG}} \\ &= Som(\overrightarrow{\alpha_\ell}) + a^2 \langle \overrightarrow{\alpha_\ell} | \overrightarrow{\alpha_\ell}\rangle \overrightarrow{\alpha_\ell} \\ &= \lambda_\ell \overrightarrow{\alpha_\ell} + a^2 ||\overrightarrow{\alpha_\ell}||^2 \overrightarrow{\alpha_\ell} \\ &= (\lambda_\ell + a^2 ||\overrightarrow{\alpha_\ell}||^2) \overrightarrow{\alpha_\ell} \end{aligned}$$

c.q.f.d.

4 Distances de Mahalanobis – Normes

Dans cette section, nous considérons la restriction des endomorphismes symétriques *Som* et Som_{P} au sous–espace vectoriel $\mathcal{L}$ associé au support $\mathcal{M}$ du nuage. *Som* et Som_{P} sont donc non singuliers.

Définition 4.1 (Norme de Mahalanobis)**.** *Si Som est non singulier, la norme, notée $|\ .\ |$, telle que :*

$$\forall \overrightarrow{\alpha} \in \mathcal{L},\ |\overrightarrow{\alpha}|^2 = \langle \overrightarrow{\alpha} | Som^{-1}(\overrightarrow{\alpha})\rangle$$

est appelée norme de Mahalanobis.

Définition 4.2 (Norme de Mahalanobis généralisée par rapport à un point P)**.** *Si Som_{P} est non singulier, la norme, notée $|\ .\ |_{\mathrm{P}}$, telle que :*

$$\forall \overrightarrow{\alpha} \in \mathcal{L},\ |\overrightarrow{\alpha}|^2_{\mathrm{P}} = \langle \overrightarrow{\alpha} | Som_{\mathrm{P}}^{-1}(\overrightarrow{\alpha})\rangle$$

est appelée norme de Mahalanobis généralisée par rapport au point P*.*

La distance de Mahalanobis, associée à l'endomorphisme *Som*, entre deux points A et B de $\mathcal{U}$ est notée $|\overrightarrow{\mathrm{AB}}|$, nous l'appelons M–*distance*. Nous définissons de même la distance entre A et B associée à l'endomorphisme Som_{P} (défini au paragraphe 3.1, p.16), nous l'appelons $\mathrm{M_P}$–*distance* et la notons $|\overrightarrow{\mathrm{AB}}|_{\mathrm{P}}$.

Propriété 4.1 (Propriété de réciprocité)**.** *Si Som est non singulier, on a la relation :*

$$|\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} = \frac{|\overrightarrow{\mathrm{PG}}|^2}{1 + |\overrightarrow{\mathrm{PG}}|^2}$$

ou, ce qui revient au même :

$$|\overrightarrow{\mathrm{PG}}|^2 = \frac{|\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}}}{1 - |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}}}$$

Démonstration. Soit $\overrightarrow{\alpha} = Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}})$, on a :

$$\begin{aligned} Som_{\mathrm{P}}(\overrightarrow{\alpha}) = \overrightarrow{\mathrm{PG}} &= Som(\overrightarrow{\alpha}) + \langle \overrightarrow{\mathrm{PG}} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{PG}} \quad \text{(propriété 3.2)} \\ \overrightarrow{\mathrm{PG}} &= Som(Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}})) + \langle \overrightarrow{\mathrm{PG}} | Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}})\rangle\ \overrightarrow{\mathrm{PG}} \\ &= Som(Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}})) + |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}}\ \overrightarrow{\mathrm{PG}} \end{aligned}$$

On en déduit :

$$Som^{-1}(\overrightarrow{\mathrm{PG}}) = Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}}) + |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}}\ Som^{-1}(\overrightarrow{\mathrm{PG}})$$

D'où :

$$\begin{aligned} |\overrightarrow{\mathrm{PG}}|^2 = \langle Som^{-1}(\overrightarrow{\mathrm{PG}}) | \overrightarrow{\mathrm{PG}}\rangle &= \langle Som_{\mathrm{P}}^{-1}(\overrightarrow{\mathrm{PG}}) | \overrightarrow{\mathrm{PG}}\rangle + |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}}\ \langle Som^{-1}(\overrightarrow{\mathrm{PG}}) | \overrightarrow{\mathrm{PG}}\rangle \\ &= |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} + |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} \times |\overrightarrow{\mathrm{PG}}|^2 \end{aligned}$$

c.q.f.d.

Propriété 4.2. *La quantité $\sum\limits_{i\in I} f_i |\overrightarrow{\mathrm{GM}}^i|^2$, appelée* M–*Variance, est égale à L.*

Démonstration. Considérons le repère principal normé $(\mathrm{G}, (\overrightarrow{\alpha_\ell})_{\ell=1,\dots L})$.
Dans ce repère, la M–*Variance* du nuage M^I, notée $\mathrm{Var_M}(\mathrm{M}^I)$ est définie par :

$$\mathrm{Var_M}(\mathrm{M}^I) = \sum_{i\in I} f_i |\overrightarrow{\mathrm{GM}}^i|^2 = \sum_{i\in I} f_i \langle Som^{-1}(\overrightarrow{\mathrm{GM}}^i) | \overrightarrow{\mathrm{GM}}^i\rangle$$

En notant y^i_ℓ la coordonnée du point M^i sur l'axe $(\mathrm{G}, (\overrightarrow{\alpha_\ell})_{\ell=1,\dots L})$, on a $\overrightarrow{\mathrm{GM}}^i = \sum\limits_{\ell=1}^{L} y^i_\ell \overrightarrow{\alpha_\ell}$ et $Som^{-1}(\overrightarrow{\mathrm{GM}}^i) = Som^{-1}(\sum\limits_{\ell=1}^{L} y^i_\ell \overrightarrow{\alpha_\ell}) = \sum\limits_{\ell=1}^{L} y^i_\ell Som^{-1}(\overrightarrow{\alpha_\ell})$.
On sait que $Som(\overrightarrow{\alpha_\ell}) = \overrightarrow{\alpha_\ell}\lambda_\ell$ (théorème 3.1), donc $Som^{-1}(\overrightarrow{\alpha_\ell}) = \overrightarrow{\alpha_\ell}/\lambda_\ell$ et $Som^{-1}(\overrightarrow{\mathrm{GM}}^i) = \sum\limits_{\ell=1}^{L} y^i_\ell \overrightarrow{\alpha_\ell}/\lambda_\ell$.
On a donc :

$$\begin{aligned}\sum_{i\in I} f_i\langle Som^{-1}(\overrightarrow{\mathrm{GM}^i})|\overrightarrow{\mathrm{GM}^i}\rangle &= \sum_{i\in I} f_i\langle \sum_{\ell=1}^{L} y_\ell^i\overrightarrow{\alpha_\ell}/\lambda_\ell| \sum_{\ell'=1}^{L} y_{\ell'}^i\overrightarrow{\alpha_{\ell'}}\rangle \\ &= \sum_{i\in I} f_i \sum_{\ell=1}^{L} y_\ell^i/\lambda_\ell \sum_{\ell'=1}^{L} y_{\ell'}^i\langle\overrightarrow{\alpha_\ell}|\overrightarrow{\alpha_{\ell'}}\rangle \\ &\quad (\forall\ \ell\neq\ell',\ \langle\overrightarrow{\alpha_\ell}|\overrightarrow{\alpha_{\ell'}}\rangle = 0 \text{ car les vecteurs } \overrightarrow{\alpha_\ell} \text{ et } \overrightarrow{\alpha_{\ell'}} \text{ sont orthogonaux,} \\ &\quad \forall\ \ell=\ell',\ \langle\overrightarrow{\alpha_\ell}|\overrightarrow{\alpha_{\ell'}}\rangle = 1 \text{ car les vecteurs } \overrightarrow{\alpha_\ell} \text{ et } \overrightarrow{\alpha_{\ell'}} \text{ sont normés}) \\ &= \sum_{\ell=1}^{L} \tfrac{1}{\lambda_\ell} \sum_{i\in I} f_i (y_\ell^i)^2\end{aligned}$$

D'après la propriété 3.3, la variance du nuage M^I dans la direction $\overrightarrow{\alpha_\ell}$, c'est–à–dire la variance de la variable y_ℓ^I est : $\mathrm{Var}\, y_\ell^I = \sum_{i\in I} f_i (y_\ell^i)^2 = \frac{\langle Som(\overrightarrow{\alpha_\ell})|\overrightarrow{\alpha_\ell}\rangle}{\|\overrightarrow{\alpha_\ell}\|^2} = \frac{\langle\lambda_\ell\overrightarrow{\alpha_\ell}|\overrightarrow{\alpha_\ell}\rangle}{\|\overrightarrow{\alpha_\ell}\|^2} = \lambda_\ell$.

D'où :

$$\sum_{i\in I} f_i\langle Som^{-1}(\overrightarrow{\mathrm{GM}^i})|\overrightarrow{\mathrm{GM}^i}\rangle = L.$$

c.q.f.d.

Propriété de réciprocité généralisée.

Soit f l'endomorphisme de l'espace vectoriel $\mathcal{V}$ défini par :

$$\forall x\in\mathcal{V},\ f(\overrightarrow{x}) = \overrightarrow{x} + a\langle\overrightarrow{v}|\overrightarrow{x}\rangle\overrightarrow{u}$$

où $\overrightarrow{u}$ et $\overrightarrow{v}$ sont deux vecteurs non orthogonaux de $\mathcal{V}$ et où a est un nombre réel.

Propriété 4.3. *Si* $1 + a\langle\overrightarrow{v}|\overrightarrow{u}\rangle \neq 0$ *alors*

$$f^{-1}(\overrightarrow{x}) = \overrightarrow{x} - \frac{a\langle\overrightarrow{v}|\overrightarrow{x}\rangle}{1+a\langle\overrightarrow{v}|\overrightarrow{u}\rangle}\overrightarrow{u}$$

Démonstration. On a $f(\overrightarrow{u}) = (1 + a\langle\overrightarrow{v}|\overrightarrow{u}\rangle)\overrightarrow{u}$, donc $\overrightarrow{u}$ est vecteur propre de f associé à la valeur propre $1 + a\langle\overrightarrow{v}|\overrightarrow{u}\rangle$.
Pour tout vecteur $\overrightarrow{y}$ orthogonal à $\overrightarrow{v}$, on a $f(\overrightarrow{y}) = \overrightarrow{y}$, donc $\overrightarrow{v}^\perp$ est sous–espace propre de f associé à la valeur propre 1.
Nous avons donc décomposé l'espace vectoriel $\mathcal{V}$ de dimension K en un sous–espace de dimension 1 et un hyperplan de dimension $K-1$.
Si $1 + a\langle\overrightarrow{v}|\overrightarrow{u}\rangle \neq 0$, toutes les valeurs propres de f étant non nulles, f est inversible et son inverse a les mêmes vecteurs propres associés aux valeurs propres inverses, d'où :
$f^{-1}(\overrightarrow{u}) = \frac{1}{1+a\langle\overrightarrow{v}|\overrightarrow{u}\rangle}\overrightarrow{u} = (1 - \frac{a\langle\overrightarrow{v}|\overrightarrow{u}\rangle}{1+a\langle\overrightarrow{v}|\overrightarrow{u}\rangle})\overrightarrow{u}$ et $f^{-1}(\overrightarrow{y}) = \overrightarrow{y}$.
$\forall\overrightarrow{x}\in\mathcal{V}$, comme $1 + a\langle\overrightarrow{v}|\overrightarrow{u}\rangle \neq 0$, on a $f^{-1}(\overrightarrow{x}) = \overrightarrow{x} - \frac{a\langle\overrightarrow{v}|\overrightarrow{x}\rangle}{1+a\langle\overrightarrow{v}|\overrightarrow{u}\rangle}\overrightarrow{u}$.
c.q.f.d.

Soient f_1 et f_2 deux endomorphismes non singuliers de $\mathcal{V}$ tels que :

$$\forall x\in\mathcal{V},\ f_1(\overrightarrow{x}) = f_2(\overrightarrow{x}) + \gamma\langle\overrightarrow{e}|\overrightarrow{x}\rangle\overrightarrow{e}$$

où $\overrightarrow{e}$ est un vecteur de $\mathcal{V}$ et $\gamma\in\mathbb{R}$. Notons $|\overrightarrow{e}|_{f_1}^2 = \langle f_1^{-1}(\overrightarrow{e})|\overrightarrow{e}\rangle$ et $|\overrightarrow{e}|_{f_2}^2 = \langle f_2^{-1}(\overrightarrow{e})|\overrightarrow{e}\rangle$.

Propriété 4.4.

$$\forall\overrightarrow{e}\in\mathcal{V},\ |\overrightarrow{e}|_{f_1}^2 = \frac{|\overrightarrow{e}|_{f_2}^2}{1+\gamma|\overrightarrow{e}|_{f_2}^2}$$

ou, ce qui revient au même :

$$\forall\overrightarrow{e}\in\mathcal{V},\ |\overrightarrow{e}|_{f_2}^2 = \frac{|\overrightarrow{e}|_{f_1}^2}{1-\gamma|\overrightarrow{e}|_{f_1}^2}$$

Démonstration. $\forall \overrightarrow{x} \in \mathcal{V}$, $\forall \overrightarrow{e} \in \mathcal{V}$,

$$\begin{aligned} f_1(\overrightarrow{x}) &= f_2(\overrightarrow{x}) + \gamma \langle \overrightarrow{e} | \overrightarrow{x} \rangle \overrightarrow{e} \\ &= f_2(\overrightarrow{x} + \gamma \langle \overrightarrow{e} | \overrightarrow{x} \rangle f_2^{-1}(\overrightarrow{e})) \\ &= f_2(f(\overrightarrow{x})) \end{aligned}$$

où $f(\overrightarrow{x})$ est analogue à l'application donnée avant la propriété 4.3, avec $a = \gamma$, $\overrightarrow{v} = \overrightarrow{e}$ et $\overrightarrow{u} = f_2^{-1}(\overrightarrow{e})$.
D'après la propriété 4.3, pour tout $\overrightarrow{e}$ tel que $1 + \gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle \neq 0$:

$$f_1^{-1}(\overrightarrow{x}) = f^{-1}(f_2^{-1}(\overrightarrow{x})) = f_2^{-1}(\overrightarrow{x}) - \frac{\gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{x}) \rangle}{1 + \gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle} f_2^{-1}(\overrightarrow{e})$$

D'où :

$$\begin{aligned} \langle f_1^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle &= \langle f_2^{-1}(\overrightarrow{e}) - \tfrac{\gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle}{1+\gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle} f_2^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle \\ &= \langle f_2^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle - \tfrac{\gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle}{1+\gamma \langle \overrightarrow{e} | f_2^{-1}(\overrightarrow{e}) \rangle} \langle f_2^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle \\ &= \tfrac{\langle f_2^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle}{1+\gamma \langle f_2^{-1}(\overrightarrow{e}) | \overrightarrow{e} \rangle} \end{aligned}$$

et donc : $|\overrightarrow{e}|_{f_1}^2 = \frac{|\overrightarrow{e}|_{f_2}^2}{1+\gamma |\overrightarrow{e}|_{f_2}^2}$.
c.q.f.d.

Remarque : Cette propriété est générale. Si f_1 et f_2 sont des endomorphismes non singuliers, symétriques et définis positifs de $\mathcal{V}$ (ou $\mathcal{L}$), ils définissent une norme.

5 Ellipsoïdes d'inertie

Définition 5.1 (κ–ellipsoïde d'inertie)**.** *L'ensemble des points* P *de* $\mathcal{M}$ *tels que :*

$$|\overrightarrow{\mathrm{GP}}| = \kappa \quad (\textit{avec } \kappa > 0)$$

est appelé κ–ellipsoïde d'inertie du nuage M^I*.*

Lorsque $\kappa = 1$, il s'agit de l'ellipsoïde indicateur. Lorsque $\kappa = \sqrt{L+2}$, il s'agit de l'ellipsoïde de concentration (*cf.* Cramér, 1946, p.283 [20] ; Malinvaud, 1980 [57] et Anderson, 1958, p.44 [3]). Pour un nuage plan, pour $\kappa = 1$ et $\kappa = 2$, on obtient respectivement l'*ellipse indicatrice* et l'*ellipse de concentration* du nuage (figure ci–après).

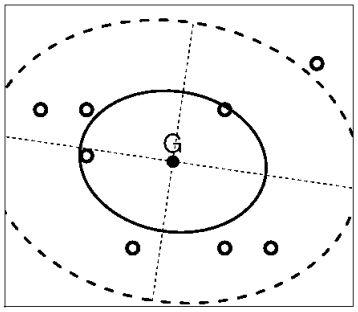


FIGURE 1.5 – Ellipse indicatrice (trait plein) et ellipse de concentration (pointillés) du nuage M^I.

Propriété 5.1. *Si* P *et* P′ *sont deux points de* $\mathcal{M}$ *appartenant au même κ–ellipsoïde d'inertie du nuage* M^I*, alors :*

$$|\overrightarrow{\mathrm{PG}}|_{\mathrm{P}} = |\overrightarrow{\mathrm{P'G}}|_{\mathrm{P'}}$$

Démonstration. Les points P et P′ appartiennent au même κ–ellipsoïde d'inertie du nuage initial si et seulement si $|\overrightarrow{\mathrm{PG}}| = |\overrightarrow{\mathrm{P'G}}| = \kappa$. D'après la propriété 4.1 de réciprocité, on a :

$$|\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} = \frac{|\overrightarrow{\mathrm{PG}}|^2}{1+|\overrightarrow{\mathrm{PG}}|^2} = \frac{\kappa^2}{1+\kappa^2}$$

et de même

$$|\overrightarrow{\mathrm{P'G}}|^2_{\mathrm{P'}} = \frac{|\overrightarrow{\mathrm{P'G}}|^2}{1+|\overrightarrow{\mathrm{P'G}}|^2} = \frac{\kappa^2}{1+\kappa^2}$$

c.q.f.d.

6 Formulations numériques et matricielles

Munissons maintenant $\mathcal{U}$ du repère cartésien $(\mathrm{O},(\overrightarrow{\delta^k})_{k\in K})$. Notons $\mathbf{Q} = [q^{kk'}]$ la matrice $K \times K$ des produits scalaires des vecteurs de la base ayant donc pour terme général $q^{kk'} = \langle \overrightarrow{\delta^k} | \overrightarrow{\delta^{k'}} \rangle$ et $(x^i_k)_{k\in K}$ les coordonnées du point M^i dans ce repère. La variable $x^I_k = (x^i_k)_{i\in I}$ est appelée k–ème *variable coordonnée* ou *initiale*.
On a :

- $\overrightarrow{\mathrm{OM}}^i = \sum\limits_{k\in K} x^i_k \, \overrightarrow{\delta^k}$
- $\overrightarrow{\mathrm{OG}} = \sum\limits_{k\in K} \overline{x}_k \, \overrightarrow{\delta^k}$ avec $\overline{x}_k = \sum\limits_{i\in I} f_i \, x^i_k$
- $v_{kk'} = \sum\limits_{i\in I} f_i (x^i_k - \overline{x}_k)(x^i_{k'} - \overline{x}_{k'})$ (covariance entre les variables coordonnées x^I_k et $x^I_{k'}$), notons $\mathbf{V} = [v_{kk'}]$ la matrice $K \times K$ de covariance.

On vérifie facilement que $\mathrm{Var}\,\mathrm{M}^I = \sum\limits_{k\in K}\sum\limits_{k'\in K} v_{kk'}\, q^{kk'} = \mathrm{Tr}(\mathbf{VQ})$.

Dans la base $(\overrightarrow{\delta^k})_{k\in K}$, la matrice de l'endomorphisme *Som* est la matrice $\mathbf{VQ}$ et celle de l'endomorphisme Som_{P} est égale à $(\mathbf{V} + \mathbf{dd'})\mathbf{Q}$, où $\mathbf{d} = [d_k]$ est le vecteur–colonne associé au vecteur $\overrightarrow{\mathrm{PG}} = \sum\limits_{k\in K} d_k \delta^k$.

La matrice $\mathbf{B}_{\mathrm{P}} = \mathbf{V} + \mathbf{dd'}$, de terme général $v_{kk'} + d_k d_{k'} = \sum\limits_{i\in I} f_i (x^i_k - a_k)(x^i_{k'} - a_{k'})$ (avec $a_k = \overline{x}_k + d_k$) est la matrice des moyennes de carrés et produits des écarts des points du nuage M^I au point P.

Cas particulier d'une base orthonormée.

Si la base est orthonormée alors $\mathbf{Q} = \mathbf{I}$: les formules se simplifient, la matrice de *Som* est égale à $\mathbf{V}$ et celle de Som_{P} à $\mathbf{B}_{\mathrm{P}}$.

Dans ce cas, si $\mathbf{V}$ est une matrice non singulière, $\forall \overrightarrow{u} \in \mathcal{V}$,

- la norme associée à l'endomorphisme *Som*, ou $\mathbf{V}$–norme est telle que :

$$|\overrightarrow{u}|^2 = \mathbf{u'V^{-1}u}$$

- pour tout $\mathrm{P} \in \mathcal{M}$, la norme associée à l'endomorphisme Som_{P}, ou $\mathbf{B}_{\mathrm{P}}$–norme, est telle que :

$$|\overrightarrow{u}|^2_{\mathrm{P}} = \mathbf{u'B_{\mathrm{P}}^{-1}u}$$

Considérons maintenant le support $\mathcal{M}$ du nuage rapporté aux axes principaux du nuage et la base principale orthonormée $(\overrightarrow{\alpha_\ell})_{\ell=1,\dots L}$ du sous–espace $\mathcal{L}$ avec donc $\lambda_\ell > 0$; $(\ell = 1, \dots L)$. L'inverse de *Som* restreint à ce sous–espace est défini par :

$$\forall \ell = 1, \dots L,\ Som^{-1}(\overrightarrow{\alpha_\ell}) = \frac{\overrightarrow{\alpha_\ell}}{\lambda_\ell}$$

avec $\overrightarrow{\alpha_\ell}$ vecteur propre de l'endomorphisme *Som* associé à la valeur propre λ_ℓ, et donc :

$$\forall \overrightarrow{u} \in \mathcal{L} \text{ tel que } \overrightarrow{u} = \sum_{\ell \in L} u_\ell \overrightarrow{\alpha_\ell},\ |\overrightarrow{u}|^2 = \sum_{\ell \in L} \frac{u_\ell^2}{\lambda_\ell}$$

En vertu de la propriété 4.1 de réciprocité, la norme du vecteur $\overrightarrow{\mathrm{PG}} = \sum_{\ell \in L} d_\ell \overrightarrow{\alpha_\ell}$ associée à Som_{P} est telle que :

$$|\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} = \frac{|\overrightarrow{\mathrm{PG}}|^2}{1 + |\overrightarrow{\mathrm{PG}}|^2} = \frac{\sum_{\ell \in L} d_\ell^2/\lambda_\ell}{1 + \sum_{\ell \in L} d_\ell^2/\lambda_\ell}$$

Remarque : Pour évaluer la grandeur d'un écart d, il est usuel en statistique univariée de rapporter cet écart d à l'écart–type $\sqrt{v}$. La quantité $d/\sqrt{v}$ est appelée écart calibré. La formule précédente montre que la distance de Mahalanobis est l'extension multivariée de l'écart calibré.

Propriété de réciprocité généralisée. Soient :

- $\mathbf{e}$ le vecteur–colonne associé au vecteur $\overrightarrow{e} \in \mathcal{V}$,
- γ un scalaire,
- $\mathbf{M_1}$ et $\mathbf{M_2}$ les deux matrices $K \times K$ non singulières associées aux endomorphismes f_1 et f_2 définis au paragraphe 4 (p.19). On a par conséquent :

$$\mathbf{M_1} = \mathbf{M_2} + \gamma \mathbf{e}\mathbf{e}'$$

Notons $|\overrightarrow{e}|_{\mathbf{M_1}}$ et $|\overrightarrow{e}|_{\mathbf{M_2}}$ les normes du vecteur $\overrightarrow{e}$ associées respectivement aux endomorphismes f_1 et f_2. On a donc $|\overrightarrow{e}|^2_{\mathbf{M_1}} = \mathbf{e}'\mathbf{M_1^{-1}}\mathbf{e}$ et $|\overrightarrow{e}|^2_{\mathbf{M_2}} = \mathbf{e}'\mathbf{M_2^{-1}}\mathbf{e}$;
La propriété 4.4 peut s'écrire :

$$\forall \overrightarrow{e} \in \mathcal{V},\ |\overrightarrow{e}|^2_{\mathbf{M_1}} = \frac{|\overrightarrow{e}|^2_{\mathbf{M_2}}}{1 + \gamma |\overrightarrow{e}|^2_{\mathbf{M_2}}}$$

ou, ce qui revient au même :

$$|\overrightarrow{e}|^2_{\mathbf{M_2}} = \frac{|\overrightarrow{e}|^2_{\mathbf{M_1}}}{1 - \gamma |\overrightarrow{e}|^2_{\mathbf{M_1}}}$$

Dans toute la suite nous considérons des nuages élémentaires, c'est–à–dire tels que :

$$\forall i \in I,\ n_i = 1$$

De même, dans les chapitres suivants, les endomorphismes considérés sont souvent restreints à $\mathcal{L}$ ou à un sous–espace de $\mathcal{L}$, ils sont donc non singuliers.

Chapitre 2

Test géométrique de typicalité

Dans ce chapitre, nous présentons le *test géométrique de typicalité*, en bref *test géométrique*. Il consiste à comparer le point moyen d'un groupe d'observations à un point de référence ou dans le cas unidimensionnel, la moyenne d'un groupe d'observations à une valeur de référence. Nous l'appelons aussi : *test de typicalité par rapport à un point de référence*.

Pour fixer les idées, prenons l'exemple du « paradigme cible » (*cf.* Le Roux & Rouanet, 2004 [50]) : on dispose d'un groupe d'observations constitué de points d'impact sur une cible. Ces points forment un nuage plan dont le point moyen est appelé G. Nous nous proposons de comparer le point G au centre O de la cible. En d'autres termes, nous nous demandons si le point effectivement visé est le point O, ou non. Le test géométrique de typicalité repose sur le principe suivant : si O est le point visé, alors chaque point observé aurait aussi bien pu se situer au niveau de son symétrique par rapport à O. Ce principe conduit à construire l'*espace des nuages possibles* engendré par la symétrie de centre O. Il s'agit ensuite de situer le groupe d'observations par rapport à cet espace, selon la statistique de test choisie.

Une des applications privilégiées du test géométrique est celle où l'on dispose des données de deux groupes appariés. Le groupe d'observations est alors constitué des données « écarts », ou « différences », entre les deux groupes, le point de référence est le point représentant l'*écart nul*. Un exemple d'une telle situation est le cas où des patients ont été observés avant et après traitement. Nous évaluons l'effet du traitement en comparant les écarts *Après* − *Avant* à l'écart nul. Le principe du test se retrouve ici : intuitivement, dire que le traitement est sans effet, c'est dire que les données observées avant traitement auraient aussi bien pu être observées après traitement. L'écart observé *Après* − *Avant* aurait donc aussi bien pu être égal à l'écart *Avant* − *Après*. L'espace des possibles se dessine ici, l'écart *Avant* − *Après* étant le symétrique de l'écart *Après* − *Avant* par rapport au point représentant l'écart nul. Nous montrons que, dans le cas de deux groupes appariés, le test géométrique est équivalent à un test de permutation (*cf.* paragraphe 1.3, p.37).

Dans ce chapitre, nous exposons le test géométrique de typicalité dans le cas multidimensionnel puis nous l'appliquons, sans perte de généralité, à des nuages à deux dimensions. Nous comparons également les résultats du test géométrique avec ceux obtenus par d'autres procédures inférentielles basées d'une part, de façon traditionnelle, sur le modèle normal et d'autre part, sur une procédure de rééchantillonnage bootstrap.

1 Typicalité d'un point moyen par rapport à un point de référence

Dans cette section, nous nous proposons d'étudier la typicalité d'un point moyen par rapport à un point de référence.

Situation de base. Considérons un ensemble $I = \{1, \ldots, i, \ldots, n\}$ de n individus formant un *groupe d'observations*. A l'ensemble I est associé le nuage M^I de n points (individus) à valeurs dans un espace affine euclidien $\mathcal{U}$ de dimension K. Le point moyen du nuage M^I est noté G, sa structure de covariance est définie par l'endomorphisme de covariance *Som* (défini au chapitre 1, p.16). Considérons également un point O de $\mathcal{U}$ et prenons le comme point de référence.

Nous nous posons la question suivante : *Le point moyen du groupe d'observations est-il atypique du point de référence ?*

Dans la suite, nous considérons que les points du nuage M^I *sont à valeurs dans* $\mathcal{M}$*, son support affine. Le sous–espace vectoriel directeur de* $\mathcal{M}$*, de dimension* L*, est noté* $\mathcal{L}$ *(cf. chapitre 1, p.12).*

Modélisation statistique. L'idée fondamentale en inférence statistique est la suivante : ce que l'on observe est une des réalisations de toutes les réalisations possibles. Dans le cas (multidimensionnel) où nous comparons le point moyen G d'un nuage à un point de référence O, le point moyen observé G est une des réalisations d'une variable *Point Moyen*, notée G (en italique), à valeurs dans un *espace des possibles* que nous construisons. La variable G est supposée varier autour d'un *vrai* point moyen Γ de telle sorte que $G = \Gamma + \overrightarrow{e}$ où $\overrightarrow{e}$ est un terme d'erreur.

1.1 Test exact

1.1.1 Principe du test

La question posée précédemment peut-être reformulée ainsi : *L'hypothèse « le vrai point moyen est le point de référence* O *»* ($\Gamma = \mathrm{O}$) *est-elle compatible avec les données, ou non ?*

Pour répondre à cette question, nous construisons le test géométrique de la façon suivante :

1. Sous l'hypothèse que le vrai point moyen est le point de référence O ($\Gamma = \mathrm{O}$), le point observé M^i pourrait aussi bien être son symétrique par rapport à O, c'est–à–dire le point N^i tel que :
$$\overrightarrow{\mathrm{ON}}^i = -\overrightarrow{\mathrm{OM}}^i$$
En effectuant les $\mathrm{J} = 2^n$ échanges possibles de $1, 2, \ldots n$ points observés avec leurs symétriques, nous obtenons 2^n nuages de n points. Soit $J = \{1, \ldots, j, \ldots \mathrm{J}\}$ l'ensemble indexant les $\mathrm{J} = 2^n$ nuages [1], le j–ème nuage est noté $(\mathrm{M}^{Ij})_{j=1,\ldots 2^n}$. L'ensemble des $\mathrm{J} = 2^n$ nuages possibles est noté $\mathcal{J}$ et appelé *espace de typicalité*.

1. La plupart du temps dans ce chapitre, l'ensemble et son cardinal sont identifiés par la même lettre, l'ensemble en italique et son cardinal en lettre droite.

2. (a) Considérons maintenant l'application C qui au nuage M^{Ij} associe son point moyen C^j :

$$\begin{array}{rccc} C: & \mathcal{J} & \to & \mathcal{M} \\ & \mathrm{M}^{Ij} & \mapsto & \mathrm{C}^j = \sum_{i\in I} \mathrm{M}^{ij}/n \end{array}$$

Le nuage C^J des J points moyens, muni de la pondération élémentaire, est appelé *nuage de typicalité*.

(b) Faisons ensuite choix d'une statistique de test (*cf. Remarque sur le choix de la statistique de test, p.29*), ici la statistique D_{O}, notée en bref D, définie de la façon suivante : au point moyen C^j est associée sa M_{O}–distance au point de référence O (*cf.* chapitre 1, p.18, avec P = O) :

$$\begin{array}{rccc} D: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{C}^j & \mapsto & |\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}} \end{array}$$

Notons d_{obs} la valeur observée de cette statistique : $d_{obs} = |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}}$.

3. Déterminons enfin la proportion des nuages M^{Ij} pour lesquels la valeur de la statistique D est supérieure ou égale à la valeur observée d_{obs}. Cette proportion définit le seuil observé du test exact (seuil exact ou p–value exacte) et est notée p_{obs} :

$$p_{obs} = p(D \geq d_{obs})$$

Nous pouvons désormais énoncer la conclusion du test en termes d'atypicalité du point moyen du groupe d'observations par rapport au point de référence.

Si $p_{obs} \leq \alpha$ (seuil $\alpha < 1/2$ fixé [2]), le résultat du test est significatif au seuil α. Pour la statistique D, nous pouvons dire que les données sont en faveur d'un point moyen vrai différent du point O. Par conséquent, nous pouvons dire que le point moyen du groupe d'observations est atypique du point de référence au seuil α.

Si $p_{obs} > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique D, nous ne pouvons pas dire que les données sont en faveur d'un point moyen vrai différent du point O. Par conséquent, nous ne pouvons pas dire que le point moyen du groupe d'observations est atypique du point de référence au seuil α.

Remarque : A partir de la distribution de la statistique de test D, nous pouvons définir la valeur critique d_α au seuil α telle que :

$$p(D \geq d_\alpha) = \alpha$$

Le résultat du test s'énonce alors comme ceci :

- si $d_{obs} \geq d_\alpha$, alors le résultat du test est significatif au seuil α,
- si $d_{obs} < d_\alpha$, alors le résultat du test n'est pas significatif au seuil α.

Exemple Cible : Afin d'illustrer les notions inhérentes au test géométrique, nous utilisons comme fil conducteur l'exemple *Cible* introduit ci–après.

2. On utilise souvent les seuils conventionnels $\alpha = .05$ et $\alpha = .01$.

Considérons le nuage plan M^I de 10 points, de point moyen G, représentant des points d'impact sur une cible de centre O [3]. Nous nous demandons si le point effectivement visé est le point O, en d'autres termes, si le point G est atypique ou non du point de référence O.

Nous considérons que l'écart descriptif entre les points O et G est [4] :

- négligeable si $|\overrightarrow{OG}| < 0.6$
- notable si $0.6 \leq |\overrightarrow{OG}| < 0.8$,
- important si $|\overrightarrow{OG}| \geq 0.8$.

Pour cet exemple, on a :

$$|\overrightarrow{OG}| = \sqrt{\mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs}} = 0.963$$

L'écart observé est donc important, nous tentons maintenant de prolonger cette conclusion descriptive en mettant en place le test géométrique.

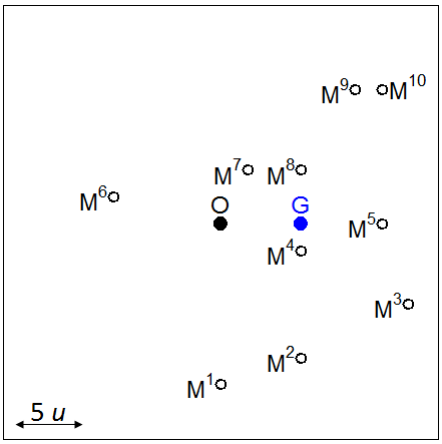


FIGURE 2.1 – *Exemple Cible.* Nuage M^I des 10 points d'impact. Pour les calculs numériques, nous prenons u comme unité de longueur.

Pour cet exemple, la construction de l'espace de typicalité et des points moyens associés est résumée sur la figure 2.2 ci–après. L'ensemble de ces points moyens constitue le nuage de typicalité.

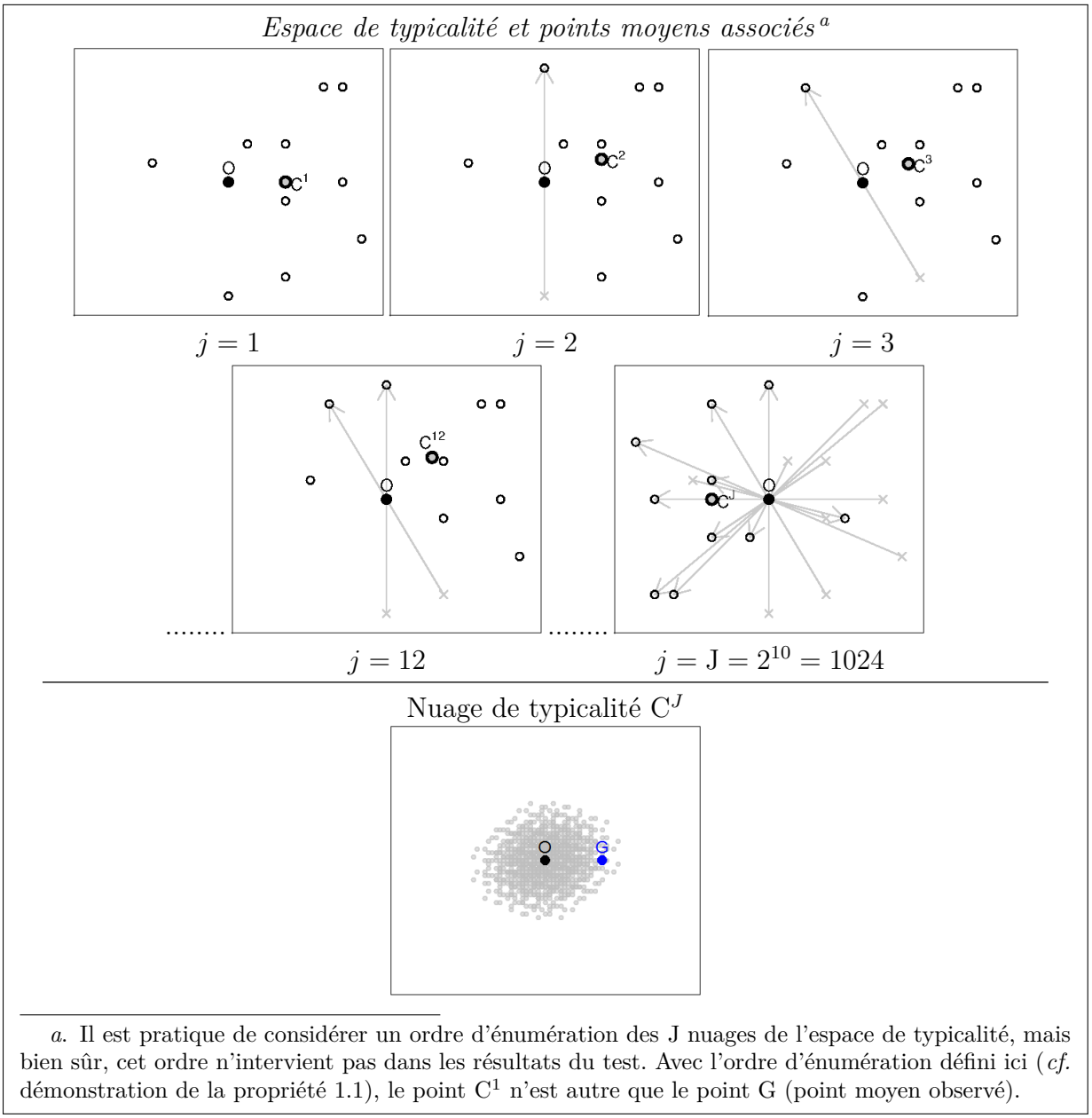
3. Tous les calculs ont été effectués en choisissant deux axes perpendiculaires, l'un horizontal (x_1), l'autre vertical (x_2), gradués selon l'unité de distance u (*cf.* figure 2.1, p.26) et dont l'origine est O, le centre de la cible. Ci–après sont reportés le tableau de nombres associés aux 10 points, la matrice de covariance du nuage M^I, les coordonnées du vecteur $\overrightarrow{OG}$ et la matrice $\mathbf{B}_O$:

x_1	0	6	14	6	12	-8	2	6	10	12	Moyenne 6
x_2	-12	-10	-6	-2	0	2	4	4	10	12	0

$\mathbf{V} = \begin{pmatrix} 40 & 8 \\ 8 & 52 \end{pmatrix}$,

$\mathbf{d}_{obs} = \overrightarrow{OG} = \begin{pmatrix} 6 \\ 0 \end{pmatrix}$, $\mathbf{B}_O = \mathbf{V} + \mathbf{d}_{obs}\ \mathbf{d}'_{obs} = \begin{pmatrix} 76 & 8 \\ 8 & 52 \end{pmatrix}$.

4. Cet indice est une généralisation multidimensionnelle de l'écart calibré (*cf.* chapitre 5, p.153). Les seuils figurant ici sont donnés dans Le Roux & Rouanet, 2004 [50].



a. Il est pratique de considérer un ordre d'énumération des J nuages de l'espace de typicalité, mais bien sûr, cet ordre n'intervient pas dans les résultats du test. Avec l'ordre d'énumération défini ici (*cf.* démonstration de la propriété 1.1), le point C^1 n'est autre que le point G (point moyen observé).

FIGURE 2.2 – *Exemple Cible.* Construction de l'espace de typicalité, des points moyens associés et du nuage de typicalité.

Caractéristiques du nuage de typicalité et de la distribution de la statistique de test D.

Propriété 1.1. *Le point moyen du nuage de typicalité* C^J *est le point de référence* O.

Démonstration. Posons $\varepsilon_i^j = -1$ si pour l'échange j le point M^i est remplacé par son symétrique et $\varepsilon_i^j = +1$ sinon. Considérons l'ordre d'énumération suivant : $\varepsilon_i^j = +1$ si dans l'expression de l'entier $j-1$ en numération binaire, le coefficient de $2^{(i-1)}$ est 0 et $\varepsilon_i^j = -1$ si ce coefficient est 1.
Considérons maintenant les $J = 2^n$ points $(C^j)_{j\in J}$ définis par :

$$\overrightarrow{OC^j} = \sum_{i\in I} \varepsilon_i^j \overrightarrow{OM^i}/n$$

Le point moyen du nuage C^J, noté $\overline{C}$, est tel que :

$$\begin{aligned}\overrightarrow{\mathrm{OC}} &= \tfrac{1}{2^n}\sum_{j\in J}\overrightarrow{\mathrm{OC}^j}\\ &= \tfrac{1}{2^n}\sum_{j\in J}(\sum_{i\in I}\tfrac{1}{n}\varepsilon_i^j\overrightarrow{\mathrm{OM}^i})\\ &= \tfrac{1}{2^n}\sum_{i\in I}\tfrac{1}{n}(\sum_{j\in J}\varepsilon_i^j\overrightarrow{\mathrm{OM}^i})\end{aligned}$$

or pour i fixé, $\sum_{j\in J}\varepsilon_i^j = 0$, donc $\overrightarrow{\mathrm{OC}} = \overrightarrow{0}$.

c.q.f.d.

Soient Som^{C} l'endomorphisme dont les vecteurs propres déterminent les directions principales du nuage de typicalité C^J et Som_{O} l'endomorphisme associé au nuage M^I tel que défini page 16 (définition 3.1 avec P = O).

Propriété 1.2. *L'endomorphisme Som^{C} est proportionnel à l'endomorphisme Som_{O} :*

$$Som^{\mathrm{C}} = \frac{Som_{\mathrm{O}}}{n}$$

Démonstration. D'après le théorème 3.1 (p.17), les directions principales du nuage C^J sont engendrées par les vecteurs propres de l'endomorphisme Som^{C} défini par :

$$\forall\overrightarrow{\alpha}\in\mathcal{L},\ Som^{\mathrm{C}}(\overrightarrow{\alpha}) = \frac{1}{\mathrm{J}}\sum_{j\in J}\langle\overrightarrow{\mathrm{OC}^j}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{OC}^j}$$

En reprenant les notations de la démonstration de la propriété 1.1 (p.27), on a :

$$\begin{aligned}Som^{\mathrm{C}}(\overrightarrow{\alpha}) &= \tfrac{1}{\mathrm{J}}\sum_{j\in J}\langle\sum_{i\in I}\tfrac{1}{n}(\varepsilon_i^j\overrightarrow{\mathrm{OM}^i})|\overrightarrow{\alpha}\rangle\sum_{i'\in I}\tfrac{1}{n}(\varepsilon_{i'}^j\overrightarrow{\mathrm{OM}^{i'}})\\ &= \tfrac{1}{\mathrm{J}}\tfrac{1}{n^2}\sum_{i\in I}\sum_{i'\in I}\Big(\sum_{j\in J}\varepsilon_i^j\varepsilon_{i'}^j\Big)\langle\overrightarrow{\mathrm{OM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{OM}^{i'}}\end{aligned}$$

Pour $i = i'$, $\sum_{j\in J}(\varepsilon_i^j)^2 = 2^n$ (car $\forall i,\forall j,(\varepsilon_i^j)^2 = 1$) ; pour $i\neq i'$, le couple $(\varepsilon_i^j,\varepsilon_{i'}^j)$ prend les valeurs $(1,1)$, $(1,-1)$, $(-1,1)$ et $(-1,-1)$, chacune 2^{n-2} fois.
D'où $\sum_{j\in J}\varepsilon_i^j\varepsilon_{i'}^j = 2^{n-2}\times\Big(1\times1+1\times(-1)+(-1)\times1+(-1)\times(-1)\Big) = 0$.

Par conséquent $Som^{\mathrm{C}}(\overrightarrow{\alpha}) = \frac{1}{2^n}\times 2^n\times\frac{1}{n}\times\Big(\sum_{i\in I}\frac{1}{n}\langle\overrightarrow{\mathrm{OM}^i}|\overrightarrow{\alpha}\rangle\overrightarrow{\mathrm{OM}^i}\Big) = \frac{1}{n}Som_{\mathrm{O}}(\overrightarrow{\alpha})$.

c.q.f.d.

Faisons maintenant choix d'un repère orthonormé $(\mathrm{O},(\overrightarrow{\varepsilon_\ell})_{\ell\in L})$ de $\mathcal{M}$. On a dans ce repère :

$$\mathbf{B}_{\mathrm{O}} = \mathbf{V} + \mathbf{d}_{obs}\ \mathbf{d}'_{obs}$$

où $\mathbf{d}_{obs}$ et $\mathbf{V}$ sont respectivement le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{OG}}$ et la matrice de covariance du nuage M^I.

Corollaire 1.1. *Dans le repère $(\mathrm{O},(\overrightarrow{\varepsilon_\ell})_{\ell\in L})$, la matrice de covariance du nuage de typicalité C^J, notée $\mathbf{V}^{\mathrm{C}}$, est telle que :*

$$\mathbf{V}^{\mathrm{C}} = \frac{\mathbf{B}_{\mathrm{O}}}{n}$$

En effet, dans ce repère orthonormé, la matrice de l'endomorphisme Som_{O} est $\mathbf{B}_{\mathrm{O}}$ et celle de l'endomorphisme Som^{C} est $\mathbf{V}^{\mathrm{C}}$, d'où la relation en appliquant la propriété 1.2.

Soient A et B deux points quelconques de $\mathcal{M}$. Notons $|\ .\ |_{Som^{\mathrm{C}}}$ la norme de Mahalanobis associée au nuage de typicalité.

Corollaire 1.2. *On a la relation :*

$$|\overrightarrow{\mathrm{AB}}|_{Som^{\mathrm{C}}} = \sqrt{n} \times |\overrightarrow{\mathrm{AB}}|_{\mathrm{O}}$$

Démonstration. D'après la propriété 1.2, on a $(Som^{\mathrm{C}})^{-1} = n \times (Som_{\mathrm{O}})^{-1}$.
D'où $|\overrightarrow{\mathrm{AB}}|_{Som^{\mathrm{C}}} = \langle \overrightarrow{\mathrm{AB}} | (Som^{\mathrm{C}})^{-1}(\overrightarrow{\mathrm{AB}})\rangle^{1/2} = \sqrt{n} \times \langle \overrightarrow{\mathrm{AB}} | (Som_{\mathrm{O}})^{-1}(\overrightarrow{\mathrm{AB}})\rangle^{1/2} = \sqrt{n} \times |\overrightarrow{\mathrm{AB}}|_{Som_{\mathrm{O}}}$.
c.q.f.d.

Propriété 1.3. *La moyenne du carré de la statistique de test D est :*

$$\mathrm{Moy}(D^2) = \frac{L}{n}$$

Démonstration. La variance du nuage C^J, calculée en fonction de la distance de Mahalanobis associée à l'endomorphisme Som^{C} est par définition égale à :

$$\frac{1}{\mathrm{J}} \sum_{j \in J} |\overrightarrow{\mathrm{OC}^j}|^2_{Som^{\mathrm{C}}} = \frac{1}{\mathrm{J}} \sum_{j \in J} \langle (Som^{\mathrm{C}})^{-1}(\overrightarrow{\mathrm{OC}^j}) | \overrightarrow{\mathrm{OC}^j} \rangle$$

Or, d'après la propriété 4.2 (chapitre 1, p.18), cette variance est égale à L.
D'après la propriété 1.2, $Som^{\mathrm{C}} = \frac{Som_{\mathrm{O}}}{n}$ donc $(Som^{\mathrm{C}})^{-1} = (Som_{\mathrm{O}})^{-1} \times n$,
d'où $L = \frac{1}{\mathrm{J}} \sum\limits_{j \in J} n \times \langle (Som_{\mathrm{O}})^{-1}(\overrightarrow{\mathrm{OC}^j}) | \overrightarrow{\mathrm{OC}^j} \rangle = \frac{1}{\mathrm{J}} \sum\limits_{j \in J} n \times |\overrightarrow{\mathrm{OC}^j}|^2_{\mathrm{O}}$
et donc $\frac{1}{\mathrm{J}} \sum\limits_{j \in J} |\overrightarrow{\mathrm{OC}^j}|^2_{\mathrm{O}} = \frac{L}{n} = \mathrm{Moy}(D^2)$.
c.q.f.d.

Interprétation géométrique du seuil observé. Rappelons que le seuil observé du test exact (*cf.* p.25) est égal à la proportion :

$$p(D \geq d_{obs}), \text{ avec } d_{obs} = |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}}$$

ou, ce qui revient au même, à la proportion :

$$p(\sqrt{n}D \geq \sqrt{n} d_{obs})$$

Considérons maintenant la famille des ellipsoïdes d'inertie du nuage de typicalité C^J et en particulier celui passant par le point G. D'après les propriétés 1.1 (p.27) et 1.2 (p.28), il est de centre O et est défini par :

$$\kappa = \sqrt{n} \times |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}}$$

Par conséquent, le seuil observé du test exact s'interprète comme la proportion des points C^j situés sur ou à l'extérieur de cet ellipsoïde.

Remarque sur le choix de la statistique de test : Pour évaluer la distance entre les points du nuage de typicalité et le point de référence, nous utilisons la statistique de test $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}}$. Or, d'après le corollaire 1.2, on a :

$$|\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}} = \frac{1}{\sqrt{n}} \times |\overrightarrow{\mathrm{OC}^j}|_{Som^{\mathrm{C}}}$$

où $|\,.\,|_{Som^{\mathrm{C}}}$ est la norme de Mahalanobis attachée au nuage de typicalité. La norme de Mahalanobis $|\,.\,|_{Som^{\mathrm{C}}}$ est privilégiée lorsque le nuage de typicalité n'est pas sphérique. De plus, elle

induit des propriétés intéressantes concernant la caractérisation de la zone de compatibilité (*cf.* paragraphe 1.1.2, p.31) et elle ne dépend pas de l'unité de distance géométrique. Les statistiques de test $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}}$ et $D' : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|_{Som\mathrm{C}} = \sqrt{n} \times |\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}}$ conduisant à des tests équivalents, nous choisissons de nous affranchir du facteur $\sqrt{n}$.

Exemple Cible : Sur la figure 2.3 est représentée la distribution de la statistique de test D et sa valeur observée

$$d_{obs} = |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}} = \sqrt{\mathbf{d}'_{obs}\mathbf{B}_{\mathrm{O}}^{-1}\mathbf{d}_{obs}} = 0.6939$$

On a $p_{obs} = p(D \geq d_{obs}) = 86/1024 = .084$. Nous pouvons donner une interprétation géométrique de p_{obs} : parmi les $2^{10} = 1024$ points C^j du nuage de typicalité, ceux vérifiant $|\overrightarrow{\mathrm{OC}^j}|_{\mathrm{O}} \geq |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}}$ sont au nombre de 86, ils sont situés sur ou à l'extérieur de l'ellipse d'inertie du nuage de typicalité C^J passant par le point G (figure 2.4).

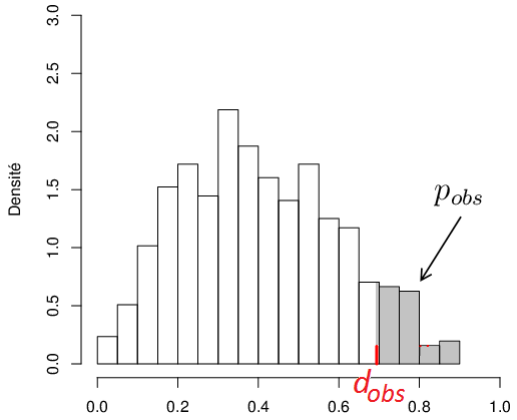


FIGURE 2.3 – *Exemple Cible.* Distribution de la statistique D ($d_{obs} = |\overrightarrow{\mathrm{OG}}|_{\mathrm{O}} = 0.6939$).

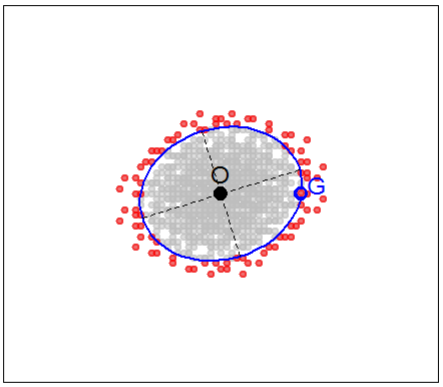


FIGURE 2.4 – *Exemple Cible.* Interprétation géométrique du seuil observé exact : proportion des points C^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage C^J passant par G (86 points rouges).

On a $p_{obs} > .05$, résultat non significatif au seuil .05. Pour la statistique D, nous ne pouvons pas dire que les données sont en faveur d'un point moyen vrai différent du point O. Par conséquent, nous ne pouvons pas dire que le point moyen G est atypique du point de référence O (au seuil .05) et que le point effectivement visé n'est pas le point O.

1.1.2 Zone de compatibilité

Considérons le point P de $\mathcal{M}$ tel que :

$$\forall \overrightarrow{u} \in \mathcal{L},\ \mathrm{P} = \mathrm{O} + \overrightarrow{u}$$

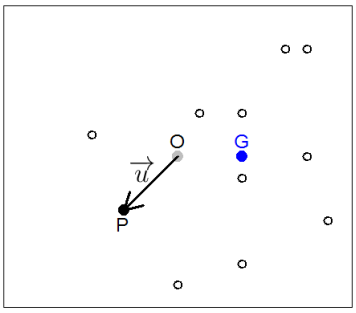


FIGURE 2.5 – *Exemple Cible.* Construction d'un point P.

En prenant le point P comme point de référence, considérons le nuage de typicalité $\mathrm{C}_{\mathrm{P}}^{J}$ défini comme au paragraphe 1.1.1, p.24 (avec P = O). Notons $|\overrightarrow{\mathrm{AB}}|_{\mathrm{P}}$ la M_{P}–distance entre deux points A et B de $\mathcal{M}$ (*cf.* chapitre 1, p.18) et D_{P} la statistique de test définie à partir de la M_{P}–distance (*cf.* p.25, avec P = O). Soit $|\overrightarrow{\mathrm{PG}}|_{\mathrm{P}}$ la valeur observée de la statistique de test D_{P}.

Définition 1.1 (Compatibilité/Incompatibilité)**.** *Les points* P *et* G *sont compatibles (resp. incompatibles) au seuil* α *si le seuil observé exact* $p_{obs} = p(D_{\mathrm{P}} \geq |\overrightarrow{\mathrm{PG}}|_{\mathrm{P}})$ *est strictement supérieur à* α *(resp. inférieur ou égal à* α*).*

Définition 1.2 (Zone de compatibilité)**.** *La zone de compatibilité au seuil* $(1-\alpha)$ *est l'ensemble des points* P *compatibles avec le point* G *au seuil* α*.*

Définition 1.3 (Point–limite d'incompatibilité)**.** *Un point* P *est point–limite d'incompatibilité s'il est juste incompatible avec le point* G *au seuil* α*, c'est–à–dire si* $p(D_{\mathrm{P}} \geq |\overrightarrow{\mathrm{PG}}|_{\mathrm{P}}) = \alpha$*.*

Caractérisation de la zone de compatibilité.

Théorème 1.1. *Si un point* P *est point–limite d'incompatibilité au seuil* α*, alors tout point de l'ellipsoïde d'inertie du nuage* M^{I} *passant par* P *est aussi point–limite d'incompatibilité au seuil* α*.*

Démonstration. Soit h_α un nombre positif tel que le point P soit point–limite d'incompatibilité au seuil α :

$$|\overrightarrow{\mathrm{PG}}|_{\mathrm{P}} = h_\alpha$$

Considérons l'ellipsoïde du nuage M^{I} passant par P et un point Q appartenant à cet ellipsoïde. D'après la propriété 5.1 (chapitre 1, p.20), on a :

$$|\overrightarrow{\mathrm{QG}}|_{\mathrm{Q}} = |\overrightarrow{\mathrm{PG}}|_{\mathrm{P}} = h_\alpha$$

Q est donc aussi un point–limite d'incompatibilité.
c.q.f.d.

Aux tracas du discret près, nous pouvons dire que les points–limites d'incompatibilité au seuil α appartiennent au même κ_α–ellipsoïde d'inertie du nuage M^I. La zone de compatibilité au seuil $1-\alpha$ est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de ce κ_α–ellipsoïde.

Dans la mesure où la compatibilité est évaluée grâce au test exact (par opposition au test approché que nous exposons dans la suite), cette zone peut également être appelée *zone de compatibilité exacte*.

Exemple Cible : Sur la figure 2.6 est représenté en rouge l'ensemble des points P juste incompatibles avec le point G aux seuils $\frac{50}{1024} = .049$ et $\frac{52}{1024} = .051$ [5] (il n'y a pas de point–limite correspondant au seuil .05, en effet 1024×0.05 n'est pas un nombre entier).

L'ellipse d'inertie du nuage M^I ajustée à cet ensemble de points (en bleu sur la figure 2.6) est telle que $\kappa = 1.094$ [6]. La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

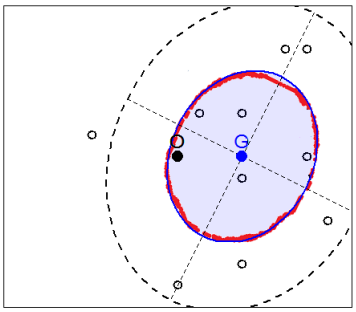


FIGURE 2.6 – *Exemple Cible.* Points–limites d'incompatibilité au seuil .05 (en rouge), zone de compatibilité au seuil .95 ($\kappa = 1.094$, en bleu) et ellipse de concentration du nuage M^I ($\kappa = 2$, en pointillés).

On voit bien sur la figure précédente que les points–limites d'incompatibilité appartiennent, aux tracas du discret près, à une ellipse d'inertie du nuage M^I. Le point O est situé à l'intérieur de la zone de compatibilité : les points O et G sont compatibles au seuil .05.

1.1.3 Cas particulier d'un nuage unidimensionnel

Dans le cas d'un nuage unidimensionnel ($L = 1$), on dispose d'un ensemble I de n individus formant un groupe d'observations auquel est associé le protocole numérique

5. *Note sur le calcul :* pour illustrer le théorème 1.1, nous avons ici voulu obtenir des points–limites d'incompatibilité au seuil .05. Pour ce faire, nous avons quadrillé l'espace et effectué le test en considérant chaque point du quadrillage comme point de référence (quadrillage horizontal : -1.5 → 13.5 par pas de 0.05, quadrillage vertical : -8 → 8 par pas de 0.05). Les points–limites obtenus ont ensuite été ajustés par une ellipse dont le κ est donné ici. Informatiquement, cette procédure est très coûteuse ; cependant, les résultats du théorème 1.1 permettent de simplifier le processus : nous procédons par approximations successives en balayant uniquement les axes principaux du nuage M^I, nous obtenons ainsi $2 \times L$ points–limites qui sont ensuite ajustés par une ellipse (*cf.* Mise en oeuvre informatique, p.196).

6. *Méthode d'ajustement :* pour un point–limite d'incompatibilité P donné, nous avons ici calculé la valeur k telle que $k = |\overrightarrow{\mathrm{PG}}|$. En effectuant ce procédé pour tous les points–limites obtenus, nous obtenons un ensemble de valeurs dont la moyenne donne le κ de l'ellipse ajustée. Nous aurions pu utiliser une autre technique d'ajustement (telle que la régression elliptique par exemple).

$x^I = (x^i)_{i\in I}$ de moyenne $\bar{x}$ et de variance v. Nous voulons étudier la typicalité de la moyenne $\bar{x}$ par rapport à une valeur de référence x_0.

Remarques :
- Le test géométrique décrit précédemment peut être effectué dans le cas d'un nuage unidimensionnel, cependant nous en donnons ici une version spécifique envisageable uniquement dans ce cas particulier.
- *Choix de la statistique de test :* afin de tenir compte du côté où se situe la moyenne du groupe d'observations par rapport à la valeur de référence (typicalité à droite ou à gauche), nous utilisons la statistique de test M (définie ci–après) qui mesure l'écart orienté à la valeur de référence, plutôt que la statistique D, utilisée précédemment, qui mesure la valeur absolue de cet écart.

Pour simplifier l'exposé, nous transformons les données de base de sorte que 0 soit la valeur de référence, nous construisons donc le protocole m^I de la façon suivante :

$$\forall i \in I,\ m^i = x^i - x_0$$

Par construction, le protocole m^I a pour moyenne $m_{obs} = \bar{x} - x_0$. Étudier la typicalité de $\bar{x}$ par rapport à x_0 revient donc à étudier la typicalité m_{obs} par rapport à 0.

Comme dans le cas multidimensionnel (*cf.* p.24), sous l'hypothèse que la vraie moyenne est la valeur de référence 0, la valeur m^i pourrait aussi bien être égale à $-m^i$. En effectuant les $\mathrm{J} = 2^n$ échanges possibles de $1, 2, \ldots n$ valeurs observées avec leurs valeurs opposées, nous obtenons un ensemble $\mathcal{J}$ de $\mathrm{J} = 2^n$ protocoles numériques $(m^{Ij})_{j\in \mathrm{J}}$ sur I (de cardinal n).

Pour étudier la typicalité de m_{obs} par rapport à 0, la statistique *Moyenne*, notée M et définie par :

$$\begin{array}{rccc} M: & \mathcal{J} & \to & \mathbb{R} \\ & m^{Ij} & \mapsto & \sum\limits_{i\in I} m^{ij}/n \end{array}$$

peut être prise comme statistique de test. Sa valeur observée est m_{obs}.

Seuil observé exact. Il s'impose d'effectuer ici un test *unilatéral* (*supérieur*, si $\bar{x} > x_0$ ou *inférieur*, si $\bar{x} < x_0$). Si $\bar{x} > x_0$ (ou $m_{obs} > 0$), on calcule le seuil observé exact supérieur p_{sup}, c'est–à–dire la proportion des valeurs de M pour lesquelles $M \geq m_{obs}$; si $\bar{x} < x_0$ (ou $m_{obs} < 0$), on calcule le seuil observé exact inférieur p_{inf}, c'est–à–dire la proportion des valeurs de M pour lesquelles $M \leq m_{obs}$. Le seuil observé exact unilatéral p_{unil} est par définition p_{sup} si $\bar{x} > x_0$ (ou $m_{obs} > 0$) ou p_{inf} si $\bar{x} < x_0$ (ou $m_{obs} < 0$).

La conclusion du test est énoncée, pour la statistique M, en termes d'atypicalité de $\bar{x}$ par rapport à x_0 (ou, ce qui est équivalent, de m_{obs} par rapport à 0), en tenant compte du signe de l'effet.

Si $\bar{x} > x_0$ (ou $m_{obs} > 0$) et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. La moyenne du groupe d'observations $\bar{x}$ est atypique (à droite) de la valeur de référence x_0 au seuil unilatéral $\alpha/2$. En d'autres termes, la moyenne du groupe d'observations est significativement supérieure à la valeur de référence au seuil unilatéral $\alpha/2$.

Si $\bar{x} < x_0$ (ou $m_{obs} < 0$) et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. La moyenne du groupe d'observations $\bar{x}$ est atypique (à gauche) de la valeur de référence x_0 au seuil unilatéral $\alpha/2$. En d'autres termes, la moyenne du groupe d'observations est significativement inférieure à la valeur de référence au seuil unilatéral $\alpha/2$.

Si $p_{unil} > \alpha/2$, le test est non significatif au seuil bilatéral α. Nous ne pouvons pas dire que la moyenne du groupe d'observations $\bar{x}$ est atypique de la valeur de référence x_0 au seuil bilatéral α.

Caractéristiques de la distribution de la statistique de test M.

Propriété 1.4. *La moyenne de la statistique Moyenne est :*

$$\mathrm{Moy}(M) = 0$$

Cette propriété est un cas particulier de la propriété 1.1 (p.27).

Propriété 1.5. *La variance de la statistique Moyenne est :*

$$\mathrm{Var}\, M = \frac{v + m_{obs}^2}{n}$$

Démonstration. En reprenant les notations de la démonstration de la propriété 1.1 (p.27), on a :

$$\begin{aligned}
\mathrm{Var}\, M &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} \sum_{i\in I} (\varepsilon_i^j \times m^i/n)^2 \quad \text{or, } \forall i \in I,\ \forall j \in J,\ (\varepsilon_i^j)^2 = 1 \\
&= \tfrac{1}{\mathrm{J}} \times \tfrac{1}{n} \sum_{j\in J} \tfrac{1}{n} \sum_{i\in I} (m^i)^2 \\
&= \tfrac{1}{\mathrm{J}} \times \tfrac{1}{n} \sum_{j\in J} (v + m_{obs}^2) \\
&= \tfrac{1}{\mathrm{J}} \times \tfrac{1}{n} \times \mathrm{J} \times (v + m_{obs}^2) \\
&= \tfrac{v+m_{obs}^2}{n}
\end{aligned}$$

c.q.f.d.

Intervalle de compatibilité. En adaptant la définition 1.1 (p.31) au cas unidimensionnel, nous pouvons dire que les valeurs m ($\in \mathbb{R}$) et m_{obs} sont compatibles au seuil unilatéral $\alpha/2$ si le seuil (unilatéral) observé associé est strictement supérieur à $\alpha/2$.

La limite inférieure (resp. supérieure) de l'intervalle de compatibilité au seuil $1-\alpha$ est alors la plus petite (resp. la plus grande) valeur m compatible avec m_{obs} au seuil unilatéral $\alpha/2$.

Remarque : Nous déduisons l'intervalle de compatibilité au seuil $1-\alpha$ correspondant aux données de base en ajoutant x_0 aux limites inférieures et supérieures de l'intervalle de compatibilité trouvé ci–avant.

1.2 Test approché

Supposons que le nuage de typicalité C^J soit ajusté par une distribution gaussienne multidimensionnelle à L dimensions (L dimension de $\mathcal{M}$), de centre O et avec la structure de covariance associée à Som_{O}/n, c'est–à–dire telle que la densité soit définie par :

$$\forall \overrightarrow{x} \in \mathcal{L},\ f(\overrightarrow{x}) = \frac{1}{(2\pi)^{L/2} \det(Som_{\mathrm{O}}/n)^{1/2}} \exp\left(-\frac{1}{2} \times n \times |\overrightarrow{x}|_{\mathrm{O}}^2\right)$$

Dans ce cas, la distribution de la statistique :

$$\mathrm{T}^2 = n \times D^2$$

est une distribution du χ^2 à L degrés de liberté (χ_L^2) (*cf.* définition de D p.25).

Seuil observé approché. Soit $t^2_{obs} = n \times d^2_{obs}$ (avec $d_{obs} = |\overrightarrow{OG}|_O$), la proportion $p(T^2 \geq t^2_{obs})$ est approximativement égale à $p(\chi^2_L \geq t^2_{obs}) = \widetilde{p}_{obs}$, seuil observé du test approché (seuil approché ou *p*–value approchée).

Nous concluons en termes d'atypicalité du point moyen du groupe d'observations par rapport au point de référence, pour la statistique T (*cf.* p.25).

Exemple Cible : Sur la figure 2.7 est représentée la distribution de la statistique $T^2 = n \times D^2$, à cette distribution est superposée celle du χ^2 à 2 degrés de liberté (en rouge).

On a $t^2_{obs} = n \times d^2_{obs} = n \times |\overrightarrow{OG}|^2_O = 10 \times 0.6939^2 = 4.815$, d'où le seuil observé approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq t^2_{obs}) = p(\chi^2_2 \geq 4.815) = .090$ (à comparer avec le seuil exact $p_{obs} = \frac{86}{1024} = .084$).

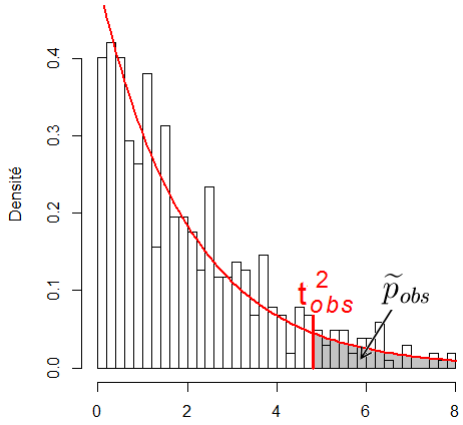


FIGURE 2.7 – *Exemple Cible.* Distributions de la statistique T^2 ($t^2_{obs} = n \times |\overrightarrow{OG}|^2_O = 4.815$) et du χ^2 à 2 degrés de liberté (en rouge).

Le test approché conduit à la même conclusion que le test exact.

1.2.1 Zone approchée de compatibilité

Rappelons que la zone de compatibilité au seuil $1-\alpha$ est définie comme étant l'ensemble des points P compatibles avec le point G au seuil α. Nous adoptons ici la même démarche que celle décrite au paragraphe 1.1.2 (p.31), la compatibilité entre les points P et G étant évidemment évaluée grâce au test approché.

Définition 1.4 (Zone–limite approchée d'incompatibilité)**.** *La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points–limites d'incompatibilité au seuil α, obtenus par le test approché.*

Soit $\chi^2_L[\alpha]$ la valeur critique de la distribution du χ^2 à L degrés de liberté au seuil supérieur α :

$$p(\chi^2_L \geq \chi^2_L[\alpha]) = \alpha$$

Théorème 1.2. *La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est le κ–ellipsoïde d'inertie du nuage M^I tel que :*

$$\kappa = \sqrt{\frac{\chi^2_L[\alpha]/n}{1 - \chi^2_L[\alpha]/n}}$$

Démonstration. La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points P vérifiant :
$n \times |\overrightarrow{\mathrm{PG}}|^2_{\mathrm{P}} = \chi^2_L[\alpha] \Leftrightarrow n \times \frac{|\overrightarrow{\mathrm{PG}}|^2}{1+|\overrightarrow{\mathrm{PG}}|^2} = \chi^2_L[\alpha]$ (d'après la propriété de réciprocité énoncée au chapitre 1, p.18).
Après calcul on obtient : $|\overrightarrow{\mathrm{PG}}|^2 = \frac{\chi^2_L[\alpha]/n}{1-\chi^2_L[\alpha]/n}$.
c.q.f.d.

Exemple Cible : Sur la figure ci–après est représentée la zone approchée de compatibilité au seuil .95, elle est définie par l'ensemble des points situés à l'intérieur de l'ellipse d'inertie du nuage M^I telle que $\kappa = \sqrt{\frac{\chi^2_2[.05]/n}{1-\chi^2_2[.05]/n}} = \sqrt{\frac{5.991/10}{1-5.991/10}} = 1.222$.

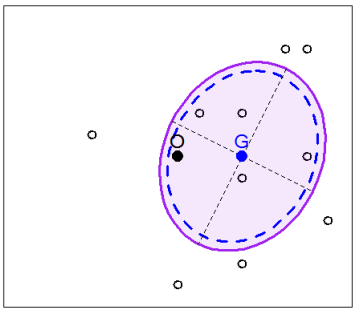


FIGURE 2.8 – *Exemple Cible.* Zone de compatibilité approchée au seuil .95 ($\kappa = 1.222$, en violet) et zone de compatibilité exacte au seuil .95 ($\kappa = 1.094$, en pointillés bleus).

Pour cet exemple, la zone de compatibilité approchée est plus large que la zone de compatibilité exacte.

1.2.2 Cas particulier d'un nuage unidimensionnel

Dans le cas où $L = 1$, le test approché consiste à ajuster la distribution de la statistique *Moyenne* M (*cf.* paragraphe 1.1.3, p.32) par une distribution gaussienne, de moyenne nulle et de variance $\mathrm{Var}\, M$ $(= (v + m^2_{obs})/n)$. Cependant, nous utilisons ici la statistique *Écart Réduit* :

$$Z = \frac{M}{\sqrt{\mathrm{Var}\, M}}$$

dont la distribution est une distribution gaussienne $\mathcal{N}(0,1)$; les statistiques M et Z conduisent à des tests équivalents.

Seuil observé approché. Soit $z_{obs} = \frac{m_{obs}}{\sqrt{(v+m^2_{obs})/n}}$ la valeur observée de la statistique *Écart Réduit*. Notons z la variable dont la distribution est une loi de Gauss centrée réduite : $z \sim \mathcal{N}(0,1)$.

Si $\bar{x} > x_0$ (ou $m_{obs} > 0$), on calcule le seuil observé approché supérieur $\widetilde{p}_{sup} = p(z \geq z_{obs})$; si $\bar{x} < x_0$ (ou $m_{obs} < 0$), on calcule le seuil observé approché inférieur $\widetilde{p}_{inf} = p(z \leq z_{obs})$. Le seuil observé approché unilatéral $\widetilde{p}_{unil}$ est par définition $\widetilde{p}_{sup}$ si $\bar{x} > x_0$ (ou $m_{obs} > 0$) ou $\widetilde{p}_{inf}$ si $\bar{x} < x_0$ (ou $m_{obs} < 0$).

Remarque : On a $\widetilde{p}_{sup} = \widetilde{p}_{inf}$ car la distribution gaussienne est symétrique.

Nous énonçons la conclusion du test, pour la statistique Z, en termes d'atypicalité de $\bar{x}$ par rapport à x_0 (ou, ce qui revient au même, de m_{obs} par rapport à 0), en tenant compte du signe de l'effet (*cf.* p.33).

Intervalle approché de compatibilité. Soit $z[\alpha]$ la valeur critique, au seuil bilatéral α, de la distribution gaussienne centrée réduite :

$$p(|z| > z[\alpha]) = \alpha$$

L'intervalle approché de compatibilité au seuil $1-\alpha$ est l'ensemble des valeurs $m \in \mathbb{R}$ compatibles avec m_{obs} au seuil bilatéral α, c'est–à–dire vérifiant :

$$\frac{|m - m_{obs}|}{\sqrt{\mathrm{Var}\, M}} < z[\alpha]$$

ou, ce qui revient au même, appartenant à l'intervalle :

$$]m_{obs} - z[\alpha]\sqrt{\mathrm{Var}\, M}\,;\, m_{obs} + z[\alpha]\sqrt{\mathrm{Var}\, D}[$$

Par construction, cet intervalle est centré sur m_{obs}.

Remarque : En ajoutant x_0 aux limites inférieures et supérieures de l'intervalle de compatibilité trouvé ci–dessus, nous obtenons l'intervalle de compatibilité approché au seuil $1-\alpha$, centré sur $\bar{x}$, construit à partir des données de base.

1.3 Cas de deux groupes appariés.

Rappelons que deux groupes d'observations sont dits « appariés » lorsqu'ils concernent un même groupe de n individus statistiques observés plusieurs fois (mesures répétées). Nous étudions dans la suite le cas où les individus sont observés deux fois : en $t1$ et en $t2$.

Le cas où l'on dispose de deux groupes d'observations appariés, dont nous voulons comparer les points moyens, est une application privilégiée du test géométrique. Nous montrons que ce dernier s'interprète aussi comme un test de permutation.

Pour chaque individu i, on a deux observations $\Big($par exemple une avant traitement $(t1)$ et une après traitement $(t2)\Big)$, c'est–à–dire deux points M^i_{t1} et M^i_{t2}. Les données de base sont donc constituées d'un ensemble de paires de points $(\mathrm{M}^i_{t1}, \mathrm{M}^i_{t2})_{i\in I}$. Nous nous proposons de comparer le point moyen du nuage M^I_{t1} au point moyen du nuage M^I_{t2}, en d'autres termes, nous nous demandons si l'écart entre ces deux points moyens est significativement différent de l'écart nul, ou non.

Nous résolvons ce problème grâce au test géométrique en construisant préalablement le nuage pertinent pour sa mise en oeuvre : à l'écart nul, nous faisons correspondre un point O quelconque de $\mathcal{M}$, pris comme point de référence. Au couple de points $(\mathrm{M}^i_{t1}, \mathrm{M}^i_{t2})$, nous associons le vecteur–écart $\overrightarrow{\mathrm{M}^i_{t1}\mathrm{M}^i_{t2}}$ et le point–écart D^i, défini par :

$$\overrightarrow{\mathrm{OD}}^i = \mathrm{O} + \overrightarrow{\mathrm{M}^i_{t1}\mathrm{M}^i_{t2}}$$

Nous comparons maintenant le point moyen du nuage D^I des n points–écarts, au point de référence O en utilisant le test géométrique.

Le test ainsi construit est équivalent au test de permutation comparant les points moyens de deux nuages appariés M^I_{t1} et M^I_{t2}. En effet, intuitivement, dire que pour l'individu i, il n'y a pas de différence entre $t1$ et $t2$, c'est dire que la/les valeur(s) observée(s) en $t1$ auraient aussi bien pu être observée(s) en $t2$: nous pourrions donc *permuter* les points M^i_{t1} et M^i_{t2}. Cette permutation revient à échanger le point–écart D^i avec son symétrique par rapport à O (Pesarin & Salmaso parlent de « symétrie induite par échangeabilité », *cf.* [64], 2010, p.15). La construction de l'ensemble des 2^n permutations possibles entre les valeurs observées en $t1$ et en $t2$ revient donc, en considérant le nuage des points–écarts D^I, à la construction de l'espace de typicalité inhérent au test géométrique (*cf.* p.24).

Le test que nous présentons ici peut être également vu comme un test d'homogénéité de deux groupes appariés vis–à–vis du point moyen. C'est pourquoi, il peut être considéré comme une généralisation multidimensionnelle du test d'homogénéité de Fisher–Pitman appliqué au cas où l'on dispose de deux groupes appariés (*cf.* Rouanet, Bernard & Le Roux, 1990, p.122 [72]).

2 Autres tests

Dans cette section, nous comparons le test géométrique de typicalité à d'autres tests : d'abord au test de Hotelling (test de signification traditionnel sous le modèle normal), puis à un test basé sur un rééchantillonnage bootstrap.

2.1 Test de Hotelling d'écart nul

Munissons l'espace $\mathcal{M}$ du repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$ et plaçons nous dans ce repère.

2.1.1 Modèle normal

Supposons que le nuage M^I soit un échantillon indépendamment et identiquement distribué d'une distribution gaussienne multidimensionnelle à L dimensions, centrée sur le vrai point moyen Γ (paramètre principal de l'inférence) et de structure de covariance $\mathbf{\Sigma}$ (paramètre secondaire de l'inférence). La matrice $\mathbf{\Sigma}$ est estimée par la matrice de covariance $\mathbf{V}$ du nuage M^I (estimateur du maximum de vraisemblance, somme des carrés divisés par n), ou par $\mathbf{S}$, matrice de covariance corrigée par le nombre q de degrés de liberté (estimateur sans biais, somme des carrés divisée par q avec $q = n - 1$ dans le cas élémentaire).

Remarques :

- Nous sortons ici du cadre combinatoire pour nous placer dans le cadre fréquentiste.
- Dans le cas de deux groupes appariés (*cf.* paragraphe 1.3, p.37), le modèle normal n'est pas posé sur les deux nuages M^I_{t1} et M^I_{t2} mais uniquement sur le nuage D^I des points–écarts.

Le modèle multinormal–Wishart met en jeu les distributions de Wishart, du χ^2 et du F de Snedecor (*cf.* par exemple Anderson, 1958 [3]).

Notons G la variable *Point Moyen* dont une réalisation est G, V la variable *Covariance* dont une réalisation est $\mathbf{V}$ et Γ le vrai point moyen (*cf.* par exemple Rouanet & Lecoutre, 1983 [75]), les propriétés d'échantillonnage sont les suivantes (Γ et $\mathbf{\Sigma}$ étant supposés fixés) :

- G et V sont statistiquement indépendants,

– $G \sim \mathcal{N}(\Gamma, \mathbf{\Sigma}/n)$: la distribution de G est une distribution gaussienne multidimensionnelle à L dimensions de point moyen Γ et de matrice de covariance $\mathbf{\Sigma}/n$,
– $n\,V \sim \mathcal{W}_L(q, \mathbf{\Sigma})$: la distribution de $n\,V$ est une distribution de Wishart à L dimensions et à q degrés de liberté ($\mathbf{\Sigma}$ est définie positive).

2.1.2 Seuil observé du test

Sous le modèle multinormal–Whishart, testons l'hypothèse nulle que le vrai point moyen est le point O ($\mathcal{H}_0 : \Gamma = \mathrm{O}$) à l'aide du test de Hotelling.

Sous l'hypothèse $\mathcal{H}_0$, le vecteur–écart $\overrightarrow{\mathrm{OG}}$ est supposé être une réalisation d'une variable *Vecteur–Écart*, notée $\overrightarrow{d}$. Soit $D^2_{\mathbf{V}}$ la statistique $\mathbf{V}$–norme de $\overrightarrow{d}$ au carré, c'est–à-dire la statistique définie par :

$$D^2_{\mathbf{V}} = |\overrightarrow{d}|^2$$

La valeur observée de la statistique $D^2_{\mathbf{V}}$ est $d^2_{\mathbf{V}\,obs} = |\overrightarrow{\mathrm{OG}}|^2$. On a la propriété suivante :

> Sous $\mathcal{H}_0$, la distribution de la statistique $T^2 = q\,D^2_{\mathbf{V}}$ est celle de $\frac{L}{k}\,F_{L,q-L+1}$ avec $k = (q - L + 1)/q$ (F usuel de Fisher–Snedecor centré avec L et $q - L + 1$ degrés de liberté).

Le seuil observé du test de Hotelling est la probabilité d'échantillonnage que la statistique T^2 dépasse sa valeur observée $t^2_{obs} = q \times d^2_{\mathbf{V}\,obs}$. Soit $F[\alpha]$ la valeur critique de la distribution $F_{L,q-L+1}$ au seuil supérieur α. Si $\frac{k}{L} t^2_{obs} = \frac{k}{L} q\, d^2_{\mathbf{V}\,obs} \geq F[\alpha]$, alors le résultat du test est significatif au seuil α.

Remarque : Pour n assez grand, k est proche de 1, et $L \times F_{L,n-L} \simeq F_{L,\infty} = \chi^2_L$, donc $T^2 \sim \chi^2_L$ (sous l'hypothèse $\mathcal{H}_0$).

Exemple Cible : On a $t^2_{obs} = q \times d^2_{\mathbf{V}\,obs} = q \times |\overrightarrow{\mathrm{OG}}|^2 = (n-1) \times \mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs} = 9 \times 0.9286 = 8.357$. On a $k = (q - L + 1)/q = 8/9$ et $p(F_{2,q-L+1} \geq \frac{k}{L} t^2_{obs}) = p(F_{2,8} \geq 3.714) = .072$, le test de Hotelling est donc non significatif au seuil $\alpha = .05$. Le test de Hotelling conduit à la même conclusion que le test géométrique pour lequel $p_{obs} = .084$.

2.1.3 Zone de confiance

La zone de confiance de Hotelling au seuil $1 - \alpha$ est définie par l'ensemble des points P de $\mathcal{M}$ tels que, sous l'hypothèse $\mathcal{H}_0 : \Gamma = \mathrm{P}$, le test de Hotelling soit non significatif au seuil α. Pour α fixé, l'ensemble des points P vérifiant cette propriété sont tels que :

$$\frac{k}{L}\, q|\overrightarrow{\mathrm{PG}}|^2 < F[\alpha]$$

Ils sont donc situés à l'intérieur de l'ellipsoïde d'inertie du nuage M^I tel que :

$$\kappa = \sqrt{\frac{L}{kq} F[\alpha]}$$

Remarque : Dans le cas élémentaire $\kappa^2 = \frac{L}{n-L} F[\alpha]$.

Exemple Cible : Sur la figure ci–après est représentée la zone de confiance de Hotelling au seuil .95, elle est définie par l'ensemble des points situés à l'intérieur de l'ellipse d'inertie du nuage M^I telle de $\kappa = \sqrt{\frac{L}{kq} \times F_{2,8}[.05]} = \sqrt{\frac{2}{8} \times 4.459} = 1.056$.

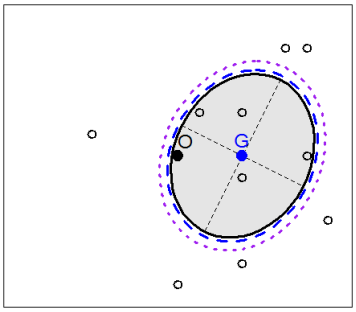


FIGURE 2.9 – *Exemple Cible.* Zone de confiance de Hotelling au seuil .95 ($\kappa = 1.056$, en gris), zone de compatibilité exacte au seuil .95 ($\kappa = 1.094$, en bleu, pointillés longs) et zone de compatibilité approchée au seuil .95 ($\kappa = 1.222$, en violet, pointillés courts).

Pour cet exemple, la zone de confiance de Hotelling est la moins large des trois zones, le test de Hotelling est donc celui qui renvoie le plus de résultats significatifs.

2.1.4 Cas particulier d'un nuage unidimensionnel : test de Student

Sous le modèle normal, on suppose que le protocole x^I (de moyenne $\bar{x}$ et de variance v) est un échantillon indépendamment et identiquement distribué d'une distribution gaussienne de moyenne μ et de variance σ^2.

Le test de Student permet de tester l'hypothèse nulle $\mathcal{H}_0 \ : \ \mu = x_0$. Si l'on prend comme statistique de test le rapport de Student :

$$t = \sqrt{n-1} \times \frac{M - \mu}{\sqrt{v}}$$

où M est la statistique *Moyenne*, alors sous $\mathcal{H}_0$, t est distribué selon une loi t_q (loi de Student à $q = n - 1$ degrés de liberté).

Seuil observé. Soit $t_{obs} = \sqrt{n-1} \times \frac{\bar{x} - x_0}{\sqrt{v}}$ la valeur observée de la statistique de Student. Le seuil observé unilatéral est alors défini comme étant la probabilité, sous $\mathcal{H}_0$, que la statistique de test t soit plus extrême, du côté des données, que la statistique observée t_{obs}.

Intervalle de confiance. Soit $t_q[\alpha]$ la valeur critique, au seuil bilatéral α, de la distribution de Student à $q = n - 1$ degrés de liberté :

$$p(|t_q| \geq t_q[\alpha]) = \alpha$$

L'intervalle de confiance de Student au seuil $1 - \alpha$ est l'ensemble des valeurs $m \in \mathbb{R}$, prises comme valeur de référence, pour lesquelles le test de Student est non significatif au seuil bilatéral α, c'est–à–dire vérifiant :

$$\sqrt{n-1} \times \frac{|\bar{x} - m|}{\sqrt{v}} < t_q[\alpha]$$

ou, ce qui revient au même, appartenant à l'intervalle :

$$]\bar{x} - t_q[\alpha](v/\sqrt{n-1}) \, ; \, \bar{x} + t_q[\alpha](v/\sqrt{n-1})[$$

Cet intervalle est centré sur $\bar{x}$.

2.2 Test du bootstrap

Principe du bootstrap. La technique du bootstrap a été introduite par Efron en 1979 ([27]). Elle s'applique lorsqu'on dispose d'un échantillon indépendamment et identiquement distribué $x^1, x^2, \ldots, x^n$ de taille n, issu d'une population inconnue dont on souhaite estimer un paramètr, par exemple la moyenne μ. Le bootstrap propose une estimation de ce paramètre inconnu basée uniquement sur les données observées, c'est–à–dire sur l'échantillon.

Classiquement, pour estimer un paramètre inconnu d'une population, on approche la distribution d'échantillonnage de la statistique d'intérêt, par exemple de la statistique *Moyenne* si l'on souhaite estimer μ (voir ci–après), par une distribution théorique, la plupart du temps basée sur le modèle normal. Les hypothèses sous–jacentes à ce procédé sont contraignantes et souvent invérifiables. Le bootstrap en propose une alternative en approchant la distribution d'échantillonnage de la statistique d'intérêt par une distribution « bootstrap », calculée uniquement à partir de l'échantillon (*cf.* Efron & Tibshirani, 1993 [28]). Le bootstrap peut donc être considéré comme une méthode de rééchantillonnage particulière, fondée sur la distribution empirique de l'échantillon initial.

La procédure est la suivante :

1. On effectue n tirages avec remise dans l'échantillon, chaque observation a alors la même probabilité $1/n$ d'être tirée $\rightarrow$ on obtient un échantillon *bootstrap* $x^{1*}, x^{2*}, \ldots, x^{n*}$ de taille n.
2. On calcule la statistique d'intérêt sur l'échantillon bootstrap. Par exemple, pour l'estimation de la moyenne μ, la statistique d'intérêt est la statistique *Moyenne* :
$$M : (x^{1*}, x^{2*}, \ldots, x^{n*}) \mapsto \frac{1}{n} \sum_{k=1}^{n} x^{k*}$$
3. On réitère les étapes 1. et 2. B fois [7] $\rightarrow$ on obtient la *distribution bootstrap* de la statistique d'intérêt. D'après la théorie sous–jacente à la procédure de rééchantillonnage bootstrap, la distribution bootstrap obtenue approche la distribution d'échantillonnage de cette même statistique [8].
4. A partir de la distribution bootstrap $\rightarrow$ on estime le paramètre inconnu et on construit un intervalle de confiance autour de cette estimation.

Remarque : Le bootstrap présenté ici est qualifié par Efron et Tibshirani de bootstrap *non paramétrique*, par opposition au bootstrap *paramétrique* qui met en jeu des distributions théoriques.

Le bootstrap peut également être utilisé dans le cadre des tests statistiques (*cf.* par exemple Efron & Tibshirani, 1993 [28] et Davison & Hinkley, 1988 [23]).

Test. Supposons maintenant que le nuage M^I, à valeur dans $\mathcal{M}$, de point moyen G, soit un échantillon indépendamment et identiquement distribué d'une population inconnue de point moyen Γ (paramètre de l'inférence). Soit O un point de $\mathcal{M}$, nous nous posons la question suivante : *L'hypothèse « le point moyen de la population est le point* O *» est-elle compatible avec les données, ou non ?*

7. Plus le nombre d'échantillons bootstrap est grand, meilleures sont les approximations, mais plus le temps de calcul est long.

8. La statistique d'intérêt doit être indépendante de l'ordre de tirage des x^{i*}.

Pour répondre à cette question, le test du bootstrap est construit de la façon suivante (*cf.* Van Aelst & Willems, 2013 [89]) :

1. On tire B échantillons bootstrap de taille n (tirages avec remise) à partir du nuage M^I. Soit $B = \{1, \ldots, b, \ldots \mathrm{B}\}$ l'ensemble indexant les B échantillons bootstrap. Notons $(\mathrm{M}^{*1b}, \mathrm{M}^{*2b}, \ldots, \mathrm{M}^{*nb})$, le nuage de n points constituant le b–ème échantillon bootstrap. L'ensemble des B échantillons bootstrap est noté $\mathcal{B}$.

   (a) On considère l'application C^* qui au nuage $(\mathrm{M}^{*1b}, \mathrm{M}^{*2b}, \ldots, \mathrm{M}^{*nb})$ associe son point moyen C^{*b} :

   $$\begin{array}{rccc} C^* : & \mathcal{B} & \to & \mathcal{M} \\ & (\mathrm{M}^{*1b}, \mathrm{M}^{*2b}, \ldots, \mathrm{M}^{*nb}) & \mapsto & \mathrm{C}^j = \sum_{k=1}^n \mathrm{M}^{*kb}/n \end{array}$$

   (b) On fait choix d'une statistique de test, ici la statistique D^*, définie de la façon suivante :

   $$\begin{array}{rccc} D^* : & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{C}^{*b} & \mapsto & |\overrightarrow{\mathrm{GC}^b}|_{*b} \end{array}$$

   où $|\ .\ |_{*b}$ est la norme de Mahalanobis attachée au nuage M^{*b}.

   Soit $d^*_{obs} = |\overrightarrow{\mathrm{GO}}|$, où $|\ .\ |$ est la norme de Mahalanobis attachée au nuage M^I.

   Remarque : En général, les tests faisant appel au bootstrap nécessitent une certaine prudence car on doit s'assurer que le rééchantillonnage se fait dans des conditions en accord avec l'hypothèse nulle (ici que le point moyen de la population est le point O). Dans le cas présenté ici, cette précaution n'est pas nécessaire. En effet, la statistique de test ne dépend pas du point O. Le test consistant à translater le nuage M^I de sorte que O soit son point moyen (on se place sous l'hypothèse nulle) et à effectuer le rééchantillonnage bootstrap à partir de ce nuage translaté (en utilisant la statistique de test D^*, avec G = O), est complètement équivalent au test du bootstrap présenté ici. La valeur observée de la statistique de test serait alors $|\overrightarrow{\mathrm{OG}}|$, égale à $d^*_{obs} = |\overrightarrow{\mathrm{GO}}|$ définie ci–avant.

2. On détermine la proportion des échantillons bootstrap, c'est–à–dire des nuages $(\mathrm{M}^{*1b}, \mathrm{M}^{*2b}, \ldots, \mathrm{M}^{*nb})$, pour lesquels la valeur de la statistique D^* est supérieure ou égale à d^*_{obs}. Cette proportion définit le seuil observé du test du bootstrap et est notée p^*_{obs}, avec $p^*_{obs} = p(D^* \geq d^*_{obs})$.

   Si $p^*_{obs} \leq \alpha$ (seuil α fixé), le résultat du test est significatif au seuil α. Pour la statistique D^*, on peut dire que les données sont en faveur d'un point moyen de la population différent du point O.

   Si $p^*_{obs} > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique D^*, on ne peut pas dire que les données sont en faveur d'un point moyen de la population différent du point O.

L'inconvénient principal du test du bootstrap présenté ici est le suivant : la norme utilisée pour le calcul de la statistique de test change pour chaque échantillon bootstrap. La matrice de covariance de chaque échantillon bootstrap doit donc être inversée. Or, il est impossible de s'assurer de l'inversibilité de cette matrice. Il est donc fort probable, surtout pour des

petites tailles de données, d'obtenir des cas de non inversibilité et de ne pas pouvoir calculer la statistique de test.

Exemple cible : Nous choisissons ici de tirer B = 1 000 échantillons bootstrap [9]. Parmi ces B = 1 000 échantillons bootstrap, il y en a 89 pour lesquels la valeur de la statistique D^* est supérieure ou égale à la valeur observée $d^*_{obs} = |\overrightarrow{\mathrm{GO}}| = \sqrt{\mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs}} = 0.9636$ (*cf.* p.26) (distribution de la statistique D^* représentée sur la figure 2.10). D'où le seuil observé du test du bootstrap : $p^*_{obs} = \frac{89}{1\,000} = 0.089$ [10], le test n'est pas significatif au seuil .05 [11]. Pour la statistique D^* et au seuil .05, nous ne pouvons pas dire que les données sont en faveur d'un point moyen de la population différent du point O.

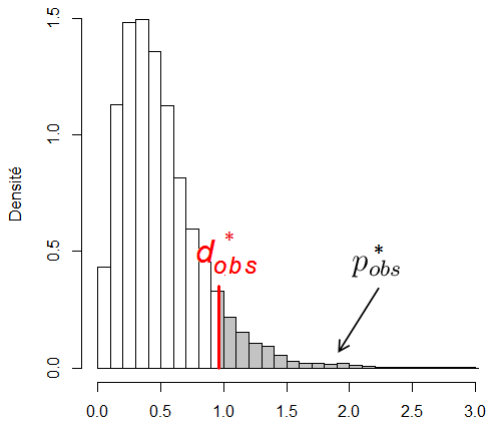


FIGURE 2.10 – *Exemple Cible.* Distribution de la statistique D^* ($d^*_{obs} = |\overrightarrow{\mathrm{GO}}| = 0.9636$).

Zone de confiance. A partir de la distribution bootstrap de la statistique D^*, nous pouvons donner un intervalle de confiance au seuil $1-\alpha$ de la valeur $|\overrightarrow{\mathrm{G\Gamma}}|$:

$$|\overrightarrow{\mathrm{G\Gamma}}| < \kappa_\alpha$$

où κ_α est le quantile d'ordre $1-\alpha$ de la distribution (*cf.* Van Aelst & Willems, 2013 [89]). On en déduit la zone de confiance bootstrap du point Γ : ensemble des points situés à l'intérieur de l'ellipse d'inertie du nuage M^I telle que $\kappa = \kappa_\alpha$.

Exemple cible : La zone de confiance bootstrap au seuil .95 est représentée sur la figure 2.12. Pour cet exemple, elle est définie par l'ellipse d'inertie du nuage M^I telle que $\kappa = 1.137$ [12] (quantile d'ordre .95 de la distribution de la statistique D^*, *cf.* figure 2.11).

9. Informatiquement, si la condition d'inversibilité pré–citée n'est pas satisfaite pour un ou plusieurs échantillon(s) bootstrap, nous ne le(s) retenons pas dans l'ensemble des B échantillons bootstrap.

10. Les seuils des tests exacts et de Hotelling sont respectivement .084 et .072.

11. Afin de s'assurer de la stabilité des résultats obtenus par bootstrap, nous avons effectué 10 tests indépendants dont les seuils observés sont donnés dans le tableau suivant :

p_{obs}	0.094	0.091	0.091	0.091	0.092	0.090	0.088	0.089	0.087	0.090

12. A comparer à $\kappa = 1.094$ pour la zone de compatibilité exacte et $\kappa = 1.056$ pour la zone de confiance de Hotelling (*cf.* figure 2.9, p.40).

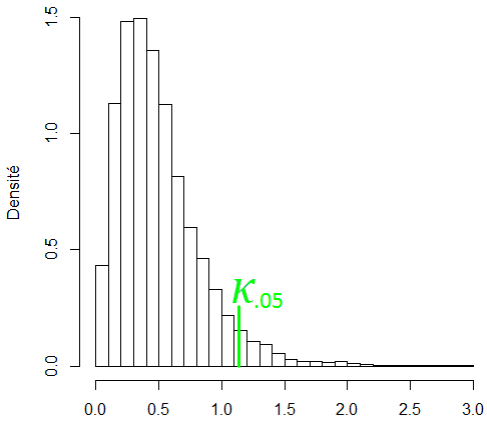


FIGURE 2.11 – *Exemple Cible.* Distribution de la statistique D^* et mise en évidence de $\kappa_{.05}$ (= 1.137).

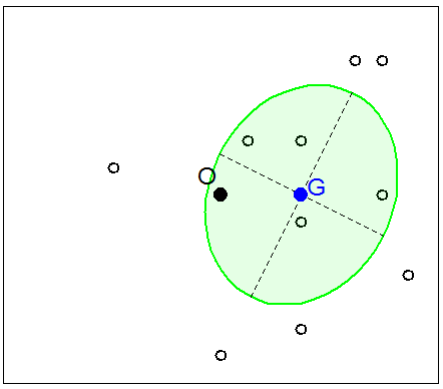


FIGURE 2.12 – *Exemple Cible.* Zone de confiance bootstrap ($\kappa = 1.137$).

3 Applications : données Parkinson (cas de deux groupes appariés)

Le test géométrique appliqué au cas de deux groupes appariés est illustré ici par l'exemple *Parkinson*.

Les données. Lors d'une recherche portant sur la marche de patients atteints de la maladie de Parkinson, Ferrandez & Blin (1991 [30]) comparent un groupe de 15 malades atteints de la maladie de Parkinson à un groupe de 45 personnes en bonne santé, de taille et d'âge comparables. Chaque observation est décrite par 6 variables numériques relatives à la marche : « vitesse », « longueur d'enjambée », « durée d'oscillation », « durée du cycle », « durée d'appui » et « durée de double appui » (*cf.* Le Roux & Rouanet, 2004 [50]). Les 15 malades sont observés deux fois : avant et après traitement.

Une ACP portant sur les données des 45 individus bien–portants et les six variables de marche est effectuée. Les 2×15 malades ont été mis en éléments supplémentaires. L'étude des variances des axes (*cf.* table 2.1 et figure 2.13), des corrélations multiples des variables initiales avec les variables principales et des qualités de représentation des individus montre que le nuage des bien–portants est presque entièrement contenu dans le premier plan principal (96.9% de la variance totale). Le premier axe s'interprète comme un axe de *performance* et le deuxième axe comme un axe de *style*.

	Valeur Propre	Pourcentage	Pourcentage cumulé
Axe 1	3.9927	66.55	66.55
Axe 2	1.8224	30.27	96.92
Axe 3	0.1711	2.85	99.77
Axe 4	0.0072	0.12	99.89
Axe 5	0.0059	0.10	99.99
Axe 6	0.01		100

TABLE 2.1 – Valeurs propres, pourcentages et pourcentages cumulés de la variance totale.

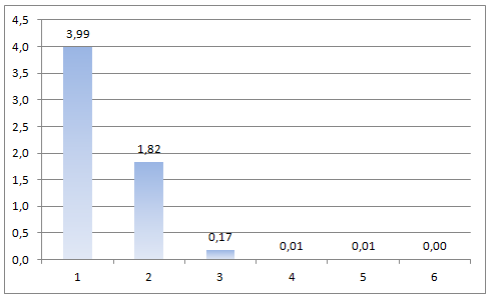


FIGURE 2.13 – Décroissance des valeurs propres.

L'ensemble I est ici composé des 15 individus atteints de la maladie de Parkinson. A cet ensemble I sont associés les nuages appariés de 2×15 points M_{t1}^I (malades observés avant traitement) et M_{t2}^I (malades observés après traitement) que nous étudions dans le premier plan principal ($L = 2$). Les coordonnées des 2×15 points $(\mathrm{M}_{t1}^i, \mathrm{M}_{t2}^i)_{i \in I}$ sur les deux premiers axes sont données dans la table 2.2, ces points sont tous très bien représentés dans le premier plan principal (*cf.* cosinus carrés, table 2.2).

	Avant traitement			Après traitement		
	$\cos^2$	x_1	x_2	$\cos^2$	x_1	x_2
1	.843	−2.212	0.928	.976	−1.286	1.854
2	.994	−1.655	−1.288	.938	−0.193	−1.562
3	.979	−0.948	−5.996	.983	−1.546	−4.651
4	.971	−0.986	−3.174	.876	+0.158	−0.912
5	.997	−1.073	−5.265	.991	−0.006	−2.335
6	.998	−1.217	−4.800	.976	+1.535	−4.765
7	.953	−1.747	−1.167	.914	−3.465	−1.430
8	.993	−3.544	−1.048	.977	+0.074	−1.737
9	.964	−5.217	−1.061	.997	−0.076	−1.814
10	.935	−5.098	−2.461	.978	−1.777	−1.188
11	.900	−3.877	−3.426	.976	−1.576	−3.270
12	.816	−1.152	−0.151	.907	−0.420	1.012
13	.967	−3.993	−1.038	.968	−5.907	1.014
14	.997	−0.759	−4.111	.997	+0.434	−4.094
15	.964	−5.820	−0.794	.929	−1.800	−1.069
Moyenne		−2.620	−2.323		−1.057	−1.663

TABLE 2.2 – Coordonnées des 15 malades (supplémentaires) dans le premier plan principal (x_1 et x_2) et qualités de représentation ($\cos^2$) dans ce plan.

La figure 2.14 suivante représente les 15 vecteurs–écarts $(\overrightarrow{\mathrm{M}_{t1}^i \mathrm{M}_{t2}^i})_{i \in I}$. En bas, à gauche de la même figure est représenté, à titre indicatif, le nuage des bien–portants, avec son point moyen O.

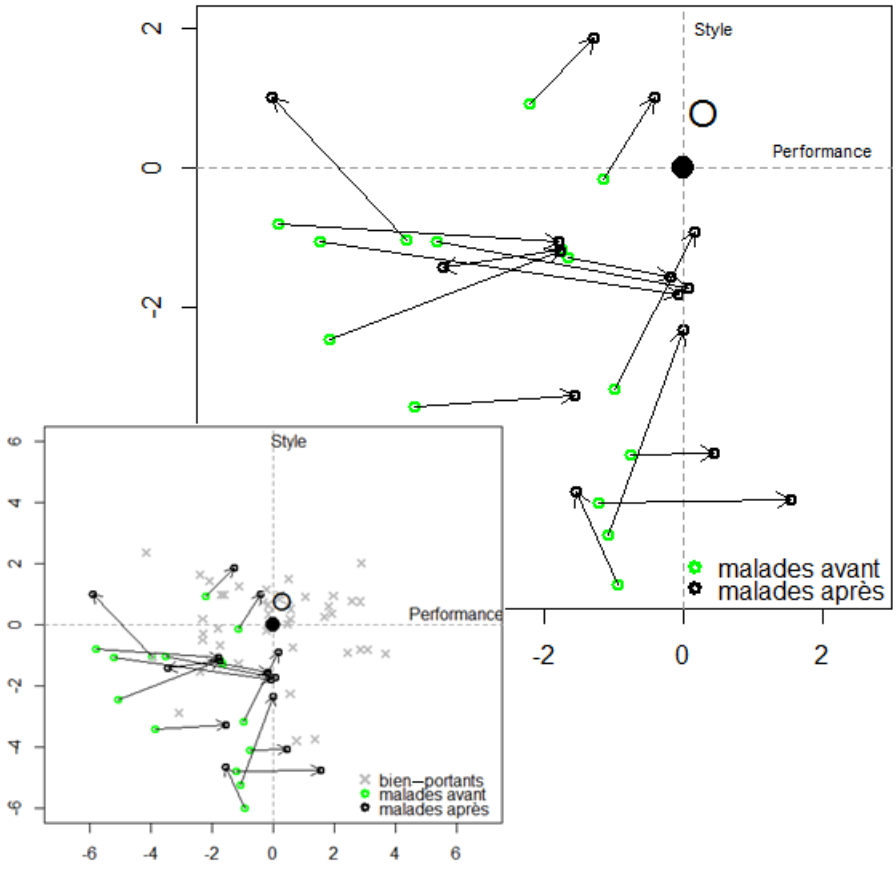


FIGURE 2.14 – *Exemple Parkinson.* Nuage des 15 vecteurs–écarts, avec en bas les 45 points des bien–portants.

Nous nous posons la question de recherche suivante : *Les données sont-elles en faveur d'un effet non nul du traitement ?*

Pour répondre à cette question, nous choisissons de comparer le point moyen du nuage M_{t1}^I (malades avant traitement) au point moyen du nuage M_{t2}^I (malades après traitement). Pour ce faire, nous mettons en oeuvre le test géométrique de typicalité, en construisant préalablement, le nuage des points–écarts pertinent pour le test (*cf.* paragraphe 1.3, p.37).

Nuage des points–écarts. Le nuage D^I des 15 *points–écarts* $(D^i)_{i\in I}$ est tel que :

$$\overrightarrow{OD}^i = O + \overrightarrow{M_{t1}^i M_{t2}^i}$$

Les étapes de sa construction sont rappelées sur la figure 2.15. Le point moyen du nuage D^I est noté D_{obs} (*cf.* figure 2.16). Les coordonnées des points–écarts dans le premier plan principal sont données dans la table 2.3.

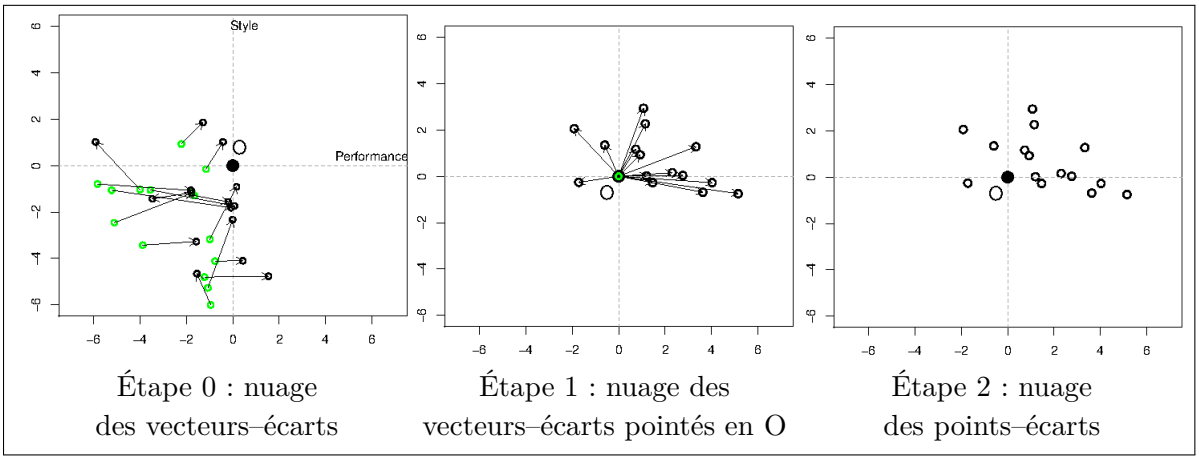


FIGURE 2.15 – *Exemple Parkinson.* Construction du nuage des points–écarts.

Remarque : Nous avons choisi de pointer les vecteurs–écarts en O, point moyen du nuage des bien–portants. En fait, dans la suite, le point O représente l'écart nul.

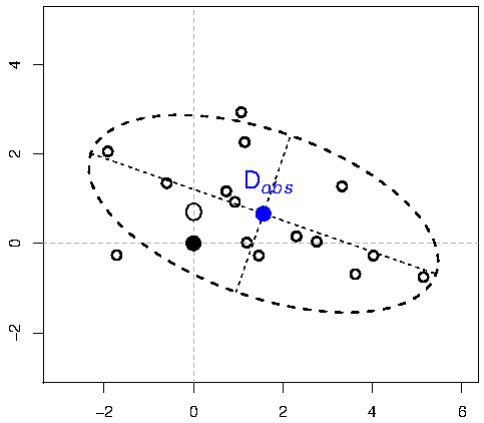


FIGURE 2.16 – *Exemple Parkinson.* Nuage des 15 points–écarts et son ellipse de concentration ($\kappa = 2$).

	x_1	x_2
1	0.926	0.926
2	1.462	−0.274
3	−0.598	1.345
4	1.144	2.262
5	1.067	2.930
6	2.752	0.035
7	−1.718	−0.263
8	3.618	−0.689
9	5.141	−0.753
10	3.321	1.273
11	2.301	0.156
12	0.732	1.163
13	−1.914	2.052
14	1.193	0.017
15	4.020	−0.275
Moyenne	1.563	0.660

TABLE 2.3 – *Exemple Parkinson.* Coordonnées des 15 points–écarts dans le premier plan principal.

Pour cet exemple, on a :

$\mathbf{d}_{obs} = \begin{pmatrix} 1.563 \\ 0.660 \end{pmatrix}$, vecteur–colonne des coordonnées de D_{obs} ;

$\mathbf{V} = \begin{pmatrix} 3.820 & -1.034 \\ -1.034 & 1.220 \end{pmatrix}$, matrice de covariance du nuage D^I, associée à Som ;

$\mathbf{B}_O = \mathbf{V} + \mathbf{d}_{obs}\ \mathbf{d}'_{obs} = \begin{pmatrix} 6.263 & -0.002 \\ -0.002 & 1.656 \end{pmatrix}$, matrice associée à Som_O.

Par conséquent, $|\overrightarrow{OD_{obs}}| = \sqrt{\mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs}} = 1.374$ et $|\overrightarrow{OD_{obs}}|_O = \sqrt{\mathbf{d}'_{obs}\mathbf{B}_O^{-1}\mathbf{d}_{obs}} = 0.809$.

Il s'agit maintenant de comparer le point moyen du nuage des points–écarts D_{obs} au point de référence O. Nous nous posons la question suivante : *le point moyen* D_{obs} *est-il atypique du point de référence* O *?*

Descriptivement, l'écart observé entre O et D_{obs} ($|\overrightarrow{OD_{obs}}| = 1.374$) est important. Tentons maintenant de prolonger cette conclusion descriptive en mettant en oeuvre le test géométrique de typicalité.

3.1 Test géométrique exact

En appliquant le test géométrique au nuage D^I des n points–écarts et en prenant le point O comme point de référence, nous obtenons les résultats suivants :

Seuil observé exact. Au j–ème nuage de l'espace de typicalité est associé son point moyen C^j, d'où le *nuage de typicalité* C^J des 2^{15} points moyens. La distribution de la statistique $D : C^j \mapsto |\overrightarrow{OC^j}|_O$ est donnée sur la figure 2.17, on a $d_{obs} = |\overrightarrow{OD_{obs}}|_O = 0.809$ et $p_{obs} = p(D \geq d_{obs}) = 36/2^{15} = 36/32\,768 = .001$. Le seuil observé exact s'interprète géométriquement comme la proportion des points du nuage de typicalité situés sur ou à l'extérieur de son ellipse d'inertie passant par D_{obs} (figure 2.18).

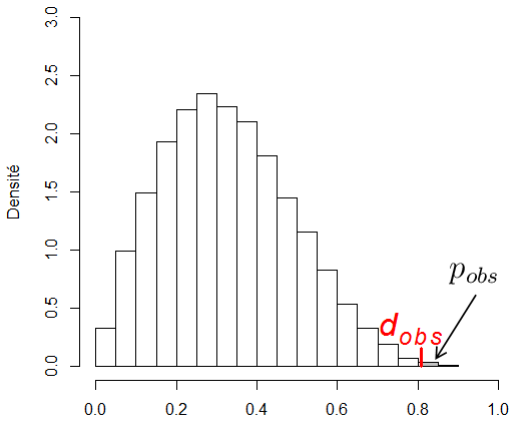


FIGURE 2.17 – *Exemple Parkinson.* Distribution de la statistique D ($d_{obs} = |\overrightarrow{OD_{obs}}|_O = 0.809$).

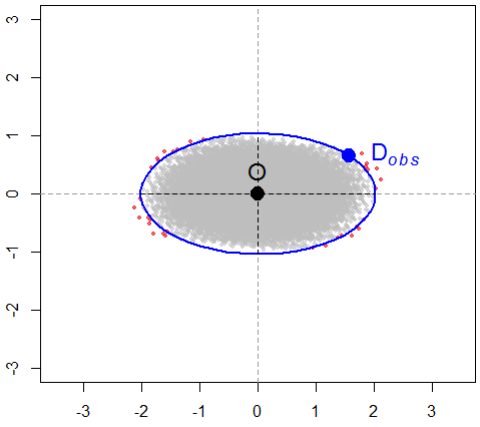


FIGURE 2.18 – *Exemple Parkinson.* Interprétation géométrique du seuil exact : proportion des points C^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage C^J passant par D_{obs} (36 points rouges).

Conclusion : On a $p_{obs} < .05$, résultat significatif au seuil .05. Pour la statistique D, nous pouvons dire que les données sont en faveur d'un point moyen vrai différent du point O. Par conséquent, nous pouvons dire que le point moyen D_{obs} est atypique du point de référence O au seuil .05. D'où la réponse à la question de recherche : les données sont en faveur d'un effet non nul du traitement.

Zone de compatibilité. En mettant en oeuvre la procédure de construction de la zone de compatibilité au seuil $1 - \alpha$ décrite au paragraphe 1.1.2 (p.31), nous avons recherché un ensemble de points P juste incompatibles avec le point D_{obs} aux seuils $\frac{1\,638}{32\,768} = 0.04999$ et

$\frac{1\,639}{32\,768} = 0.05002$ (il n'y a pas de point–limite correspondant au seuil .05, en effet $32\,768 \times 0.05$ n'est pas un nombre entier). Sur la figure 2.19 est représentée en bleu l'ellipse ajustée à cet ensemble de points, il s'agit de l'ellipse d'inertie du nuage des points–écarts telle que $\kappa = 0.767$ (la méthode d'ajustement est la même que celle décrite en bas de la page 32). La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

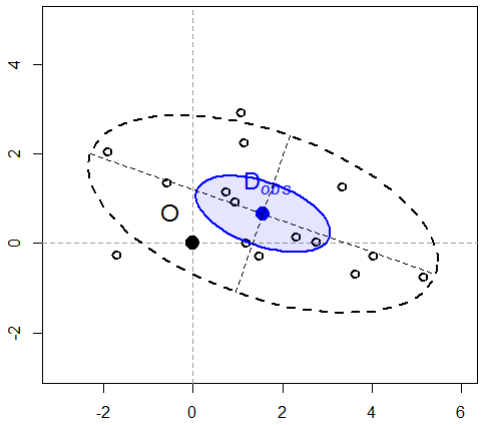


FIGURE 2.19 – *Exemple Parkinson.* Zone de compatibilité au seuil .95 ($\kappa = 0.767$, en bleu) et ellipse de concentration du nuage des points–écarts ($\kappa = 2$, en pointillés).

Le point O est situé à l'extérieur de la zone de compatibilité : les points O et D_{obs} sont incompatibles au seuil .05.

3.2 Test géométrique approché

Seuil observé approché. On a $t^2_{obs} = 15 \times d^2_{obs} = 15 \times 0.809^2 = 9.806$ (*cf.* p.35), d'où le seuil observé approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq 9.806) = .007$ (à comparer avec le seuil exact $p_{obs} = 36/2^{15} = .001$).

Le test approché conduit à la même conclusion que le test exact.

Zone de compatibilité approchée. On a $\chi^2_2[.05] = 5.991$, d'après le théorème 1.2 (p.35), la zone–limite approchée d'incompatibilité au seuil .95 est l'ellipse d'inertie du nuage des points–écarts telle que $\kappa = \sqrt{\frac{5.991/15}{1-5.991/15}} = 0.815$.

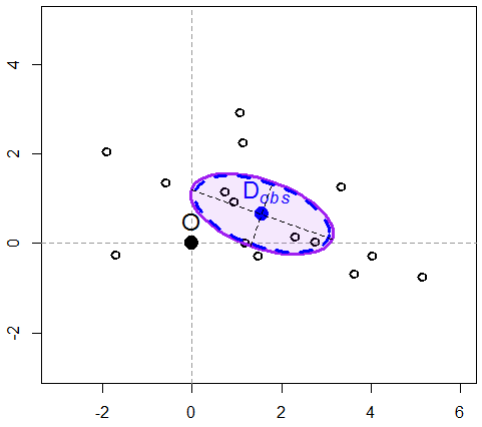


FIGURE 2.20 – *Exemple Parkinson.* Zone de compatibilité approchée au seuil .95 ($\kappa = 0.815$, en violet) et zone de compatibilité exacte au seuil .95 ($\kappa = 0.767$, en pointillés bleus).

Pour cet exemple, la zone de compatibilité approchée est plus large que la zone de compatibilité exacte. Le point O est situé à l'extérieur de la zone de compatibilité approchée : les points O et D_{obs} sont incompatibles au seuil .05.

A des fins de comparaisons des résultats, nous effectuons maintenant le test de Hotelling et le test du bootstrap.

3.3 Test de Hotelling

Seuil observé. On a $q = 15 - 1 = 14$, la valeur observée de la statistique de Hotelling est donc $t^2_{obs} = 14 \times |\overrightarrow{\mathrm{OD}_{obs}}|^2 = 14 \times \mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs} = 14 \times 1.888 = 26.432$.
On a $k = (q - L + 1)/q = 13/14$, on calcule donc $p(F_{L,q-L+1} \geq \frac{k}{L} \times t^2_{obs}) = p(F_{2,13} \geq \frac{13}{14\times 2} \times 26.432) = p(F_{2,13} \geq 12.272) = .001$.

Le seuil observé du test de Hotelling est égal, au millième près, au seuil observé du test géométrique exact. Les deux tests conduisent donc à la même conclusion.

Zone de confiance. La zone de confiance du vrai point moyen au seuil .95, appelée zone de confiance de Hotelling, est l'ensemble des points situés à l'intérieur de l'ellipse d'inertie du nuage des points–écarts telle que $\kappa = \sqrt{\frac{L}{n-L} \times F_{2,13}[.05]} = \sqrt{\frac{2}{13} \times 3.806} = 0.765$.

Cette zone de confiance est très proche de la zone de compatibilité du test géométrique exact ($\kappa = 0.767$) (*cf.* figure 2.19).

3.4 Test du bootstrap

Seuil observé. Nous choisissons ici de tirer $\mathrm{B} = 1\,000$ échantillons bootstrap. Parmi ces $\mathrm{B} = 1\,000$ échantillons bootstrap, il y en a 3 pour lesquels la valeur de la statistique D^* est supérieure ou égale à la valeur observée $d^*_{obs} = |\overrightarrow{\mathrm{D}_{obs}\mathrm{O}}| = \sqrt{\mathbf{d}'_{obs}\mathbf{V}^{-1}\mathbf{d}_{obs}} = 1.374$, d'où le seuil observé du test du bootstrap : $p^*_{obs} = \frac{3}{1\,000} = 0.003$.

Le test géométrique exact, le test de Hotelling et le test du bootstrap conduisent à la même conclusion.

Zone de confiance. A partir de la distribution bootstrap de la statistique D^*, nous pouvons construire la zone de confiance bootstrap du vrai point moyen au seuil .95 (*cf.* p.43). Pour cet exemple, la zone de confiance bootstrap au seuil .95 est définie par l'ellipse d'inertie du nuage M^I telle que $\kappa = 0.804$ (à comparer avec $\kappa = 0.767$ pour la zone de compatibilité exacte et avec $\kappa = 0.765$ pour la zone de confiance de Hotelling).

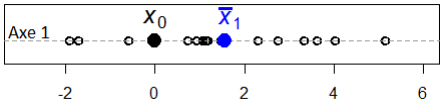
3.5 Tests unidimensionnels

En complément de l'étude de l'effet global du traitement (à partir des malades projetés sur le premier plan factoriel), nous pouvons nous poser les questions orientées suivantes :

1. Les données sont–elles en faveur d'une amélioration de la *performance* de la marche (interprétation de l'axe 1) après traitement ?
2. Les données sont–elles en faveur d'une amélioration du *style* de la marche (interprétation de l'axe 2) après traitement ?

Répondons à ces questions en mettant en oeuvre le test géométrique.

Pour traiter la première question, le protocole unidimensionnel x_1^I pertinent pour le test est le protocole constitué des coordonnées des points–écarts $(\mathrm{D}^i)_{i\in I}$ sur l'axe 1 (*cf.* colonne x_1 de la table 2.3, p.47). Il s'agit maintenant de comparer la moyenne observée $\bar{x_1} = 1.563$ à la valeur de référence $x_0 = 0$ (figure ci–après).

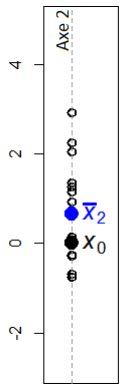


Les résultats des tests géométriques exact et approché et du test de Student sont donnés dans le tableau suivant (tests unilatéraux) :

$\bar{x}$	x_0	Test géométrique				Test de Student	
		Exact		Approché			
		p–value	I.C.	p–value	I.C.	p–value	I.C.
1.563	0	.005	]0.436; 2.689[	.008	]0.297 ; 2.830[	.005	]0.443 ; 2.683[

Conclusion : Les données sont en faveur d'une amélioration de la *performance* de la marche après traitement.

De la même façon, pour répondre à la seconde question, le protocole unidimensionnel x_2^I pertinent pour la mise en oeuvre du test géométrique est le protocole constitué des coordonnées des points–écarts $(\mathrm{D}^i)_{i\in I}$ sur l'axe 2 (*cf.* colonne x_2 de la table 2.3, p.47). Comparons maintenant la moyenne observée $\bar{x_2} = 0.660$ à la valeur de référence $x_0 = 0$ (figure ci–après).



Les résultats des tests géométriques exact et approché et du test de Student sont donnés dans le tableau suivant (tests unilatéraux) :

$\bar{x}$	x_0	Test géométrique				Test de Student	
		Exact		Approché			
		p–value	I.C.	p–value	I.C.	p–value	I.C.
0.660	0	.021	$]0.027; 1.304[$	.023	$]0.009\,;\,1.311[$	.021	$]0.027\,;\,1.293[$

Conclusion : Les données sont en faveur d'une amélioration du *style* de la marche après traitement.

Chapitre 3

Test ensembliste de typicalité

Dans ce chapitre, nous présentons le *test de typicalité ensembliste*, en bref *test de typicalité*. Il consiste à comparer un groupe d'observations à une population de référence, c'est–à–dire à étudier la *typicalité* du groupe d'observations par rapport à la population de référence. Pour ce test, le groupe peut, ou non, appartenir à la population de référence.

La logique du test de typicalité est la suivante : intuitivement, dire qu'un groupe d'observations de taille n est typique d'une population de référence de taille N, c'est dire qu'il est assimilable à la majorité des échantillons de taille n pouvant être obtenus à partir de cette population de référence. Ce principe conduit à construire l'*espace des échantillons* engendré par l'*ensemble des* $\binom{N}{n}$ *échantillons possibles*. Il s'agit ensuite de situer le groupe d'observations par rapport à l'espace des échantillons, selon la statistique de test choisie. Par exemple, en unidimensionnel, nous prenons la statistique *Moyenne* comme statistique de test. Nous considérons que le groupe d'observations est atypique de la population de référence si sa moyenne est extrême par rapport à la distribution de la statistique *Moyenne*.

Dans ce chapitre, nous exposons le test ensembliste de typicalité dans le cas multidimensionnel (le cas unidimensionnel est traité comme un cas particulier) ; une solution approchée, c'est–à–dire une approximation de la distribution de la statistique de test, est aussi fournie. Nous comparons également le test ensembliste au test plus traditionnel basé sur le modèle normal et l'appliquons, sans perte de généralité, à des nuages bi–dimensionnels.

1 Typicalité d'un groupe d'observations par rapport à une population de référence

Dans cette section, nous nous proposons d'étudier la typicalité d'un groupe d'observations par rapport à une population de référence.

Situation de base. Considérons un ensemble $I = \{1, \ldots, i, \ldots, N\}$ de N individus, formant une *population de référence* et un ensemble $I' = \{1, \ldots, i', \ldots, n\}$ de n individus (avec $n \leq N$), constituant un *groupe d'observations* (*cf.* Le Roux, 1998 [48] et Rouanet & Le Roux, 1990 [72], 2004 [50]). A l'ensemble I est associé le nuage M^I de N points (individus), appelé par la suite *nuage de référence*. Le nuage M^I est à valeurs dans un espace affine euclidien $\mathcal{U}$ de dimension K, son point moyen est noté O et sa structure de covariance est définie par l'endomorphisme de covariance *Som* (défini au chapitre 1, p.16). De la même façon, à l'ensemble I' est associé le nuage $\mathrm{M}^{I'}$ de n points, à valeurs dans le même espace $\mathcal{U}$ et dont

le point moyen est noté G.

Dans la suite, nous considérons que les points du nuage M^I *sont à valeurs dans* $\mathcal{M}$*, son support affine. Le sous–espace vectoriel directeur de* $\mathcal{M}$*, de dimension* L*, est noté* $\mathcal{L}$ *(cf. chapitre 1, p.12).*

1.1 Test exact

1.1.1 Principe du test

Nous nous posons la question suivante : *Le groupe d'observations est-il atypique de la population de référence ?*

Remarque : Le groupe d'observations peut, ou non, appartenir à la population de référence.

Pour répondre à cette question, nous construisons le test ensembliste de typicalité de la façon suivante :

1. Un échantillon de taille n est défini comme un sous–ensemble à n éléments de la population de référence I, de taille N. Considérons l'ensemble des $\mathrm{J} = \binom{N}{n}$ échantillons possibles issus de la population de référence. Soit $J = \{1, \ldots, j, \ldots, \mathrm{J}\}$ l'ensemble indexant les J échantillons [1]. Notons $I<j>$ l'ensemble des individus appartenant au j–ème échantillon et $\mathrm{M}^{I<j>}$ le sous–nuage qui lui est associé. L'ensemble des $\mathrm{J} = \binom{N}{n}$ échantillons possibles est appelé *ensemble des échantillons*, l'ensemble des $\mathrm{J} = \binom{N}{n}$ sous–nuages possibles est noté $\mathcal{J}$ et appelé *espace des échantillons*.
2. (a) Considérons maintenant l'application C qui au sous–nuage $\mathrm{M}^{I<j>}$ associe son point moyen C^j :

$$\begin{array}{rccl} C: & \mathcal{J} & \to & \mathcal{M} \\ & \mathrm{M}^{I<j>} & \mapsto & \mathrm{C}^j = \sum\limits_{i \in I<j>} \mathrm{M}^i/n \end{array}$$

   Le nuage C^J des J points moyens, muni de la pondération élémentaire, est appelé *nuage d'échantillonnage*.

   (b) Faisons ensuite choix d'une statistique de test (*cf. Remarque sur le choix de la statistique de test p.*59), ici la statistique notée D définie de la façon suivante : au point moyen C^j est associée sa M–distance [2] au point moyen O de la population de référence :

$$\begin{array}{rccl} D: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{C}^j & \mapsto & |\overrightarrow{\mathrm{OC}^j}| \end{array}$$

   Notons d_{obs} la valeur observée de cette statistique : $d_{obs} = |\overrightarrow{\mathrm{OG}}|$, où G est le point moyen du groupe d'observations, de taille n, dont nous voulons étudier la typicalité.

1. La plupart du temps dans ce chapitre, l'ensemble et son cardinal sont identifiés par la même lettre, l'ensemble en italique et son cardinal en lettre droite.
2. Distance de Mahalanobis attachée au nuage M^I (*cf.* chapitre 1, p.18).

3. Déterminons enfin la proportion des nuages $M^{I<j>}$ pour lesquels la valeur de la statistique D est supérieure ou égale à la valeur observée d_{obs}. Cette proportion définit le seuil observé du test exact (seuil exact ou p–value exacte) et est notée p_{obs} :

$$p_{obs} = p(D \geq d_{obs})$$

Nous pouvons désormais énoncer la conclusion du test en termes d'atypicalité du groupe d'observations par rapport à la population de référence.

Si $p_{obs} \leq \alpha$ (seuil $\alpha < 1/2$ fixé [3]), le résultat du test est significatif au seuil α. Pour la statistique D, nous pouvons dire que le groupe d'observations est atypique de la population de référence au seuil α.

Si $p_{obs} > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique D, nous ne pouvons pas dire que le groupe d'observations est atypique de la population de référence au seuil α.

Remarque : A partir de la distribution de la statistique de test D, nous pouvons définir la valeur critique d_α au seuil α telle que :

$$p(D \geq d_\alpha) = \alpha$$

Le résultat du test s'énonce alors comme ceci :

- si $d_{obs} \geq d_\alpha$, alors le résultat du test est significatif au seuil α,
- si $d_{obs} < d_\alpha$, alors le résultat du test n'est pas significatif au seuil α.

Exemple « Typicalité » : Afin d'illustrer les notions inhérentes au test de typicalité ensembliste, nous utilisons l'exemple suivant comme fil conducteur.

Considérons l'ensemble I de $N = 15$ individus : $I = \{i1, \ldots, i15\}$, formant une population de référence, et le groupe des $n = 3$ observations : $\{i13, i14, i15\}$ [4]. A l'ensemble I est associé le nuage de référence M^I : nuage plan ($L = 2$) de 15 points, de point moyen O. De même, au groupe $\{i13, i14, i15\}$ est associé le nuage des 3 points $\{M^{13}, M^{14}, M^{15}\}$, de point moyen G (*cf.* figure 3.1).

Nous nous demandons si le groupe d'observations est atypique de la population de référence.

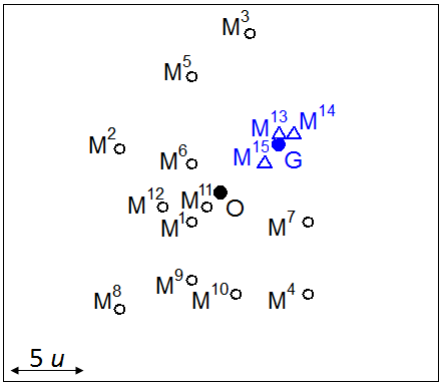
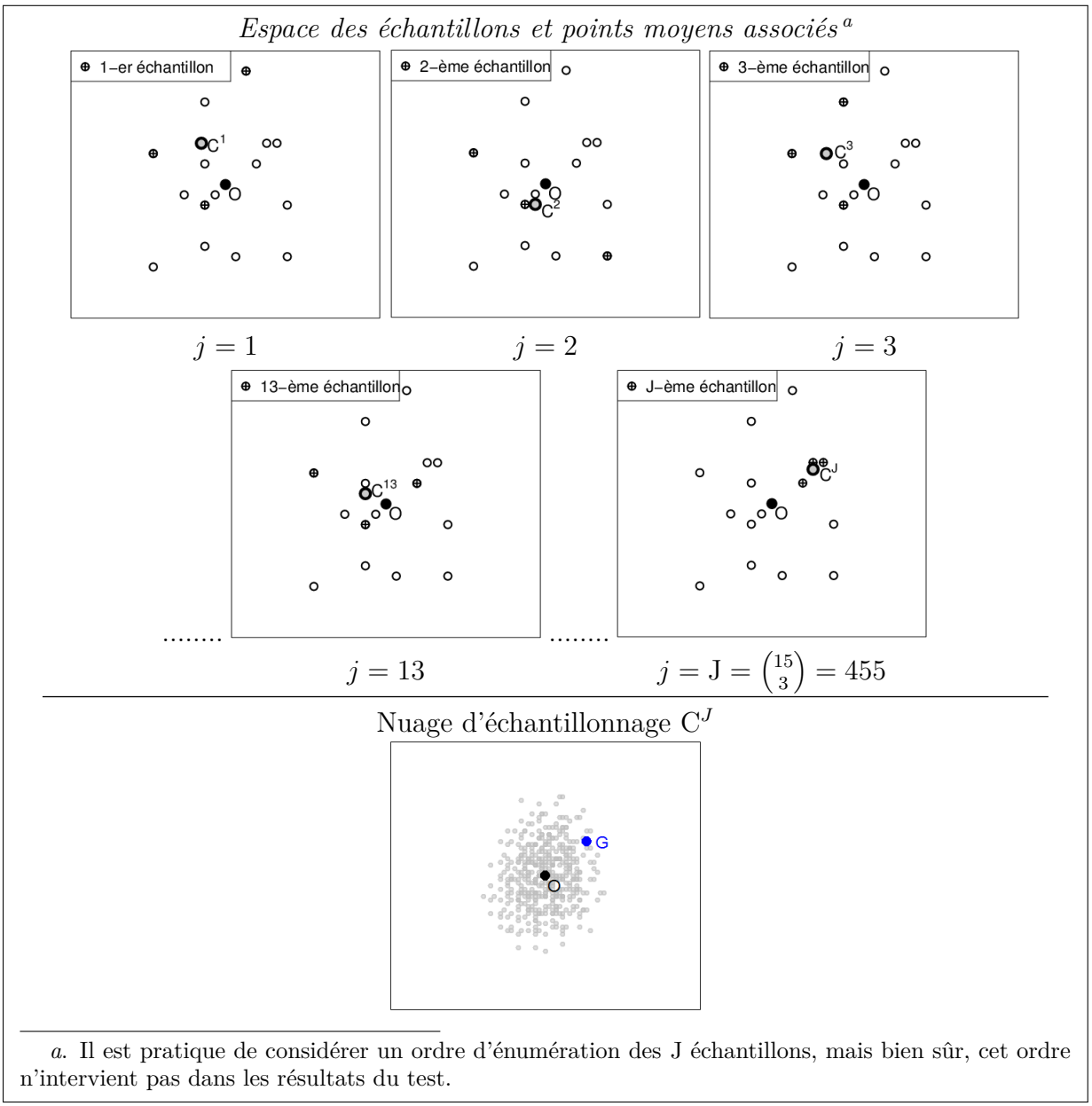


FIGURE 3.1 – *Exemple « Typicalité ».* Nuage de référence M^I et son point moyen O ; nuage des 3 points $\{M^{13}, M^{14}, M^{15}\}$ (triangles bleus) et son point moyen G. Pour les calculs numériques, nous prenons u comme unité de longueur.

3. On utilise souvent les seuils conventionnels $\alpha = .05$ et $\alpha = .01$.
4. Dans cet exemple, le groupe d'observations appartient à la population de référence.

Pour cet exemple, la construction de l'espace des échantillons et des points moyens associés est résumée sur la figure 3.2 ci–après. L'ensemble de ces points moyens constitue le nuage d'échantillonnage.



a. Il est pratique de considérer un ordre d'énumération des J échantillons, mais bien sûr, cet ordre n'intervient pas dans les résultats du test.

FIGURE 3.2 – *Exemple « Typicalité ».* Construction de l'espace des échantillons, des points moyens associés et du nuage d'échantillonnage.

Caractéristiques du nuage d'échantillonnage et de la distribution de la statistique de test.

Propriété 1.1. *Le point moyen du nuage d'échantillonnage* C^J *est le point moyen du nuage de référence* M^I*.*

Démonstration.
Posons $\varepsilon_i^j = 1$ si l'individu i appartient au j–ème échantillon et $\varepsilon_i^j = 0$, sinon.

Considérons maintenant les $\mathrm{J} = \binom{N}{n}$ points $(\mathrm{C}^j)_{j\in J}$ définis par :

$$\overrightarrow{\mathrm{OC}^j} = \sum_{i\in I} \varepsilon_i^j \overrightarrow{\mathrm{OM}^i}/n$$

Le point moyen du nuage C^J, noté $\overline{\mathrm{C}}$, est tel que :

$$\begin{aligned}\overrightarrow{\mathrm{OC}} &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} \overrightarrow{\mathrm{OC}^j} \\ &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} (\sum_{i\in I} \tfrac{1}{n} \varepsilon_i^j \overrightarrow{\mathrm{OM}^i}) \\ &= \tfrac{1}{\mathrm{J}} \sum_{i\in I} \tfrac{1}{n} (\sum_{j\in J} \varepsilon_i^j \overrightarrow{\mathrm{OM}^i})\end{aligned}$$

or pour i fixé, $\sum_{j\in J} \varepsilon_i^j = \binom{N-1}{n-1}$ (nombre de fois que l'on prend $n-1$ éléments autres que i parmi les $N-1$ éléments restants), d'où

$$\begin{aligned}\overrightarrow{\mathrm{OC}} &= \tfrac{1}{\mathrm{J}} \sum_{i\in I} \tfrac{1}{n} \tbinom{N-1}{n-1} \overrightarrow{\mathrm{OM}^i} \\ &= \tfrac{1}{\mathrm{J}} \tbinom{N-1}{n-1} \tfrac{1}{n} \sum_{i\in I} \overrightarrow{\mathrm{OM}^i}\end{aligned}$$

O étant le point moyen du nuage M^I, on a $\frac{1}{n} \sum_{i\in I} \overrightarrow{\mathrm{OM}^i} = \overrightarrow{0}$ (*cf.* caractérisation barycentrique du point moyen, p.13) et donc :

$$\overrightarrow{\mathrm{OC}} = \overrightarrow{0}.$$

c.q.f.d.

Lemme 1.1. *On a la relation :*

$$\sum_{i\in I} \sum_{\substack{i'\in I \\ i'\neq i}} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^{i'}} = -N Som(\overrightarrow{\alpha})$$

Démonstration. En raison de la caractérisation barycentrique du point moyen (p.13), on a $\sum_{i'\in I} \overrightarrow{\mathrm{OM}^{i'}} = \overrightarrow{0}$ et donc : $\sum_{i\in I} \sum_{i'\in I} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^{i'}} = \overrightarrow{0}$.

Or $\sum_{i\in I} \sum_{i'\in I} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^{i'}} = \sum_{i\in I} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^i} + \sum_{i\in I} \sum_{\substack{i'\in I \\ i'\neq i}} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^{i'}}$

et $\sum_{i\in I} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^i} = N Som(\overrightarrow{\alpha})$. D'où $\sum_{i\in I} \sum_{\substack{i'\in I \\ i'\neq i}} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OM}^{i'}} = -N Som(\overrightarrow{\alpha})$.

c.q.f.d.

Soit Som^{C} l'endomorphisme dont les vecteurs propres déterminent les directions principales du nuage d'échantillonnage C^J.

Propriété 1.2. *L'endomorphisme Som^{C} est proportionnel à l'endomorphisme de covariance Som associé au nuage de référence M^I :*

$$Som^{\mathrm{C}} = \frac{1}{n} \times \frac{N-n}{N-1} \times Som$$

Démonstration. D'après le théorème 3.1 (p.17), les directions principales du nuage C^J sont engendrées par les vecteurs propres de l'endomorphisme Som^{C} défini par :

$$\forall \overrightarrow{\alpha} \in \mathcal{L},\ Som^{\mathrm{C}}(\overrightarrow{\alpha}) = \frac{1}{\mathrm{J}} \sum_{j\in J} \langle \overrightarrow{\mathrm{OC}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OC}^j}$$

En reprenant les notations de la démonstration de la propriété 1.1 (p.56), on a :

$$\begin{aligned} Som^{\mathrm{C}}(\overrightarrow{\alpha}) &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} \langle \sum_{i\in I} \tfrac{1}{n}(\varepsilon_i^j \overrightarrow{\mathrm{OM}^i}) | \overrightarrow{\alpha}\rangle \sum_{i'\in I} \tfrac{1}{n}(\varepsilon_{i'}^j \overrightarrow{\mathrm{OM}^{i'}}) \\ &= \tfrac{1}{\mathrm{J}} \tfrac{1}{n^2} \sum_{i\in I} \sum_{i'\in I} \Big(\sum_{j\in J} \varepsilon_i^j \varepsilon_{i'}^j \Big) \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{OM}^{i'}} \end{aligned}$$

Pour $i = i'$, $\sum_{j\in J} (\varepsilon_i^j)^2 = \binom{N-1}{n-1}$; pour $i \neq i'$, $\sum_{j\in J} \varepsilon_i^j \varepsilon_{i'}^j = \binom{N-2}{n-2}$ (nombre de fois que l'on prend $n-2$ éléments autres que i et i' parmi les $N-2$ éléments restants).

Par conséquent $Som^{\mathrm{C}}(\overrightarrow{\alpha}) = \tfrac{1}{\mathrm{J}} \tfrac{1}{n^2} \Big(\binom{N-1}{n-1} \sum_{i\in I} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{OM}^i} + \binom{N-2}{n-2} \sum_{i\in I} \sum_{\substack{i'\in I \\ i'\neq i}} \langle \overrightarrow{\mathrm{OM}^i} | \overrightarrow{\alpha}\rangle \overrightarrow{\mathrm{OM}^{i'}} \Big)$.

D'après le lemme 1.1, on a :

$$\begin{aligned} Som^{\mathrm{C}}(\overrightarrow{\alpha}) &= \tfrac{1}{\mathrm{J}} \tfrac{1}{n^2} \Big(\tbinom{N-1}{n-1} N Som(\overrightarrow{\alpha}) - \tbinom{N-2}{n-2} N Som(\overrightarrow{\alpha}) \Big) \\ &= \tfrac{1}{n^2} \Big(\tfrac{\binom{N-1}{n-1} N}{\binom{N}{n}} - \tfrac{\binom{N-2}{n-2} N}{\binom{N}{n}} \Big) Som(\overrightarrow{\alpha}) \quad \text{(en remplaçant J par } \tbinom{N}{n}) \\ &= \tfrac{N}{n^2} \Big(\tfrac{n}{N} - \tfrac{n(n-1)}{N(N-1)} \Big) Som(\overrightarrow{\alpha}) \\ &= \tfrac{1}{n} (1 - \tfrac{n-1}{N-1}) Som(\overrightarrow{\alpha}) \\ &= \tfrac{1}{n} \times \tfrac{N-n}{N-1} \times Som(\overrightarrow{\alpha}). \end{aligned}$$

c.q.f.d.

Faisons maintenant choix d'un repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell\in L})$ de $\mathcal{M}$. Soit $\mathbf{V}$ la matrice de covariance du nuage de référence M^I dans ce repère.

Corollaire 1.1. *Dans le repère $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell\in L})$, la matrice de covariance du nuage d'échantillonnage C^J, notée $\mathbf{V}^{\mathrm{C}}$, est telle que :*

$$\mathbf{V}^{\mathrm{C}} = \frac{N-n}{N-1} \times \frac{\mathbf{V}}{n}$$

En effet, dans ce repère orthonormé, la matrice de l'endomorphisme *Som* est $\mathbf{V}$ et celle de l'endomorphisme Som^{C} est $\mathbf{V}^{\mathrm{C}}$, d'où la relation en appliquant la propriété 1.2.

Notons $|\ .\ |_{Som^{\mathrm{C}}}$ la norme de Mahalanobis attachée au nuage d'échantillonnage C^J et rappelons que $|\ .\ |$ désigne la norme de Mahalanobis attachée au nuage de référence M^I.

Corollaire 1.2. *On a la relation :*

$$|\ .\ |_{Som^{\mathrm{C}}} = \sqrt{n \times \frac{N-1}{N-n}} \times |\ .\ |$$

Tout ellipsoïde d'inertie du nuage de référence est donc ellipsoïde d'inertie du nuage d'échantillonnage.

La figure suivante, construite à partir de l'exemple précédent, illustre ce corollaire :

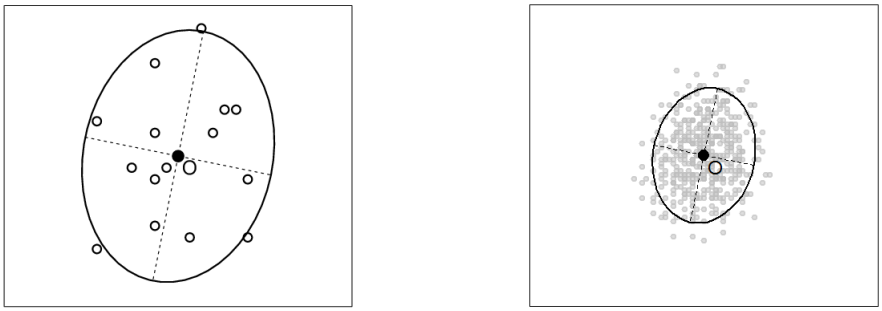


FIGURE 3.3 – *Exemple « Typicalité ».* Nuage de référence M^I (gauche) et nuage d'échantillonnage C^J (droite) avec leurs ellipses de concentration ($\kappa = 2$).

Propriété 1.3. *La moyenne du carré de la statistique de test D est :*

$$\mathrm{Moy}(D^2) = \frac{L}{n} \times \frac{N-n}{N-1}$$

Démonstration. D'après la propriété 4.2 (chapitre 1, p.18), la variance du nuage C^J, calculée en fonction de la distance de Mahalanobis associée à l'endomorphisme Som^{C} est égale à L.
On a donc $\frac{1}{\mathrm{J}} \sum_{j \in J} \langle (Som^{\mathrm{C}})^{-1}(\overrightarrow{\mathrm{OC}^j}) | \overrightarrow{\mathrm{OC}^j} \rangle = L$.
Or, d'après la propriété 1.2, $Som^{\mathrm{C}} = \frac{1}{n} \times \frac{N-n}{N-1} \times Som$, donc $(Som^{\mathrm{C}})^{-1} = n \times \frac{N-1}{N-n} \times (Som)^{-1}$,
d'où $L = \frac{1}{\mathrm{J}} \sum_{j \in J} n \times \frac{N-1}{N-n} \times \langle (Som)^{-1}(\overrightarrow{\mathrm{OC}^j}) | \overrightarrow{\mathrm{OC}^j} \rangle = \frac{1}{\mathrm{J}} \sum_{j \in J} n \times \frac{N-1}{N-n} \times |\overrightarrow{\mathrm{OC}^j}|^2$
et donc $\frac{1}{\mathrm{J}} \sum_{j \in J} |\overrightarrow{\mathrm{OC}^j}|^2 = \frac{L}{n} \times \frac{N-n}{N-1} = \mathrm{Moy}(D^2)$.
c.q.f.d.

Interprétation géométrique du seuil observé. Rappelons que le seuil observé du test exact (*cf.* p.55) est égal à la proportion des points C^j tels que $|\overrightarrow{\mathrm{OC}^j}| \geq |\overrightarrow{\mathrm{OG}}|$. Il s'interprète donc géométriquement comme la proportion des points C^j situés sur ou à l'extérieur de l'ellipsoïde d'inertie du nuage de référence M^I passant par le point G, c'est–à–dire défini par :

$$\kappa = |\overrightarrow{\mathrm{OG}}|$$

D'après le corollaire 1.2, cet ellipsoïde appartient également à la famille des ellipsoïdes d'inertie du nuage d'échantillonnage C^J.

Remarque sur le choix de la statistique de test : Pour évaluer la distance entre les points du nuage d'échantillonnage et le point O (point moyen de la population de référence), nous utilisons la statistique de test $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|$. Or, d'après le corollaire 1.2, on a :

$$|\overrightarrow{\mathrm{OC}^j}| = \sqrt{\frac{1}{n} \times \frac{N-n}{N-1}} \times |\overrightarrow{\mathrm{OC}^j}|_{Som^{\mathrm{C}}}$$

où $|\,.\,|_{Som^{\mathrm{C}}}$ est la norme de Mahalanobis attachée au nuage d'échantillonnage. La norme de Mahalanobis $|\,.\,|_{Som^{\mathrm{C}}}$ est privilégiée lorsque le nuage d'échantillonnage n'est pas sphérique, de plus, elle induit des propriétés intéressantes concernant la caractérisation de la zone de compatibilité (*cf.* paragraphe 1.1.2, p.61) et elle ne dépend pas de l'unité de distance géométrique. Les statistiques de test $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|$ et $D' : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|_{Som^{\mathrm{C}}} = \sqrt{n \times \frac{N-1}{N-n}} \times$

$|\overrightarrow{\mathrm{OC}^j}|$ conduisant à des tests équivalents, nous choisissons de nous affranchir du facteur $\sqrt{n \times \frac{N-1}{N-n}}$.

Exemple « Typicalité »[5] *:* On a :

$$|\overrightarrow{\mathrm{OG}}| = \sqrt{\begin{pmatrix}4 & 10/3\end{pmatrix} \times \begin{pmatrix}17.2 & 2.6\\ 2.6 & 29.5\end{pmatrix}^{-1} \times \begin{pmatrix}4\\ 10/3\end{pmatrix}} = 1.089$$

L'écart descriptif est important (≥ 0.8), nous sommes donc fondés à effectuer le test. Sur la figure 3.4 est représentée la distribution de la statistique de test D et sa valeur observée $d_{obs} = |\overrightarrow{\mathrm{OG}}|$. On a $p_{obs} = p(D \geq d_{obs}) = 61/455 = .134$. Nous pouvons donner une interprétation géométrique de p_{obs} : parmi les $\binom{15}{3} = 455$ points C^j du nuage d'échantillonnage, ceux vérifiant $|\overrightarrow{\mathrm{OC}^j}| \geq |\overrightarrow{\mathrm{OG}}|$ sont au nombre de 61, ils sont situés sur ou à l'extérieur de l'ellipse d'inertie du nuage de référence M^I passant par le point G (figure 3.5).

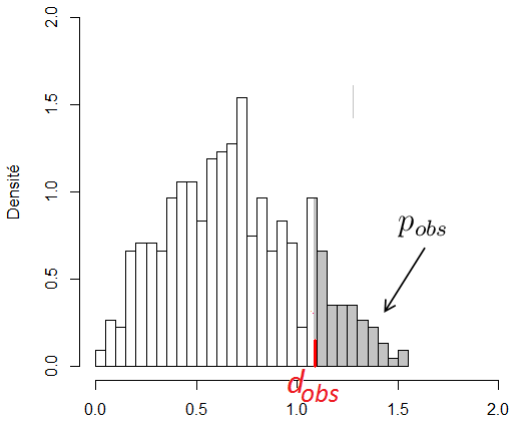


FIGURE 3.4 – *Exemple « Typicalité ».* Distribution de la statistique D ($d_{obs} = |\overrightarrow{\mathrm{OG}}| = 1.089$).

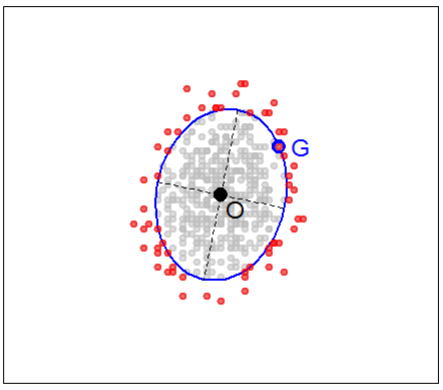


FIGURE 3.5 – *Exemple « Typicalité ».* Interprétation géométrique du seuil observé exact : proportion des points C^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par G (61 points rouges).

On a $p_{obs} > .05$, résultat non significatif au seuil .05. Pour la statistique D, nous ne pouvons pas dire que le groupe d'observations est atypique de la population de référence au seuil .05.

5. Tous les calculs ont été effectués en choisissant deux axes perpendiculaires, l'un horizontal (x_1), l'autre vertical (x_2), gradués selon l'unité de distance u (*cf.* figure 3.1, p.55) et dont l'origine est O, point moyen de la population de référence. Ci–après sont reportés le tableau de nombres associés aux 15 points, la matrice de covariance du nuage M^I et les coordonnées du vecteur $\overrightarrow{\mathrm{OG}}$:

	M^1	M^2	M^3	M^4	M^5	M^6	M^7	M^8	M^9	M^{10}	M^{11}	M^{12}	M^{13}	M^{14}	M^{15}	Moy
x_1	-2	-7	2	6	-2	-2	6	-7	-2	1	-1	-4	4	5	3	0
x_2	-2	3	11	-7	8	2	-2	-8	-6	-7	-1	-1	4	4	2	0

$$\mathbf{V} = \begin{pmatrix}17.2 & 2.6\\ 2.6 & 29.5\end{pmatrix}, \quad \overrightarrow{\mathrm{OG}} = \begin{pmatrix}4\\ 10/3\end{pmatrix}.$$

1.1.2 Zone de compatibilité

Considérons le point P de $\mathcal{M}$ tel que :

$$\forall \overrightarrow{u} \in \mathcal{L},\ \mathrm{P} = \mathrm{O} + \overrightarrow{u}$$

Considérons également le nuage P^I translaté du nuage M^I de telle sorte que :

$$\forall\ i \in I,\ \mathrm{P}^i = \mathrm{M}^i + \overrightarrow{\mathrm{OP}}$$

Par construction, le nuage P^I admet le point P pour point moyen et a la même structure de covariance que le nuage M^I. La norme de Mahalanobis attachée au nuage P^I est donc égale à la norme de Mahalanobis attachée au nuage M^I, notée $|\ .\ |$.

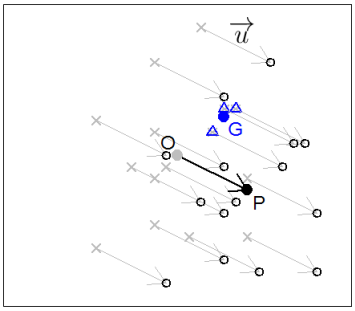


FIGURE 3.6 – *Exemple « Typicalité ».* Nuage P^I de point moyen P déduit du nuage M^I par translation de vecteur $\overrightarrow{\mathrm{OP}}$.

En prenant le nuage P^I comme nuage de référence, considérons le nuage d'échantillonnage C_P^J défini comme au paragraphe 1.1.1 (p.54), avec P = O.

Le nuage C_P^J admet P pour point moyen (propriété 1.1, p.56, avec O = P) et a la même structure de covariance que le nuage P^I (propriété 1.2, p.56) ; le nuage C_P^J a donc également la même structure de covariance que le nuage M^I et donc que le nuage d'échantillonnage C^J. Notons $D_\mathrm{P} : \mathrm{C}_\mathrm{P}^j \mapsto |\overrightarrow{\mathrm{PC}_\mathrm{P}^j}|$, la statistique de test écrite en fonction du point P. Soit $|\overrightarrow{\mathrm{PG}}|$ la valeur observée de la statistique de test D_P.

Définition 1.1 (Compatibilité/Incompatibilité)**.** *Les points* P *et* G *sont compatibles (resp. incompatibles) au seuil* α *si le seuil observé exact* $p_{obs} = p(D_\mathrm{P} \geq |\overrightarrow{\mathrm{PG}}|)$ *est strictement supérieur à* α *(resp. inférieur ou égal à* α*).*

Remarque : On a par construction : $\overrightarrow{\mathrm{PC}_\mathrm{P}^j} = \overrightarrow{\mathrm{OC}^j}$. La proportion des points C_P^j tels que $|\overrightarrow{\mathrm{PC}_\mathrm{P}^j}| \geq |\overrightarrow{\mathrm{PG}}|$ est donc égale à la proportion des points C^j tels que $|\overrightarrow{\mathrm{OC}^j}| \geq |\overrightarrow{\mathrm{PG}}|$.

Définition 1.2 (Zone de compatibilité)**.** *La zone de compatibilité au seuil* $(1-\alpha)$ *est l'ensemble des points* P *compatibles avec le point* G *au seuil* α*.*

Définition 1.3 (Point–limite d'incompatibilité)**.** *Un point* P *est point–limite d'incompatibilité s'il est juste incompatible avec le point* G *au seuil* α*, c'est–à–dire si* $p(D_\mathrm{P} \geq |\overrightarrow{\mathrm{PG}}|) = \alpha$*.*

Caractérisation de la zone de compatibilité.

Théorème 1.1. *Soit* P *un point–limite d'incompatibilité au seuil* α. *Si* $|\overrightarrow{PG}| = \kappa_\alpha$, *alors tout point de l'ellipsoïde défini par l'ensemble des points* Q *tels que* $|\overrightarrow{QG}| = \kappa_\alpha$ *est aussi point–limite d'incompatibilité au seuil* α.

Démonstration. Soit κ_α un nombre positif tel que le point P soit point–limite d'incompatibilité au seuil α : $|\overrightarrow{PG}| = \kappa_\alpha$ et soit l'ellipsoïde de centre G défini par l'ensemble des points Q tels que $|\overrightarrow{QG}| = \kappa_\alpha$. Soit Q′ un point appartenant à cet ellipsoïde : $|\overrightarrow{Q'G}| = \kappa_\alpha$.
Etudions la compatibilité entre les points G et Q′ : la proportion des points $C^j_{Q'}$ tels que $|\overrightarrow{Q'C^j_{Q'}}| \geq |\overrightarrow{Q'G}|$ est égale à la proportion des points C^j_P tels que $|\overrightarrow{PC^j_P}| \geq \kappa_\alpha$ (d'après la remarque ci–avant). Comme P est point–limite d'incompatibilité, cette proportion est égale à α.
Q′ est donc aussi un point–limite d'incompatibilité.
c.q.f.d.

Nous pouvons dire que les points–limites d'incompatibilité au seuil α appartiennent, aux tracas du discret près, au même κ_α–ellipsoïde de centre G, translaté d'un ellipsoïde d'inertie du nuage de référence M^I. La zone de compatibilité au seuil $1-\alpha$ est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de ce κ_α–ellipsoïde.

Dans la mesure où la compatibilité est évaluée grâce au test exact (par opposition au test approché que nous exposons dans la suite), cette zone peut également être appelée *zone de compatibilité exacte*.

Exemple « Typicalité » : Sur la figure 3.7 est représenté en rouge l'ensemble des points P juste incompatibles avec le point G aux seuils $\frac{22}{455} = .048$ et $\frac{23}{455} = .051$ [6] (il n'y a pas de point–limite correspondant au seuil .05, en effet 455×0.05 n'est pas un nombre entier).

L'ellipse ajustée à cet ensemble de points (en bleu sur la figure 3.7) est l'ellipse de centre G, translatée de l'ellipse d'inertie du nuage M^I telle que $\kappa = 1.266$ [7]. La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

6. *Note sur le calcul :* pour illustrer le théorème 1.1, nous avons ici voulu obtenir des points–limites d'incompatibilité au seuil .05. Pour ce faire, nous avons quadrillé l'espace et effectué le test en considérant chaque point du quadrillage comme point moyen de la population de référence (quadrillage horizontal : -1.5 → 13.5 par pas de 0.05, quadrillage vertical : -8 → 8 par pas de 0.05). Les points–limites obtenus ont ensuite été ajustés par une ellipse dont le κ est donné ici. Informatiquement, cette procédure est très coûteuse ; cependant, les résultats du théorème 1.1 permettent de simplifier le processus : nous procédons par approximations successives en balayant uniquement les axes parallèles aux axes principaux du nuage M^I qui se coupent en G. Nous obtenons ainsi $2 \times L$ points–limites qui sont ensuite ajustés par une ellipse.

7. *Méthode d'ajustement :* pour un point–limite d'incompatibilité P donné, nous avons ici calculé la valeur k telle que $k = |\overrightarrow{PG}|$. En effectuant ce procédé pour tous les points–limites obtenus, nous obtenons un ensemble de valeurs dont la moyenne donne le κ de l'ellipse ajustée. Nous aurions pu utiliser une autre technique d'ajustement.

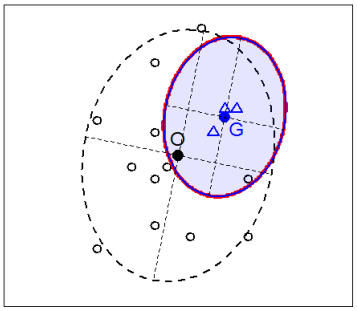


FIGURE 3.7 – *Exemple « Typicalité ».* Points–limites d'incompatibilité au seuil .05 (en rouge), zone de compatibilité au seuil .95 ($\kappa = 1.266$, en bleu) et ellipse de concentration du nuage M^I ($\kappa = 2$, en pointillés).

On voit bien sur la figure précédente que les points–limites d'incompatibilité appartiennent, aux tracas du discret près, à une ellipse de centre G, translatée d'une ellipse d'inertie du nuage M^I. Le point O est situé à l'intérieur de la zone de compatibilité : les points O et G sont compatibles au seuil .05.

1.1.3 Cas particulier d'un nuage unidimensionnel

Dans le cas d'un nuage unidimensionnel ($L = 1$), on dispose d'un ensemble I de N individus formant une population de référence à laquelle est associé le protocole numérique $x^I = (x^i)_{i \in I}$ de moyenne x_0 et de variance v ; et d'un groupe d'observations de taille n dont la moyenne observée est notée $\bar{x}$. Nous voulons étudier la typicalité du groupe d'observations par rapport à la population de référence.

Remarques :

- Le test de typicalité décrit précédemment peut être effectué dans le cas d'un nuage unidimensionnel, cependant nous en donnons ici une version spécifique envisageable uniquement dans ce cas particulier.
- *Choix de la statistique de test :* afin de tenir compte du côté où se situe la moyenne du groupe d'observations par rapport à moyenne de la population de référence (typicalité à droite ou à gauche), nous utilisons la statistique de test M (définie ci–après) qui mesure l'écart orienté entre ces deux moyennes, plutôt que la statistique D, utilisée précédemment, qui mesure la valeur absolue de cet écart.

Comme pour le cas multidimensionnel (*cf.* p.54), nous construisons l'ensemble $\mathcal{J}$ des $\mathrm{J} = \binom{N}{n}$ échantillons possibles de taille n issus de la population de référence. Notons $x^{I<j>}$ le j–ème échantillon.

Le principe du test est exactement le même que précédemment mais la statistique de test est, dans ce cas, la statistique *Moyenne*, notée M, définie par :

$$\begin{array}{rccc} M : & \mathcal{J} & \to & \mathbb{R} \\ & x^{I<j>} & \mapsto & \displaystyle\sum_{i \in I<j>} x^i/n \end{array}$$

La valeur observée de la statistique M est $\bar{x}$.

Seuil observé exact. Il s'impose d'effectuer ici un test *unilatéral* (*supérieur*, si $\bar{x} > x_0$ ou *inférieur*, si $\bar{x} < x_0$). Si $\bar{x} > x_0$, on calcule le seuil observé exact supérieur p_{sup}, c'est–à–dire la proportion des valeurs de M pour lesquelles $M \geq \bar{x}$; si $\bar{x} < x_0$, on calcule le seuil observé exact inférieur p_{inf}, c'est–à–dire la proportion des valeurs de M pour lesquelles $M \leq \bar{x}$. Le seuil observé exact unilatéral p_{unil} est par définition p_{sup} si $\bar{x} > x_0$ ou p_{inf} si $\bar{x} < x_0$.

La conclusion du test est énoncée, pour la statistique M, en termes d'atypicalité du groupe d'observations par rapport à la population de référence, en tenant compte du signe de l'effet.

Si $\bar{x} > x_0$ et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. Le groupe d'observations est atypique (à droite) de la population de référence au seuil unilatéral $\alpha/2$. La moyenne du groupe d'observations est significativement supérieure à la moyenne de la population de référence au seuil unilatéral $\alpha/2$.

Si $\bar{x} < x_0$ et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. Le groupe d'observations est atypique (à gauche) de la population de référence au seuil unilatéral $\alpha/2$. La moyenne du groupe d'observations est significativement inférieure à la moyenne de la population de référence au seuil unilatéral $\alpha/2$.

Si $p_{unil} > \alpha/2$, le test est non significatif au seuil bilatéral α. Nous ne pouvons pas dire que le groupe d'observations est atypique de la population de référence au seuil bilatéral α.

Caractéristiques de la distribution de la statistique de test M. D'après la théorie classique d'échantillonnage dans une population finie (*cf.* par exemple Morin, 1993 [60]), on a les propriétés suivantes :

Propriété 1.4. *La moyenne de la statistique Moyenne est :*

$$\mathrm{Moy}(M) = x_0$$

Propriété 1.5. *La variance de la statistique Moyenne est :*

$$\mathrm{Var}\, M = \frac{N-n}{N-1} \times \frac{v}{n}$$

Intervalle de compatibilité. En adaptant la définition 1.1 (p.61) au cas unidimensionnel, nous pouvons dire que les valeurs m ($\in \mathbb{R}$) et $\bar{x}$ sont compatibles au seuil unilatéral $\alpha/2$ si le seuil (unilatéral) observé associé est strictement supérieur à $\alpha/2$.

La limite inférieure (resp. supérieure) de l'intervalle de compatibilité au seuil $1-\alpha$ est alors la plus petite (resp. la plus grande) valeur m compatible avec $\bar{x}$ au seuil unilatéral $\alpha/2$.

1.2 Test approché

Supposons que le nuage d'échantillonnage C^J soit ajusté par une distribution gaussienne multidimensionnelle à L dimensions (L dimension de $\mathcal{M}$), de centre O et avec la structure de covariance associée à :

$$\frac{1}{n} \times \frac{N-n}{N-1} \times Som$$

c'est–à–dire telle que la densité soit définie par :

$$\forall\ \overrightarrow{x} \in \mathcal{L},\ f(\overrightarrow{x}) = \frac{1}{(2\pi)^{L/2} \det(Som \times \frac{1}{n} \times \frac{N-n}{N-1})^{1/2}} \exp\left(-\frac{1}{2} \times n \times \frac{N-1}{N-n} |\overrightarrow{x}|^2\right)$$

Dans ce cas, la distribution de la statistique :

$$\mathrm{T}^2 = n \times \frac{N-1}{N-n} \times D^2$$

est une distribution du χ^2 à L degrés de liberté (χ^2_L) (*cf.* définition de D p.54).

Seuil observé approché Soit $\mathrm{t}^2_{obs} = n \times \frac{N-1}{N-n} \times d^2_{obs}$ (avec $d_{obs} = |\overrightarrow{\mathrm{OG}}|$), la proportion $p(\mathrm{T}^2 \geq \mathrm{t}^2_{obs})$ est approximativement égale à $p(\chi^2_L \geq \mathrm{t}^2_{obs}) = \widetilde{p}_{obs}$, seuil observé du test approché (seuil approché ou p–value approchée).

Nous concluons en termes d'atypicalité du groupe d'observations par rapport à la population de référence, pour la statistique T (*cf.* p.55).

Exemple « Typicalité » : Sur la figure 3.8 est représentée la distribution de la statistique $\mathrm{T}^2 = n \times \frac{N-1}{N-n} \times D^2$, à cette distribution est superposée celle du χ^2 à 2 degrés de liberté (en rouge).

On a $\mathrm{t}^2_{obs} = n \times \frac{N-1}{N-n} \times d^2_{obs} = n \times \frac{N-1}{N-n} \times |\overrightarrow{\mathrm{OG}}|^2 = 3 \times \frac{14}{12} \times 1.089^2 = 4.152$, d'où le seuil observé approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq \mathrm{t}^2_{obs}) = p(\chi^2_2 \geq 4.152) = .125$ (à comparer avec le seuil exact $p_{obs} = \frac{61}{455} = .134$).

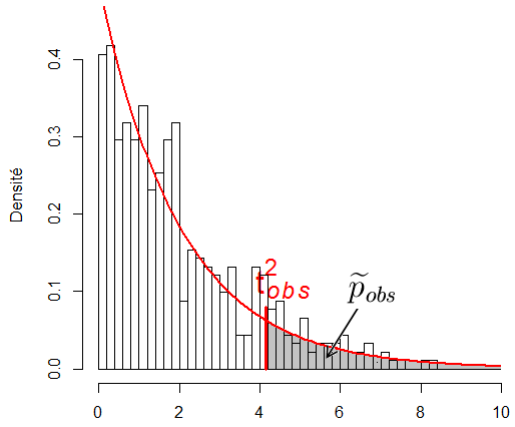


FIGURE 3.8 – *Exemple « Typicalité ».* Distributions de la statistique T^2 ($\mathrm{t}^2_{obs} = n \times \frac{N-1}{N-n} \times |\overrightarrow{\mathrm{OG}}|^2 = 4.152$) et du χ^2 à 2 degrés de liberté (en rouge).

Le test approché conduit à la même conclusion que le test exact.

1.2.1 Zone approchée de compatibilité

Rappelons que la zone de compatibilité au seuil $1-\alpha$ est définie comme étant l'ensemble des points P compatibles avec le point G au seuil α. Nous adoptons ici la même démarche que celle décrite au paragraphe 1.1.2 (p.61), la compatibilité entre les points P et G étant évidemment évaluée grâce au test approché.

Définition 1.4 (Zone–limite approchée d'incompatibilité)**.** *La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points–limites d'incompatibilité au seuil α, obtenus par le test approché.*

Soit $\chi^2_L[\alpha]$ la valeur critique de la distribution du χ^2 à L degrés de liberté au seuil supérieur α :

$$p(\chi^2_L \geq \chi^2_L[\alpha]) = \alpha$$

Théorème 1.2. *La zone–limite approchée d'incompatibilité au seuil* $1-\alpha$ *est le* κ*–ellipsoïde de centre* G*, translaté de l'ellipsoïde d'inertie du nuage de référence* M^I *tel que :*

$$\kappa = \sqrt{\frac{\chi^2_L[\alpha]}{n} \times \frac{N-n}{N-1}}$$

Démonstration. La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points P vérifiant :
$n \times \frac{(N-1)}{N-n} \times |\overrightarrow{\mathrm{PG}}|^2 = \chi^2_L[\alpha] \Leftrightarrow |\overrightarrow{\mathrm{PG}}|^2 = \frac{\chi^2_L[\alpha]}{n} \times \frac{N-n}{N-1}$.
c.q.f.d.

Exemple « Typicalité » : Sur la figure ci–après est représentée la zone approchée de compatibilité au seuil .95, elle est définie par l'ensemble des points situés à l'intérieur de l'ellipse de centre G, translatée de l'ellipse d'inertie du nuage M^I telle que $\kappa = \sqrt{\frac{\chi^2_L[.05]}{n} \times \frac{N-n}{N-1}} = \sqrt{\frac{5.991}{3} \times \frac{12}{14}} = 1.308$.

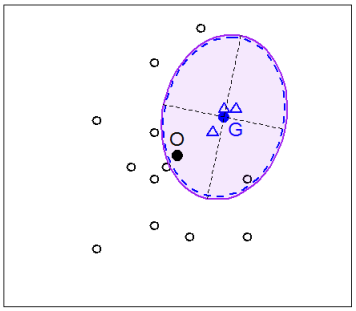


FIGURE 3.9 – *Exemple « Typicalité ».* Zone de compatibilité approchée au seuil .95 ($\kappa = 1.308$, en violet) et zone de compatibilité exacte au seuil .95 ($\kappa = 1.266$, en pointillés bleus).

Pour cet exemple, la zone de compatibilité approchée est plus large que la zone de compatibilité exacte.

1.2.2 Cas particulier d'un nuage unidimensionnel : test Z approché

Dans le cas où $L = 1$, le test approché consiste à ajuster la distribution d'échantillonnage de la statistique *Moyenne* M (*cf.* paragraphe 1.1.3, p.63) par une distribution gaussienne, de moyenne x_0 et de variance $\mathrm{Var}\, M (= \frac{N-n}{N-1} \times \frac{v}{n})$ et à utiliser la statistique *Écart Réduit* :

$$Z = \frac{M - x_0}{\sqrt{\mathrm{Var}\, M}}$$

dont la distribution est une distribution gaussienne $\mathcal{N}(0,1)$. Les deux statistiques Z et M conduisent à des tests équivalents.

Seuil observé approché. Soit $z_{obs} = \frac{\bar{x}-x_0}{\sqrt{\frac{v}{n}\times\frac{N-n}{N-1}}}$ la valeur observée de la statistique *Écart Réduit*. Notons z la variable dont la distribution est une loi de Gauss centrée réduite : $z \sim \mathcal{N}(0,1)$.

Si $\bar{x} > x_0$, on calcule le seuil observé approché supérieur $\widetilde{p}_{sup} = p(z \geq z_{obs})$; si $\bar{x} < x_0$, on calcule le seuil observé approché inférieur $\widetilde{p}_{inf} = p(z \leq z_{obs})$. Le seuil observé approché unilatéral $\widetilde{p}_{unil}$ est par définition $\widetilde{p}_{sup}$ si $\bar{x} > x_0$ ou $\widetilde{p}_{inf}$ si $\bar{x} < x_0$.

Remarque : On a $\widetilde{p}_{sup} = \widetilde{p}_{inf}$ car la distribution gaussienne est symétrique.

Nous énonçons la conclusion du test en termes d'atypicalité du groupe d'observations par rapport à la population de référence, pour la statistique Z, en tenant compte du signe de l'effet (*cf.* p.64).

Remarque : z_{obs} est appelée *valeur–test* par Lebart, Morineau & Piron, 1995 [54]. La valeur t^2_{obs} (*cf.* paragraphe 1.2, p.65) en est une généralisation multidimensionnelle.

Intervalle approché de compatibilité. Soit $z[\alpha]$ la valeur critique, au seuil bilatéral α, de la distribution gaussienne centrée réduite :

$$p(|z| > z[\alpha]) = \alpha$$

L'intervalle approché de compatibilité au seuil $1-\alpha$ est l'ensemble des valeurs $m \in \mathbb{R}$ compatibles avec $\bar{x}$ au seuil bilatéral α, c'est–à–dire vérifiant :

$$\frac{|m-\bar{x}|}{\sqrt{\text{Var}\, M}} < z[\alpha]$$

ou, ce qui revient au même, appartenant à l'intervalle :

$$]\bar{x} - z[\alpha]\sqrt{\text{Var}\, M}\,;\, \bar{x} + z[\alpha]\sqrt{\text{Var}\, D}[$$

Cet intervalle est centré sur $\bar{x}$.

2 Typicalité d'un groupe d'observations par rapport à une distribution de référence gaussienne : test Z

Au lieu de considérer une population de référence, nous considérons ici une *distribution de référence* L–gaussienne, de point moyen O et de structure de covariance définie par l'endomorphisme *Som*. Nous souhaitons étudier la typicalité du groupe d'observations, de taille n et de point moyen G, par rapport à cette distribution de référence. Pour ce faire, nous utilisons le test Z multidimensionnel présenté ci–après.

D'après la théorie de l'échantillonnage dans une distribution, la distribution d'échantillonnage de la statistique *Point Moyen* (ensemble des points moyens des échantillons de taille n issus de la distribution de référence) est une distribution L–gaussienne, centrée sur O et de structure de covariance $\frac{1}{n} \times Som$. La statistique $Z^2 : n \times D^2$ est alors distribuée selon un χ^2 à L degrés de liberté (*cf.* définition de D p.54).

Remarque : Ce test peut être qualifié de test *exact* car il s'agit ici d'échantillonner dans une distribution gaussienne et non d'approcher la distribution d'échantillonnage par une distribution gaussienne.

2.1 Seuil observé du test

Le seuil observé du test est la probabilité d'échantillonnage que la statistique Z^2 ($\sim \chi^2_L$) dépasse sa valeur observée $z^2_{obs} = n \times d^2_{obs} = n \times |\overrightarrow{\mathrm{OG}}|^2$.

Exemple « Typicalité » : On a $z^2_{obs} = n \times |\overrightarrow{\mathrm{OG}}|^2 = 3 \times 1.089^2 = 3.558$ et $p(\chi^2_2 \geq 3.558) = .169$, le test Z est donc non significatif (au seuil $\alpha = .05$). Le test Z conduit à la même conclusion que le test de typicalité pour lequel $p_{obs} = .134$.

2.2 Zone de compatibilité

Translatons maintenant la distribution de référence de telle sorte que le point P ($\in \mathcal{M}$) soit le point moyen de cette distribution de référence translatée. Si le test Z, comparant la distribution de référence translatée au groupe d'observations, est non significatif au seuil α, alors nous pouvons dire que les points P et G sont compatibles au seuil α. La zone de compatibilité entre le point moyen G du groupe d'observations et le point moyen de la distribution de référence, au seuil $1 - \alpha$, est alors définie comme l'ensemble des points P compatibles avec le point G au seuil α, nous l'appelons *zone de compatibilité du test Z*.

Pour α fixé, les points P vérifiant cette propriété sont tels que :

$$n \times |\overrightarrow{\mathrm{PG}}|^2 < \chi^2_L[\alpha]$$

où $\chi^2_L[\alpha]$ est la valeur critique de la distribution du χ^2 à L degrés de liberté au seuil supérieur α. Ils sont donc situés à l'intérieur de l'ellipsoïde de centre G défini par l'ensemble des points Q ($\in \mathcal{M}$) tels que :

$$|\overrightarrow{\mathrm{QG}}| = \sqrt{\frac{\chi^2_L[\alpha]}{n}}$$

c'est–à–dire à l'intérieur de l'ellipsoïde de centre G, translaté de l'ellipsoïde d'inertie du nuage de référence M^I tel que :

$$\kappa = \sqrt{\frac{\chi^2_L[\alpha]}{n}}$$

Exemple « Typicalité » : Sur la figure ci–après est représentée la zone de compatibilité du test Z au seuil .95, elle est définie par l'ensemble des points situés à l'intérieur de l'ellipse de centre G, translatée de l'ellipse d'inertie du nuage M^I telle de $\kappa = \sqrt{\frac{\chi^2_2[.05]}{n}} = \sqrt{\frac{5.991}{3}} = 1.413$.

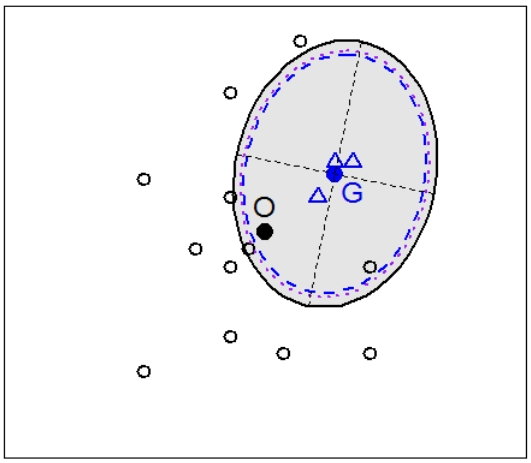


FIGURE 3.10 – *Exemple « Typicalité ».* Zone de compatibilité du test Z au seuil .95 ($\kappa = 1.413$, en gris), zone de compatibilité exacte au seuil .95 ($\kappa = 1.266$, en bleu, pointillés longs) et zone de compatibilité approchée au seuil .95 ($\kappa = 1.308$, en violet, pointillés courts).

Pour cet exemple, la zone de compatibilité du test Z est la plus large des trois zones, ici le test Z est celui qui renvoie le moins de résultats significatifs.

2.3 Cas particulier d'un nuage unidimensionnel : test Z unidimensionnel

Considérons une distribution de référence gaussienne, de moyenne x_0 et de variance v. Nous voulons étudier la typicalité du groupe d'observations, de taille n et de moyenne $\bar{x}$, par rapport à cette distribution de référence. Pour ce faire, nous utilisons le test Z unidimensionnel décrit ci–après.

D'après la théorie de l'échantillonnage dans une distribution, la distribution d'échantillonnage de la statistique *Moyenne* est gaussienne, et plus précisément on a :

$$M \sim \mathcal{N}(x_0, \frac{v}{n})$$

soit encore, en posant $Z = \frac{M - x_0}{\sqrt{v/n}}$ (statistique *Écart Réduit*) :

$$Z \sim \mathcal{N}(0, 1)$$

Seuil observé. La valeur observée de la statistique *Écart Réduit* est :

$$z_{obs} = \frac{\bar{x} - x_0}{\sqrt{v/n}}$$

Le seuil observé unilatéral est alors défini comme étant la probabilité que la statistique de test Z soit plus extrême, du côté des données, que la statistique observée z_{obs}.

Intervalle de compatibilité. Comme pour le cas multidimensionnel, nous translatons la distribution de référence de telle sorte que la valeur m ($\in \mathbb{R}$) soit la moyenne de cette distribution de référence translatée. Si le test Z, comparant la distribution de référence

translatée au groupe d'observations, est non significatif au seuil unilatéral $\alpha/2$, alors nous pouvons dire que les valeurs m et $\bar{x}$ sont compatibles au seuil unilatéral $\alpha/2$. L'intervalle de compatibilité entre la moyenne du groupe d'observations $\bar{x}$ et la moyenne de la distribution de référence, au seuil $1-\alpha$, est alors défini comme l'ensemble des valeurs m compatibles avec $\bar{x}$ au seuil unilatéral $\alpha/2$.

Soit $z[\alpha]$ la valeur critique, au seuil bilatéral α, de la distribution gaussienne centrée réduite, l'intervalle de compatibilité au seuil $1-\alpha$ est l'ensemble des valeurs m vérifiant :

$$\frac{|\bar{x}-x_0|}{\sqrt{v/n}} < z[\alpha]$$

ou, ce qui revient au même, appartenant à l'intervalle :

$$]\bar{x} - z[\alpha](\sqrt{v/n})\,;\, \bar{x} + z[\alpha](\sqrt{v/n})[$$

Cet intervalle est centré sur $\bar{x}$.

3 Applications

Dans cette section, nous illustrons le test ensembliste par les exemples *Races canines* et *Parkinson*. Dans l'exemple *Races canines*, le groupe d'observations appartient à la population de référence ; ce n'est pas le cas pour l'exemple *Parkinson*.

3.1 Exemple des races canines

Les données. Dans cette partie, nous nous intéressons aux données des *races canines*, provenant de l'exemple tiré de Tenenhaus (*cf.* [84] 2007, p.254). 27 races canines sont décrites par 6 variables catégorisées : « taille » (3 modalités), « poids » (3 modalités), « vélocité » (3 modalités), « intelligence » (3 modalités), « affection » (2 modalités) et « agressivité » (2 modalités). Une septième variable « fonction » permet de répartir les 27 races en trois groupes, selon trois catégories de fonctions :

- Groupe 1 : « compagnie »,
- Groupe 2 : « chasse »,
- Groupe 3 : « utilité ».

Une analyse des correspondances multiples (ACM) portant sur les données des 27 individus (races canines) et les six variables caractéristiques est effectuée. L'étude des variances des axes montre que les deux premiers axes sont les plus importants. En effet, l'écart entre la deuxième et la troisième valeur propre est grand et la décroissance des valeurs propres, à partir de la troisième, est régulière (*cf.* table 3.1 et figure 3.11). De plus, en considérant les deux premiers axes, nous arrivons à un taux modifié cumulé (*cf.* Benzécri, 1984, p.306 [6]) de 98.68%. Dans la suite, nous nous intéressons donc aux individus (races canines) projetés sur le premier plan principal et aux trois groupes définis par la variable « fonction » (figure 3.12).

Pour cet exemple, l'ensemble I caractérisant la *population de référence* est l'ensemble des 27 races canines. A cette population de référence est associé le nuage de référence M^I de 27 points et de point moyen O, que nous étudions dans le premier plan principal ($L=2$) (*cf.* figure 3.12). Les coordonnées des 27 points $(\mathrm{M}^i)_{i\in I}$ sur les deux premiers axes et la répartition des races canines dans les trois groupes sont données dans la table 3.2.

	Valeur Propre	Taux modifié	Taux modifié cumulé
Axe 1	0.4816	0.6671	0.6671
Axe 2	0.3847	0.3197	0.9868
Axe 3	0.2110	0.0132	1.0000
Axe 4	0.1576		
Axe 5	0.1501		
Axe 6	0.1233		
Axe 7	0.0815		
Axe 8	0.0457		
Axe 9	0.0235		
Axe 10	0.0077		

TABLE 3.1 – Valeurs propres et taux modifiés.

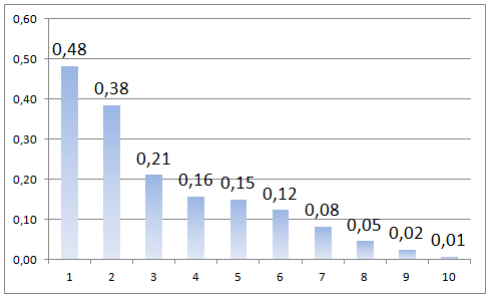


FIGURE 3.11 – Décroissance des valeurs propres.

	x_1	x_2	Groupe
1–Beauceron	0.317	−0.418	3
2–Basset	−0.254	1.101	2
3–Berger Allemand	0.486	−0.464	3
4–Boxer	−0.447	−0.882	1
5–Bull-Dog	−1.013	0.550	1
6–Bull-Mastiff	0.753	0.547	3
7–Caniche	−0.912	−0.016	1
8–Chihuahua	−0.841	0.844	1
9–Cocker	−0.733	0.079	1
10–Colley	0.117	−0.526	1
11–Dalmatien	−0.647	−0.990	1
12–Dobermann	0.873	−0.315	3
13–Dogue Allemand	1.047	0.507	3
14–Epagneul Breton	−0.478	−1.037	2
15–Epagneul Français	0.145	−0.516	2
16–Fox-Hound	0.877	0.025	2
17–Fox-Terrier	−0.882	0.139	1
18–Grand bleu de Gascogne	0.517	−0.113	2
19–Labrador	−0.647	−0.990	2
20–Lévrier	0.677	−0.083	2
21–Mastiff	0.756	0.888	3
22–Pékinois	−0.841	0.844	1
23–Pointer	0.673	−0.424	2
24–Saint-Bernard	0.583	0.594	3
25–Setter	0.504	−0.377	2
26–Teckel	−1.013	0.550	1
27–Terre-Neuve	0.384	0.485	3
Moyenne	0.000	0.000	

TABLE 3.2 – Coordonnées des 27 races canines sur le premier plan principal (x_1 et x_2) et partition en trois groupes.

Les individus 5 et 26 ont des profils identiques, ils sont donc représentés par deux points confondus sur le premier plan principal, il en est de même pour les individus 8 et 22 et pour les individus 11 et 19 (*cf.* figure 3.12). Pour rendre compte de cette pondération, les points du nuage M^I ont une taille proportionnelle à leurs poids.

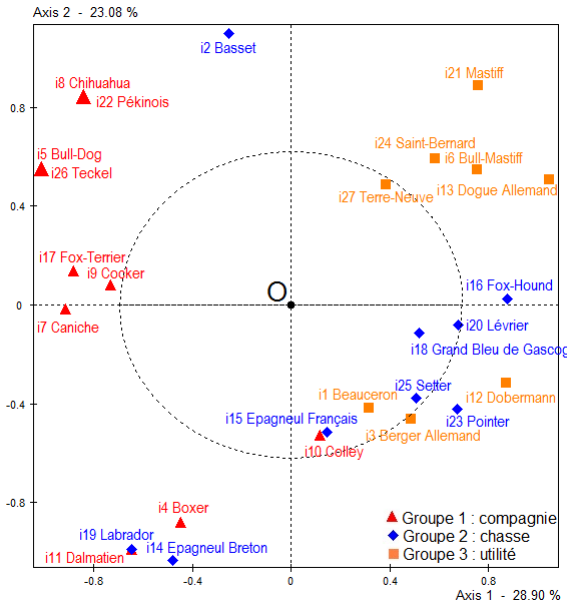


FIGURE 3.12 – *Exemple des Races canines.* Nuage des 27 individus dans le plan 1-2 avec son point moyen O et son ellipse indicatrice ($\kappa = 1$).

Considérons les 3 sous–nuages des $n_1 = 10$ points du groupe 1 (figure 3.13, gauche), des $n_2 = 9$ points du groupe 2 (figure 3.13, centre) et des $n_3 = 8$ points du groupe 3 (figure 3.13, droite) avec leurs points moyens respectifs G^1, G^2 et G^3.

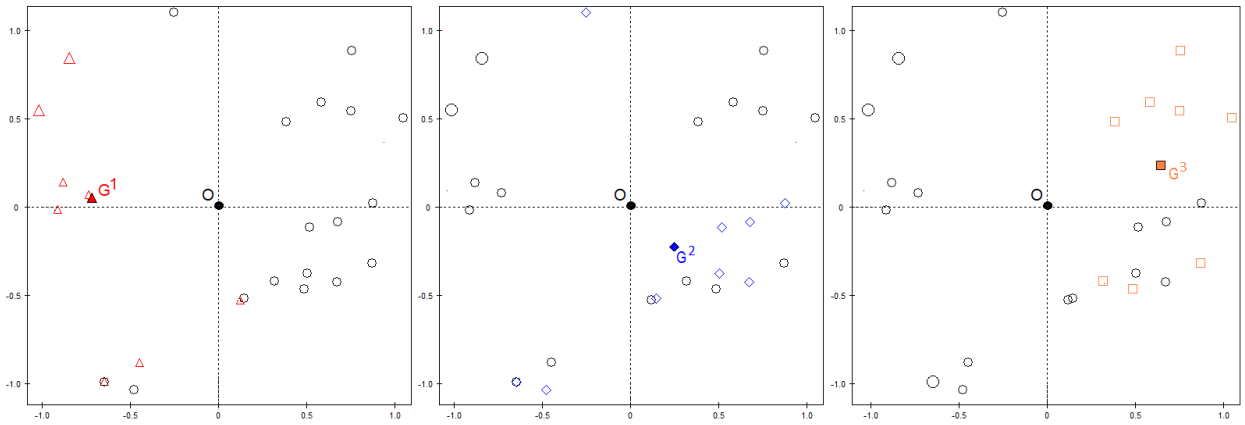


FIGURE 3.13 – *Exemple des Races canines.* Sous–nuages des trois groupes et points moyens G^1, G^2 et G^3.

Pour cet exemple, on a :

$\overrightarrow{OG^1} = \begin{pmatrix} -0.721 \\ 0.059 \end{pmatrix}$, $\overrightarrow{OG^2} = \begin{pmatrix} 0.224 \\ -0.268 \end{pmatrix}$ et $\overrightarrow{OG^3} = \begin{pmatrix} 0.650 \\ 0.228 \end{pmatrix}$;

$\mathbf{V} = \begin{pmatrix} 0.481 & 0 \\ 0 & 0.385 \end{pmatrix}$, matrice de covariance du nuage M^I, associée à *Som*.

D'où, $|\overrightarrow{OG^1}| = \sqrt{\begin{pmatrix} -0.721 & 0.059 \end{pmatrix} \times \begin{pmatrix} 0.481 & 0 \\ 0 & 0.385 \end{pmatrix}^{-1} \times \begin{pmatrix} -0.721 \\ 0.059 \end{pmatrix}} = 1.044$ (écart descriptif important), $|\overrightarrow{OG^2}| = 0.539$ (écart descriptif négligeable) et $|\overrightarrow{OG^3}| = 1.006$ (écart descriptif important).

Nous nous proposons de répondre à la question suivante : *le groupe 1 (« compagnie ») est-il atypique de la population de référence des 27 races canines ?* (illustrée par le plan gauche de la figure 3.13)

Remarque : Le groupe d'observations (groupe 1) appartient ici à la population de référence (ensemble des 27 races canines).

3.1.1 Test exact

En appliquant le test ensembliste comparant la population de référence des 27 races canines au groupe 1, nous obtenons les résultats suivants :

Seuil observé exact. Au j–ème sous–nuage de l'espace des échantillons est associé son point moyen C^j, d'où le nuage d'échantillonnage C^J des $\binom{27}{10}$ points moyens. La distribution de la statistique $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|$ est donnée sur la figure 3.14, on a $d_{obs} = |\overrightarrow{\mathrm{OG}^1}| = 1.044$ et $p_{obs} = p(D \geq d_{obs}) = 182/\binom{27}{10} = 182/8\,436\,285 = .00002$. Le seuil observé exact s'interprète géométriquement comme la proportion des points du nuage d'échantillonnage situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par G^1 (figure 3.15).

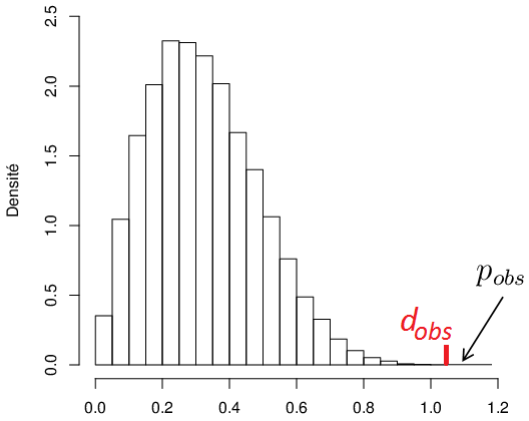


FIGURE 3.14 – *Exemple des Races canines, typicalité du groupe 1.* Distribution de la statistique D ($d_{obs} = |\overrightarrow{\mathrm{OG}^1}| = 1.044$).

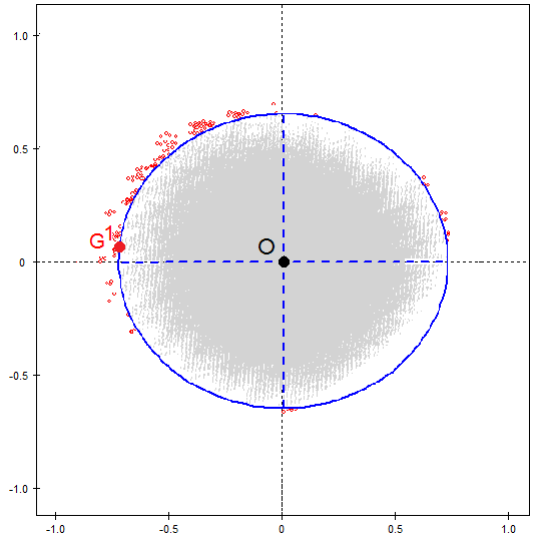


FIGURE 3.15 – *Exemple des Races canines, typicalité du groupe 1.* Interprétation géométrique du seuil exact : proportion des points C^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par G^1 (182 points rouges).

Conclusion : On a $p_{obs} < .05$, résultat significatif au seuil .05. Pour la statistique D, nous pouvons dire que le groupe 1 (« compagnie ») est atypique de la population de référence des 27 races canines au seuil .05.

Remarque : Nous étudions de la même façon la typicalité du groupe 3 « utilité » (figure 3.13, plan de droite) par rapport à la population de référence des 27 races canines : $p_{obs} = 0.002$ (résultat significatif au seuil .05).

Le test comparant le groupe 2 par rapport à la population de référence n'est pas effectué car l'écart descriptif séparant le point G^2 du point O ($|\overrightarrow{OG^2}| = 0.539$) est négligeable.

Zone de compatibilité. En mettant en oeuvre la procédure de construction de la zone de compatibilité au seuil $1-\alpha$ décrite au paragraphe 1.1.2 (p.61), nous avons recherché un ensemble de points P juste incompatibles avec le point G^1 aux seuils $\frac{421\,814}{8\,436\,285} = 0.04999997$ et $\frac{421\,815}{8\,436\,285} = 0.05000009$ [8] (il n'y a pas de point–limite correspondant au seuil .05, en effet $8\,436\,285 \times 0.05$ n'est pas un nombre entier). Sur la figure 3.16 est représentée en bleu l'ellipse ajustée à cet ensemble de points, il s'agit de l'ellipse de centre G^1, translatée de l'ellipse d'inertie du nuage de référence telle que $\kappa = 0.619$ (la méthode d'ajustement est la même que celle décrite en bas de la page 62). La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

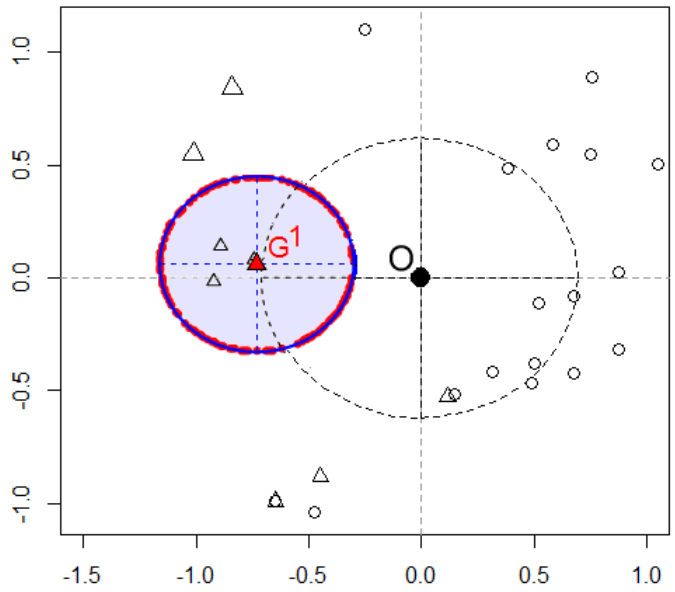


FIGURE 3.16 – *Exemple des Races canines, typicalité du groupe 1.* Points–limites d'incompatibilité au seuil .05 (en rouge), zone de compatibilité au seuil .95 ($\kappa = 0.619$, en bleu) et ellipse indicatrice du nuage de référence ($\kappa = 1$, en pointillés).

Le point O est situé à l'extérieur de la zone de compatibilité, les points G^1 et O sont incompatibles au seuil .05.

3.1.2 Test approché

Seuil observé approché. On a $t^2_{obs} = \frac{10\times(27-1)}{27-10} \times d^2_{obs} = \frac{260}{17} \times 1.044^2 = 16.670$ (*cf.* p.64), d'où le seuil approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq 16.670) = .00024$ (à comparer avec le seuil exact $p_{obs} = 182/\binom{27}{10} = .00002$).

Le test approché conduit à la même conclusion que le test exact.

8. *cf.* note sur le calcul p.62. Quadrillage horizontal : -1.2 → -0.2 par pas de 0.01, quadrillage vertical : -0.4 → 0.5 par pas de 0.01.

Zone de compatibilité approchée. On a $\chi^2_2[.05] = 5.991$, d'après le théorème 1.2 (p.66), la zone–limite approchée d'incompatibilité au seuil .95 est l'ellipse de centre G^1, translatée de l'ellipse d'inertie du nuage de référence telle que $\kappa = \sqrt{\frac{5.991}{10} \times \frac{27-10}{27-1}} = 0.626$.

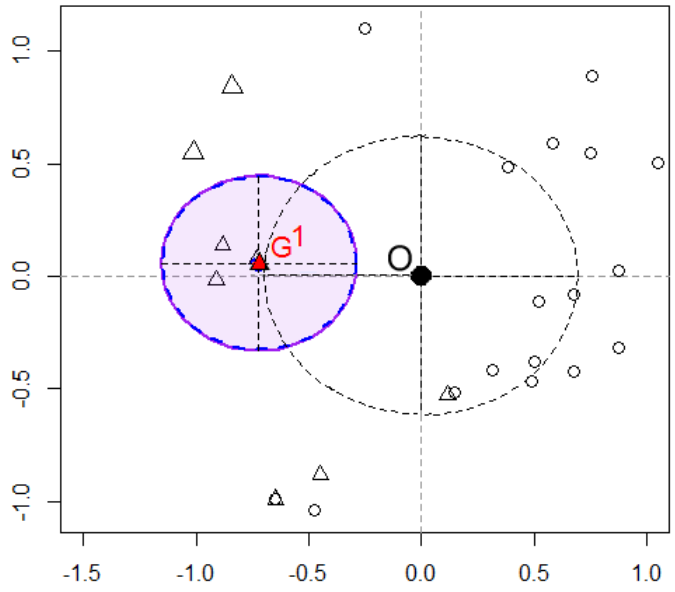


FIGURE 3.17 – *Exemple des Races canines, typicalité du groupe 1.* Zone de compatibilité approchée au seuil .95 ($\kappa = 0.626$, en violet), zone de compatibilité exacte au seuil .95 ($\kappa = 0.619$, en pointillés bleus) et ellipse indicatrice du nuage de référence M^I ($\kappa = 1$, en pointillés noirs).

Pour cet exemple, la zone de compatibilité approchée est plus large que la zone de compatibilité exacte. Le point O est situé à l'extérieur de la zone de compatibilité approchée : les points O et G^1 sont incompatibles au seuil .05.

3.2 Données Parkinson

Dans ce paragraphe, nous considérons les données de l'exemple *Parkinson* présenté au chapitre 2 (p.44).

Nous avons précédemment évalué l'effet du traitement grâce au test géométrique (*cf.* chapitre 2). Nous nous demandons maintenant si les malades observés après traitement peuvent être considérés comme étant similaires aux bien–portants. Pour ce faire, nous utilisons le test ensembliste de typicalité permettant de les comparer.

Les coordonnées des 45 bien–portants et des 15 malades observés après traitement sur les deux premiers axes principaux sont données dans la table 3.3.

L'ensemble I caractérisant la *population de référence* est ici l'ensemble des 45 individus bien–portants. A cette population de référence est associé le nuage de référence M^I de 45 points et de point moyen O, que nous étudions dans le premier plan principal ($L = 2$). Considérons également le groupe des 15 malades observés après traitement et le nuage de 15 points, de point moyen G, qui lui est associé (*cf.* figure 3.18).

	Bien–portants	
	x_1	x_2
1	3.091	$-0,828$
2	-2.420	$1,632$
3	0.501	$1,492$
4	-0.153	$0,651$
5	1.040	$0,911$
6	2.847	$0,756$
7	1.350	$-3,758$
8	$1,899$	$0,362$
9	0.180	$0,840$
10	-2.319	$-0,501$
11	-0.354	$0,750$
12	0.575	$0,515$
13	2.568	$0,774$
14	-1.828	$-0,136$
15	-2.105	$1,425$
16	2.887	$2,015$
17	-0.233	$-0,185$
18	2.437	$-0,923$
19	1.961	$0,951$
20	0.475	$0,632$
21	-2.080	$-1,215$
22	-1.703	$0,985$
23	1.802	$0,593$
24	-3.105	$-2,882$
25	-2.412	$-1,535$
26	2.553	$0,786$
27	$-3,967$	$-1,042$
28	$-1,599$	$0,983$
29	$-1,757$	$-0,690$
30	$1,667$	$0,258$
31	$0,464$	$0,027$
32	$-4,165$	$2,373$
33	$0,546$	$-2,253$
34	$0,751$	$-3,800$
35	$-1,144$	$-1,261$
36	$0,198$	$0,981$
37	$-2,319$	$0,186$
38	$0,643$	$-0,751$
39	$-0,170$	$0,343$
40	$-0,225$	$1,165$
41	$0,545$	$0,182$
42	$-1,142$	$1,243$
43	$3,669$	$-0,944$
44	$-2,300$	$-0,290$
45	$2,853$	$-0,815$
Moyenne	0.000	0.000

	Malades après traitement	
	x_1	x_2
1	$-1,286$	$1,854$
2	$-0,193$	$-1,562$
3	$-1,546$	$-4,651$
4	$0,158$	$-0,912$
5	$-0,006$	$-2,335$
6	$1,535$	$-4,765$
7	$-3,465$	$-1,430$
8	$0,074$	$-1,737$
9	$-0,075$	$-1,814$
10	$-1,777$	$-1,188$
11	$-1,576$	$-3,270$
12	$-0,420$	$1,012$
13	$-5,907$	$1,014$
14	$0,434$	$-4,095$
15	$-1,800$	$-1,069$
Moyenne	-1.057	-1.663

TABLE 3.3 – Coordonnées des 45 bien–portants et des 15 malades observés après traitement sur le premier plan principal. x_1 : coordonnées sur l'axe 1 ; x_2 : coordonnées sur l'axe 2.

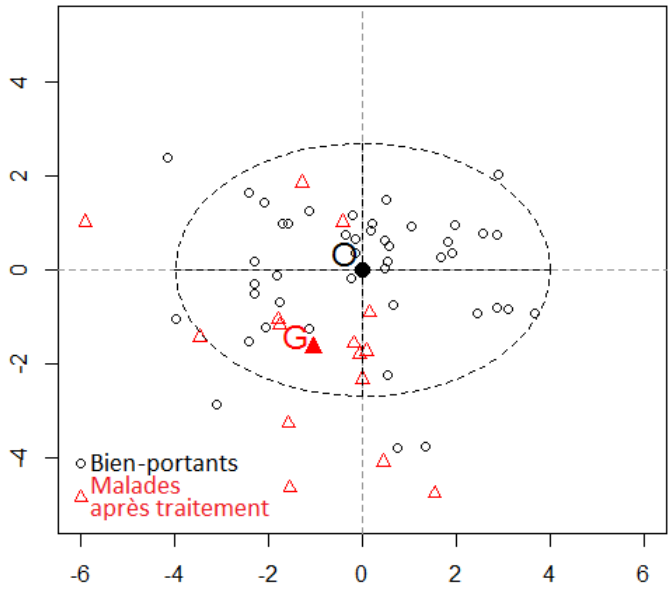


FIGURE 3.18 – *Exemple Parkinson.* Nuage des 45 bien–portants (ronds noirs) avec son point moyen O et son ellipse de concentration ($\kappa = 2$) ; et nuage des 15 malades après traitement (triangles rouges) avec son point moyen G.

On a $\overrightarrow{\mathrm{OG}} \begin{pmatrix} -1.057 \\ -1.663 \end{pmatrix}$ et $\mathbf{V} = \begin{pmatrix} 3.993 & 0 \\ 0 & 1.822 \end{pmatrix}$, matrice de covariance du nuage M^I, associée à *Som.*

D'où, $|\overrightarrow{\mathrm{OG}}| = \sqrt{\begin{pmatrix} -1.057 & -1.663 \end{pmatrix} \times \begin{pmatrix} 3.993 & 0 \\ 0 & 1.822 \end{pmatrix}^{-1} \times \begin{pmatrix} -1.057 \\ -1.663 \end{pmatrix}} = 1.341$ (écart descriptif important).

Nous nous posons la question de recherche suivante : *le groupe des malades observés après traitement est-il atypique de la population de référence des bien–portants ?*

Remarque : Dans le cas présent, le groupe d'observations (malades après traitement) n'appartient pas à la population de référence (bien–portants).

3.2.1 Test exact

Ensemble et espace des échantillons. Pour cet exemple, le nombre combinatoire $\binom{45}{15} = 3.44 \times 10^{11}$ est trop important pour que l'ensemble des échantillons « complet » soit construit. Nous utilisons la méthode de Monte Carlo afin d'engendrer un ensemble « restreint » des échantillons possibles. Pour cet exemple, nous choisissons de considérer un ensemble « restreint » de J = 50 000 échantillons de taille 15 issus de la population de référence de taille 45 [9]. A cet ensemble « restreint », nous associons l'espace des échantillons correspondant (*cf.* paragraphe 1.1.1, p.54).

Seuil observé exact. Au j–ème sous–nuage de l'espace des échantillons « restreint » est associé son point moyen C^j, d'où le nuage d'échantillonnage « restreint » de = 50 000 points

9. Pour des raisons de temps de calcul, nous ne vérifions pas le fait que les échantillons engendrés soient différents. En effet, la probabilité d'obtenir plusieurs fois le même échantillon est très faible.

moyens. La distribution de la statistique $D : \mathrm{C}^j \mapsto |\overrightarrow{\mathrm{OC}^j}|$ est donnée sur la figure 3.19, on a $d_{obs} = |\overrightarrow{\mathrm{OG}}| = 1.341$ et $p_{obs} = p(D \geq d_{obs}) = 0$ [10]. Aucun point du nuage d'échantillonnage n'est situé sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par G (interprétation géométrique du seuil observé, figure 3.20).

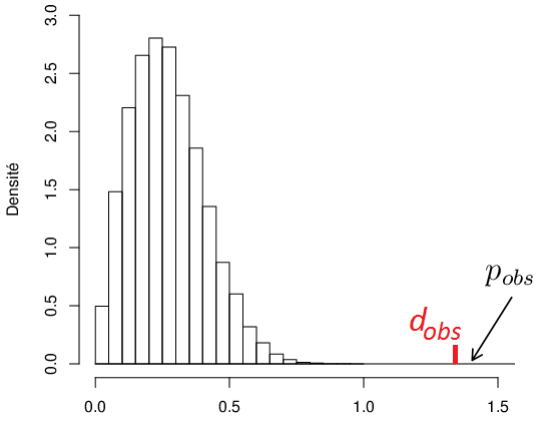


FIGURE 3.19 – *Exemple Parkinson.* Distribution de la statistique D ($d_{obs} = |\overrightarrow{\mathrm{OG}^1}| = 1.341$).

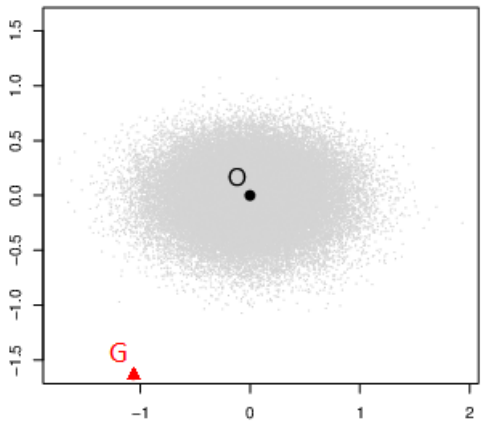


FIGURE 3.20 – *Exemple Parkinson.* Interprétation géométrique du seuil exact : proportion des points C^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par G (aucun point ici).

Conclusion : On a $p_{obs} < .05$, résultat significatif au seuil .05. Pour la statistique D, nous pouvons dire que le groupe des malades observés après traitement est atypique de la population de référence des bien–portants. Même si les données sont en faveur d'un effet du traitement (*cf.* chapitre 2, p.48), les malades n'acquièrent pas une marche similaire à celle des bien–portants après la prise du traitement.

Zone de compatibilité. Nous avons recherché un ensemble de points P juste incompatibles avec le point G au seuil $\frac{2\,500}{50\,000} = 0.05$ [11] (rappelons que nous considérons ici un nuage d'échantillonnage « restreint » de 50 000 points), ces points sont représentés en rouge sur la figure 3.21. Sur la même figure est représentée en bleu l'ellipse ajustée à cet ensemble de points, il s'agit de l'ellipse de centre G, translatée de l'ellipse d'inertie du nuage de référence telle que $\kappa = 0.515$ (la méthode d'ajustement est la même que celle décrite en bas de la page 62). La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

10. Afin de s'assurer de la stabilité du seuil observé obtenu par la méthode de Monte Carlo, nous avons effectué 10 simulations indépendantes ; pour cet exemple, les 10 seuils observés sont tous égaux à 0.

11. *cf.* note sur le calcul p.62. Quadrillage horizontal : -2.5 → 0 par pas de 0.05, quadrillage vertical : -2.5 → -0.5 par pas de 0.01.

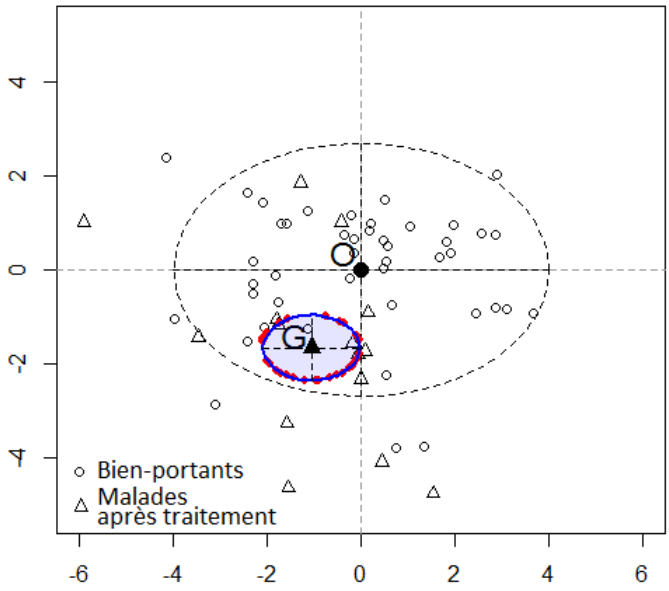


FIGURE 3.21 – *Exemple Parkinson.* Zone de compatibilité au seuil .95 ($\kappa = 0.515$, en bleu) et ellipse de concentration du nuage de référence ($\kappa = 2$, en pointillés).

Le point O est situé à l'extérieur de la zone de compatibilité, les points G et O sont incompatibles au seuil .05.

3.2.2 Test approché

Seuil observé approché. On a $t^2_{obs} = \frac{15\times(45-1)}{45-15} \times d^2_{obs} = 22 \times 1.341^2 = 39.562$ (*cf.* p.64), d'où le seuil approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq 39.562) = 2.57 \times 10^{-9}$ (à comparer avec le seuil exact $p_{obs} = 0$).

Le test approché conduit à la même conclusion que le test exact.

Zone de compatibilité approchée. On a $\chi^2_2[.05] = 5.991$, d'après le théorème 1.2 (p.66), la zone–limite approchée d'incompatibilité au seuil .95 est l'ellipse de centre G, translatée de l'ellipse d'inertie du nuage de référence telle que $\kappa = \sqrt{\frac{5.991}{15} \times \frac{45-15}{45-1}} = 0.522$.

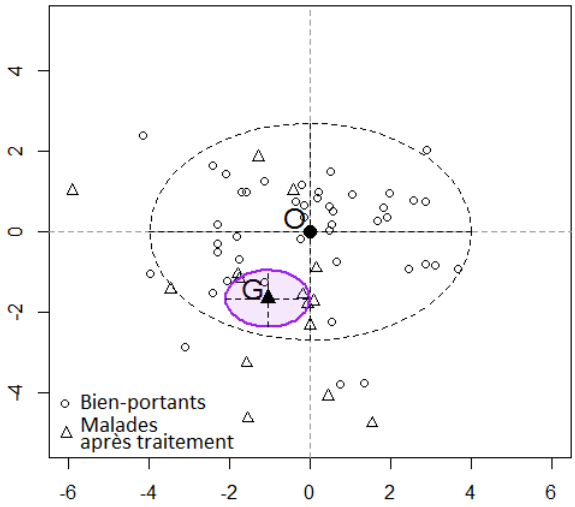


FIGURE 3.22 – *Exemple Parkinson.* Zone de compatibilité approchée au seuil .95 ($\kappa = 0.522$, en violet) et ellipse de concentration du nuage de référence ($\kappa = 2$, en pointillés).

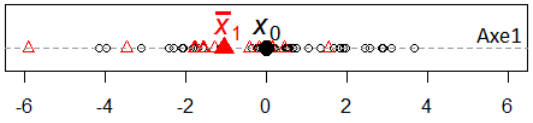
La zone de compatibilité approchée est à comparer avec la zone de compatibilité exacte au seuil .95 ($\kappa = 0.515$). Pour cet exemple, la zone de compatibilité approchée est plus large que la zone de compatibilité exacte. Le point O est situé à l'extérieur de la zone de compatibilité approchée : les points O et G sont incompatibles au seuil .05.

3.2.3 Tests unidimensionnels.

Le test présenté précédemment compare les malades observés après traitement aux bien–portants sur le premier plan principal, c'est–à–dire en tenant compte de la *performance* et du *style* de la marche (interprétation des axes 1 et 2). En complément, nous pouvons aussi nous poser les questions suivantes :

1. Les malades observés après traitement sont-ils atypiques des bien–portants pour la *performance* de la marche ? (interprétation de l'axe 1)
2. Les malades observés après traitement sont-ils atypiques des bien–portants pour le *style* de la marche ? (interprétation de l'axe 2)

Pour traiter la première question, considérons d'une part, le protocole unidimensionnel constitué des coordonnées sur l'axe 1 des 45 bien–portants (ronds noirs sur la figure ci–après), de moyenne $\bar{x_0} = 0$ et de variance $v = 3.993$; et d'autre part, celui constitué des coordonnées sur l'axe 1 des 15 malades observés après traitement (triangles rouge sur la figure ci–après), de moyenne $\bar{x_1} = -1.057$ (*cf.* table 3.3, p.76). Il s'agit maintenant de comparer la *performance* de la marche des malades après traitement à celle des bien–portants.



On a $\bar{x_1} < x_0$ et $d_{cal} = (\bar{x_1} - x_0)/\sqrt{v} = -1.057/\sqrt{3.993} = -0.529$, l'écart observé est notable (*cf.* p.153 pour la définition de l'écart calibré d_{cal}). Descriptivement, la performance

de la marche des malades après traitement est moins bonne que celle des bien–portants. La question 1 énoncée ci–avant peut donc être reformulée de la façon suivante :

La performance de la marche des malades après traitement est-elle significativement moins bonne que celle des bien–portants ?

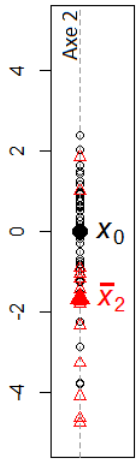
Les résultats des tests ensembliste de typicalité exact et approché et du test Z répondant à cette question sont donnés dans le tableau suivant (tests unilatéraux) :

$\bar{x}$	x_0	Test ensembliste				Test Z	
		Exact		Approché			
		p–value	I.C.	p–value	I.C.	p–value	I.C.
-1.057	0	.0061	$]-1.757;-0.357[$	.0065 [a]	] -1.757 ;-0.356 [	.020	] -1.905 ;-0.208 [

a. Appelée *valeur–test* par Lebart, Morineau & Piron, 1995 [54].

Conclusion : Pour l'ensemble de ces tests, le seuil observé est inférieur à .025. Nous pouvons dire que la *performance* de la marche des malades observés après traitement est moins bonne que celle des bien–portants.

De la même façon, pour répondre à la seconde question, considérons d'une part, le protocole unidimensionnel constitué des coordonnées sur l'axe 2 des 45 bien–portants (ronds noirs sur la figure ci–après), de moyenne $\bar{x_0} = 0$ et de variance $v = 1.822$; et d'autre part, celui constitué des coordonnées sur l'axe 2 des 15 malades observés après traitement (triangles rouges sur la figure ci–après), de moyenne $\bar{x_2} = -1.663$ (*cf.* table 3.3, p.76). Il s'agit maintenant de comparer le *style* de la marche des malades après traitement à celui des bien–portants.



On a $\bar{x_2} < x_0$ et $d_{cal} = (\bar{x_2} - x_0)/\sqrt{v} = -1.663/\sqrt{1.822} = -1.232$, l'écart observé est important. Descriptivement, le style de la marche des malades après traitement est moins bon que celui des bien–portants. La question 2 énoncée ci–avant peut donc être reformulée de la façon suivante :

Le style de la marche des malades après traitement est-il significativement moins bon que celui des bien–portants ?

Les résultats des tests ensembliste de typicalité exact et approché et du test Z répondant à cette question sont donnés dans le tableau suivant (tests unilatéraux) :

$\bar{x}$	x_0	Test ensembliste				Test Z	
		Exact		Approché			
		p–value	I.C.	p–value	I.C.	p–value	I.C.
-1.663	0	0.000	$]-2.137;-1.191[$	3.8×10^{-9}	] -2.136 ; -1.190[	9.1×10^{-7}	] -2.237 ;-1.090 [

Conclusion : Pour l'ensemble de ces tests, le seuil observé est inférieur à .025. Nous pouvons dire que le *style* de la marche des malades observés après traitement est moins bon que celui des bien–portants.

Chapitre 4

Test d'homogénéité

Les principes de l'inférence combinatoire peuvent être utilisés pour les tests de comparaison de plusieurs groupes, regroupés sous le terme de *tests d'homogénéité*.

Pour la comparaison de plusieurs groupes (données structurées), la logique des tests d'homogénéité est immédiate, elle repose sur le principe *d'échangeabilité* : dire que deux groupes sont homogènes c'est dire que chaque valeur observée aurait pu appartenir à l'un ou l'autre des deux groupes.

Nous nous intéressons ici à un ensemble d'individus statistiques répartis dans plusieurs groupes et plus précisément à une partition de cet ensemble appelée *emboîtement* (défini de façon formelle au paragraphe suivant). Le principe d'échangeabilité conduit à construire l'*espace des emboîtements possibles*, engendré par un *ensemble de permutations* associé à la structure de l'emboîtement observé. Il s'agit ensuite de situer le protocole observé par rapport à l'espace des emboîtements, selon la statistique de test choisie. Ainsi, si, en unidimensionnel, on étudie l'homogénéité de deux groupes indépendants d'effectifs n_1 et n_2, un espace de $\binom{n_1+n_2}{n_1} = \frac{(n_1+n_2)!}{n_1!\,n_2!}$ emboîtements possibles est construit en échangeant les observations entre les deux groupes. Nous pouvons alors prendre (c'est ce qui est fait dans ce chapitre) la statistique *Différence des moyennes* comme statistique de test. De même, si l'on considère une partition d'une population en C groupes (ou classes) que nous souhaitons comparer, l'espace des emboîtements est construit en échangeant les observations entre les groupes. Nous pouvons prendre alors la statistique *Variance*[1] comme statistique de test. Dans la mesure où les observations sont *échangées*, les tests d'homogénéité appartiennent à la famille des tests de permutation.

Dans la suite, nous nous plaçons dans le cas multidimensionnel, nous étudions d'abord l'homogénéité de plusieurs groupes, puis nous traitons en détail le cas particulier de l'homogénéité de deux groupes. Nous comparons ensuite les résultats du test avec ceux obtenus par d'autres procédures inférentielles basées d'une part, de façon traditionnelle, sur le modèle normal et d'autre part, sur une procédure de rééchantillonnage bootstrap. Enfin, nous appliquons le test d'homogénéité, sans perte de généralité, à des nuages à deux dimensions.

Considérations préliminaires. En analyse de la variance, on considère souvent un ensemble $I = \{1, \ldots, i, \ldots, N\}$ d'individus (ou de sujets) et un emboîtement de I dans un

1. Le choix de cette statistique de test conduit à des propriétés intéressantes permettant notamment la construction d'une zone de compatibilité dans le cas de la comparaison de deux groupes (*cf.* paragraphe 2.1.2, p.100).

facteur, noté A, à A modalités [2]. Du point de vue de la formalisation ensembliste, on a donc une surjection $f : I \to A$. Notons $I<a>$ l'ensemble $f^{-1}(a) \subset I$, c'est–à–dire l'ensemble des individus possédant la modalité a (ou appartenant au groupe a). On dit qu'un emboîtement est de type n_A si l'effectif du groupe a est égal à n_a ($a \in A$). Les groupes $(a)_{a\in A}$ étudiés ici sont appelés *groupes indépendants*.

Dans la suite, nous nous intéressons à l'homogénéité de C′ *groupes d'intérêt*. Ils peuvent être :

- des groupes *élémentaires*, correspondants aux groupes $a1$, $a2$, ..., c'est–à–dire aux groupes définis par les modalités $a1$, $a2$, ... de A,
- des groupes *composites*, correspondants à des regroupements de modalités, par exemple le groupe $a1$–$a2$ constitué de la réunion des groupes $a1$ et $a2$.

Notons C l'ensemble composé :
- des C′ groupes d'intérêt,
- du groupe c_r consistant en la réunion des groupes restants.

L'ensemble C définit une partition de I en C groupes.

1 Homogénéité de plusieurs groupes indépendants

Dans cette section, nous nous proposons d'étudier l'homogénéité de plusieurs groupes indépendants.

Situation de base. Considérons un ensemble $I = \{1, \ldots, i, \ldots, N\}$ de N individus et un emboîtement de I dans C groupes (ou classes). L'ensemble indexant les groupes est noté $C = \{1, \ldots, c, \ldots, \mathrm{C}\}$, chaque groupe $c \in C$ est d'effectif n_c. A l'ensemble I est associé le nuage M^I de N points (individus) à valeurs dans un espace affine euclidien $\mathcal{U}$ de dimension K. Le point moyen du nuage M^I est noté O, sa structure de covariance est définie par l'endomorphisme de covariance *Som* (défini au chapitre 1, p.16). De même, à l'ensemble des individus $I<c>$ appartenant au groupe c est associé le sous–nuage $\mathrm{M}^{I<c>} = (\mathrm{M}^i)_{i\in I<c>}$ de n_c points, admettant G^c pour point moyen.

Dans la suite, nous étudions l'homogénéité de C′ groupes d'intérêt (C′ ≥ 2). Posons :

$$N' = \sum_{c\in C'} n_c \quad (C' \subseteq C)$$

Si les groupes d'intérêt forment une partition de I, alors C′ = C et $N' = N$, sinon C′ = C − 1.

Lorsque C′ = C, nous comparons tous les groupes, nous effectuons alors une *comparaison globale*. Par opposition, lorsque C′ = C − 1, nous effectuons une *comparaison partielle*. Si maintenant, nous comparons C′ groupes, avec C′ = C − 1, en ne considérant que les données de ces C′ groupes, nous effectuons une *comparaison spécifique*.

Dans la suite, nous considérons que les points du nuage M^I *sont à valeurs dans* $\mathcal{M}$*, son support affine. Le sous–espace vectoriel directeur de* $\mathcal{M}$*, de dimension* L*, est noté* $\mathcal{L}$ *(cf. chapitre 1, p.12).*

2. La plupart du temps dans ce chapitre, l'ensemble et son cardinal sont identifiés par la même lettre, l'ensemble en italique et son cardinal en lettre droite.

1.1 Test exact

1.1.1 Principe du test

Nous nous posons la question suivante : *les* C′ *groupes sont-ils hétérogènes*[3] *?*

Pour répondre à cette question, nous construisons le test d'homogénéité de la façon suivante :

1. Considérons d'abord l'ensemble des $\mathrm{J} = \frac{N!}{\prod_{c\in C} n_c!}$ affectations possibles des N individus dans C groupes d'effectifs $(n_c)_{c\in C}$, c'est–à–dire l'ensemble des J emboîtements de type $(n_c)_{c\in C}$ possibles.

   Remarques :
   – Il s'agit en fait de considérer toutes les affectations possibles des N individus dans les C′ groupes d'intérêt d'effectifs $(n_c)_{c\in C'}$ et dans le groupe restant d'effectif $N - N'$. Si $\mathrm{C}' = \mathrm{C}$ (comparaison globale), le groupe restant c_r est vide.
   – Les J emboîtements sont obtenus en effectuant toutes les *permutations* possibles entre les individus des C groupes.

   Soit $J = \{1, \ldots, j, \ldots, \mathrm{J}\}$ l'ensemble indexant les J emboîtements, notons $I < cj >$ l'ensemble des individus appartenant au groupe d'effectif n_c du j–ème emboîtement et $\mathrm{M}^{I<cj>}$ le sous–nuage qui lui est associé[4].
   L'ensemble des $\mathrm{J} = \frac{N!}{\prod_{c\in C} n_c!}$ emboîtements possibles est appelé *ensemble des emboîtements*.
   Pour un j donné, les nuages $(\mathrm{M}^{I<cj>})_{c\in C}$ forment une partition du nuage M^I. L'espace des C × J sous–nuages $\mathrm{M}^{I<cj>}$, muni de cette propriété, est appelé *espace des emboîtements* et noté $\mathcal{J}$.

2. (a) Considérons maintenant l'application $H : \mathcal{J} \to \mathcal{M}$ qui au sous–nuage $\mathrm{M}^{I<cj>}$ associe son point moyen H^{cj} :

$$\begin{array}{rccc} H: & \mathcal{J} & \to & \mathcal{M} \\ & \mathrm{M}^{I<cj>} & \mapsto & \mathrm{H}^{cj} = \sum\limits_{i\in I<cj>} \mathrm{M}^i/n_c \end{array}$$

   Le point moyen H^{cj} a pour poids n_c.

   (b) Faisons ensuite choix d'une statistique de test, ici la statistique notée V_{M} définie de la façon suivante : à l'emboîtement j est associé la moyenne, pondérée par n_c, des carrés des M–distances[5] entre les C′ points moyens $(\mathrm{H}^{cj}, n_c)_{c\in C'}$ et leur point moyen $\mathrm{O}^j = \sum_{c\in C'} \frac{n_c}{N'} \mathrm{H}^{cj}$:

$$\begin{array}{rccc} V_{\mathrm{M}}: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{H}^{Cj} & \mapsto & \sum\limits_{c\in C'} \frac{n_c}{N'} \, |\overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}}|^2 \end{array}$$

3. En analyse de la variance, on dit qu'on a une comparaison à $\mathrm{C}' - 1$ degrés de liberté. Cette comparaison peut être décomposée en $\mathrm{C}' - 1$ comparaisons à 1 degré de liberté associées à $\mathrm{C}' - 1$ contrastes orthogonaux.

4. Dans la notation $I < cj >$, le nom et la place des indices sont signifiants : le premier (c) indique le groupe et le second (j) le numéro de l'emboîtement.

5. Distance de Mahalanobis attachée au nuage M^I (*cf.* chapitre 1, p.18).

Remarques :
- Si C′ = C, alors on a :
$$\forall j \in J,\ \mathrm{O}^j = \mathrm{O}$$
- Pour l'emboîtement j, la statistique de test V_{M} est la variance du nuage des C′ points $(\mathrm{H}^{cj})_{c\in C'}$, obtenue en prenant comme distance entre les points la M–distance, c'est pourquoi nous l'appelons aussi M–*Variance*.
- Pour l'emboîtement j, la statistique de test V_{M} peut également s'écrire en fonction des M–distances entre les couples de points du nuage $\mathrm{H}^{C'j} = (\mathrm{H}^{cj})_{c\in C'}$.

$$\begin{array}{rccl} V_{\mathrm{M}} : & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{H}^{Cj} & \mapsto & \frac{1}{2}\sum\limits_{c\in C'}\sum\limits_{c'\in C'} \frac{n_c n_{c'}}{N'^2}\ |\overrightarrow{\mathrm{H}^{cj}\mathrm{H}^{c'j}}|^2 \end{array}$$

Cette dernière expression de V_{M} montre que le point O^j n'intervient pas directement, seules interviennent les M–distances entre les points moyens pris deux à deux.

Notons $v_{\mathrm{M}\,obs}$ la valeur observée de cette statistique :
$$v_{\mathrm{M}\,obs} = \frac{1}{2}\sum_{c\in C'}\sum_{c'\in C'} \frac{n_c n_{c'}}{N'^2}\ |\overrightarrow{\mathrm{G}^c\mathrm{G}^{c'}}|^2$$
où les points $(\mathrm{G}^c)_{c\in C'}$ sont les points moyens des groupes dont nous souhaitons étudier l'homogénéité.

3. Déterminons enfin la proportion des emboîtements pour lesquels la valeur de la statistique V_{M} est supérieure ou égale à la valeur observée $v_{\mathrm{M}\,obs}$. Cette proportion définit le seuil observé du test exact (seuil exact ou p–value exacte) et est notée p_{obs} :
$$p_{obs} = p(V_{\mathrm{M}} \geq v_{\mathrm{M}\,obs})$$

Nous pouvons désormais énoncer la conclusion du test en termes d'hétérogénéité des groupes.

Si $p_{obs} \leq \alpha$ (seuil $\alpha < 1/2$ fixé [6]), le résultat du test est significatif au seuil α. Pour la statistique M–*Variance*, nous pouvons dire que les groupes sont hétérogènes au seuil α.

Si $p_{obs} > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique M–*Variance*, nous ne pouvons pas dire que les groupes sont hétérogènes au seuil α.

Remarque : A partir de la distribution de la statistique de test V_{M}, nous pouvons définir la valeur critique v_{M_α} au seuil α telle que :
$$p(V_{\mathrm{M}} \geq v_{\mathrm{M}_\alpha}) = \alpha$$

Le résultat du test s'énonce alors comme ceci :
- si $v_{\mathrm{M}\,obs} \geq v_{\mathrm{M}_\alpha}$, alors le résultat du test est significatif au seuil α,
- si $v_{\mathrm{M}\,obs} < v_{\mathrm{M}_\alpha}$, alors le résultat du test n'est pas significatif au seuil α.

6. On utilise souvent les seuils conventionnels $\alpha = .05$ et $\alpha = .01$.

Exemple « Homogénéité » : Afin d'illustrer les notions inhérentes au test d'homogénéité, nous utilisons l'exemple suivant comme fil conducteur.

Considérons l'ensemble de $N = 8$ individus : $I = \{i1, \ldots, i8\}$ et la partition en A = 4 groupes suivante : $I{<}a1{>} = \{i1, i2, i3\}$, $I{<}a2{>} = \{i4, i5\}$, $I{<}a3{>} = \{i6, i7\}$ et $I{<}a4{>} = \{i8\}$. On a donc $n_1 = 3$, $n_2 = 2$, $n_3 = 2$ et $n_4 = 1$. A l'ensemble I est associé le nuage M^I : nuage plan ($L = 2$) de 8 points, de point moyen O. Le nuage M^I et sa partition en 4 sous–nuages sont représentés sur la figure 4.1 ci–après.

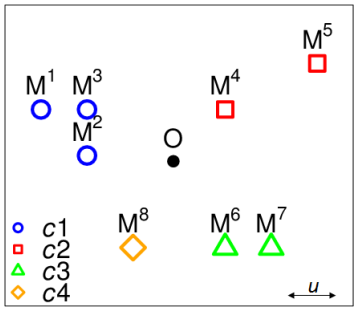
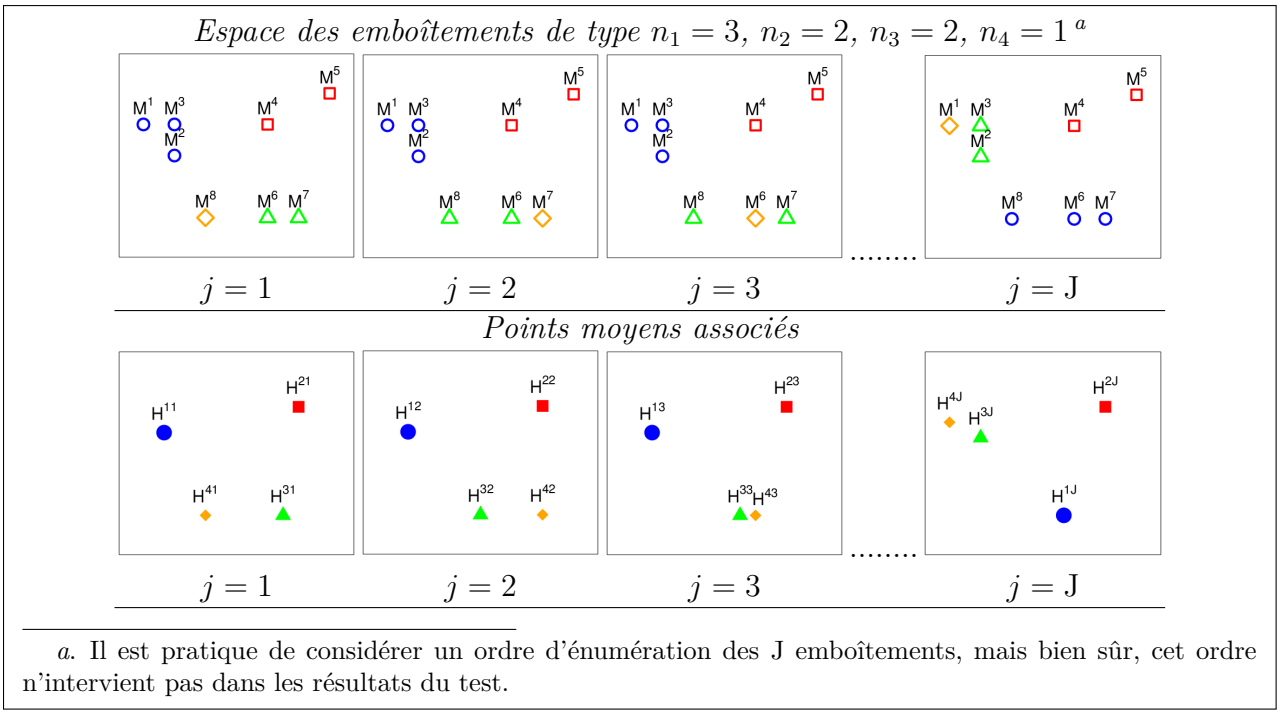


FIGURE 4.1 – *Exemple « Homogénéité ».* Partition du nuage M^I en 4 sous–nuages. Pour les calculs numériques, nous prenons u comme unité de longueur.

• *Cas 1* : nous nous intéressons à l'homogénéité des $\mathrm{C}' = 4$ groupes d'intérêt $a1$, $a2$, $a3$ et $a4$. On a ici $C = \{c1, c2, c3, c4\}$ avec $c1 = a1$, $c2 = a2$, $c3 = a3$ et $c4 = a4$ et donc C = 4. Les groupes d'intérêt forment une partition de I ($\mathrm{C}' = \mathrm{C}$) : nous effectuons une *comparaison globale.*

Nous construisons d'abord l'ensemble des $\mathrm{J} = \frac{N!}{n_1!\times n_2!\times n_3!\times n_4!} = \frac{8!}{3!\times 2!\times 2!\times 1!} = 1680$ emboîtements possibles de type $n_1 = 3$, $n_2 = 2$, $n_3 = 2$ et $n_4 = 1$ et l'espace des emboîtements associé. La figure suivante résume la construction de l'espace des emboîtements et des points moyens $(\mathrm{H}^{1j}, \mathrm{H}^{2j}, \mathrm{H}^{3j}, \mathrm{H}^{4j})_{j\in J}$:



a. Il est pratique de considérer un ordre d'énumération des J emboîtements, mais bien sûr, cet ordre n'intervient pas dans les résultats du test.

FIGURE 4.2 – *Exemple « Homogénéité ».* Cas 1 : comparaison globale (C′ = C). Construction de l'espace des emboîtements et des points moyens associés.

Puis, pour l'emboîtement j, nous calculons la statistique de test V_{M} : moyenne pondérée des carrés des M–distances entre les 4 points moyens H^{1j}, H^{2j}, H^{3j} et H^{4j} (munis des poids n_1, n_2, n_3 et n_4) et le point moyen O. La figure suivante permet d'illustrer ce calcul :

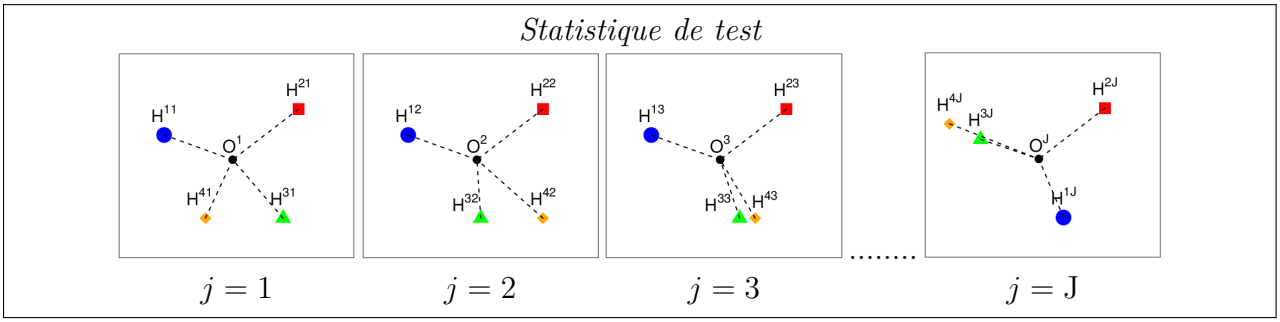


FIGURE 4.3 – *Exemple « Homogénéité ».* Cas 1 : comparaison globale (C′ = C). Construction de la statistique de test. Pour tout emboîtement j de J, on a $\mathrm{O}^j = \mathrm{O}$ (*cf.* p.86).

Seuil observé du test[7] *:* Sur la figure 4.4 est représentée la distribution de la statistique de test V_{M} et sa valeur observée $v_{\mathrm{M}\,obs}$; on a $p_{obs} = p(V_{\mathrm{M}} \geq v_{\mathrm{M}\,obs}) = 2/1680 = .00119$.

7. Tous les calculs ont été effectués en choisissant deux axes perpendiculaires, l'un horizontal (x_1), l'autre vertical (x_2), gradués selon l'unité de distance u (*cf.* figure 4.1, p.87) et dont l'origine est O, point moyen du nuage M^I. Ci–après sont reportés le tableau de nombres associés aux 8 points, la matrice de covariance du nuage M^I et les coordonnées des points G^1, G^2, G^3, G^4 :

	$c1$			$c2$		$c3$		$c4$
	M^1	M^2	M^3	M^4	M^5	M^6	M^7	M^8
x_1	1	2	2	5	7	5	6	3
x_2	4	3	4	4	5	1	1	1

$$\mathbf{V} = \begin{pmatrix} 4.109 & -0.266 \\ -0.266 & 2.359 \end{pmatrix},$$

$$\mathrm{G}^1 \begin{pmatrix} 5/3 \\ 11/3 \end{pmatrix}, \quad \mathrm{G}^2 \begin{pmatrix} 6 \\ 9/2 \end{pmatrix}, \quad \mathrm{G}^3 \begin{pmatrix} 11/2 \\ 1 \end{pmatrix}, \quad \mathrm{G}^4 \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

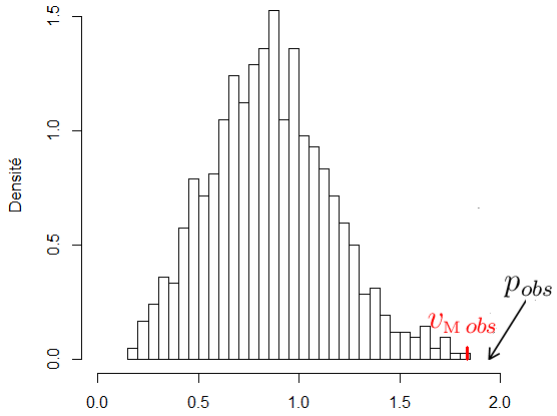


FIGURE 4.4 – *Exemple « Homogénéité ».* Cas 1. Distribution de la statistique V_{M} ($v_{\mathrm{M}\,obs} = 1.836$).

On a $p_{obs} \leq .05$, résultat significatif au seuil .05. Pour la statistique M–*Variance*, nous pouvons dire que les quatre groupes sont hétérogènes au seuil .05.

• *Cas 2* : nous nous intéressons maintenant à l'homogénéité des $\mathrm{C}' = 2$ groupes d'intérêt $a1$ et $a2$. On a ici $C = \{c1, c2, c_r\}$ avec $c1 = a1$, $c2 = a2$ et $c_r = a3$–$a4$ (réunion des groupes $a3$ et $a4$) et donc $\mathrm{C} = 3$. Les groupes d'intérêt ne forment pas une partition de I ($\mathrm{C}' = \mathrm{C}-1$) : nous effectuons une *comparaison partielle*.

Nous construisons d'abord l'ensemble des $\mathrm{J} = \frac{N!}{n_1!\times n_2!\times(N-(n_1+n_2))!} = \frac{8!}{3!\times 2!\times 3!} = 560$ emboîtements possibles de type $n_1 = 3$, $n_2 = 2$ et $n_3 + n_4 = 3$.

Nous nous intéressons ensuite aux points moyens $(\mathrm{H}^{1j})_{\ j\in J}$ et $(\mathrm{H}^{2j})_{\ j\in J}$ des groupes d'effectifs n_1 et n_2 associés à ces emboîtements, leur construction est résumée sur la figure suivante.

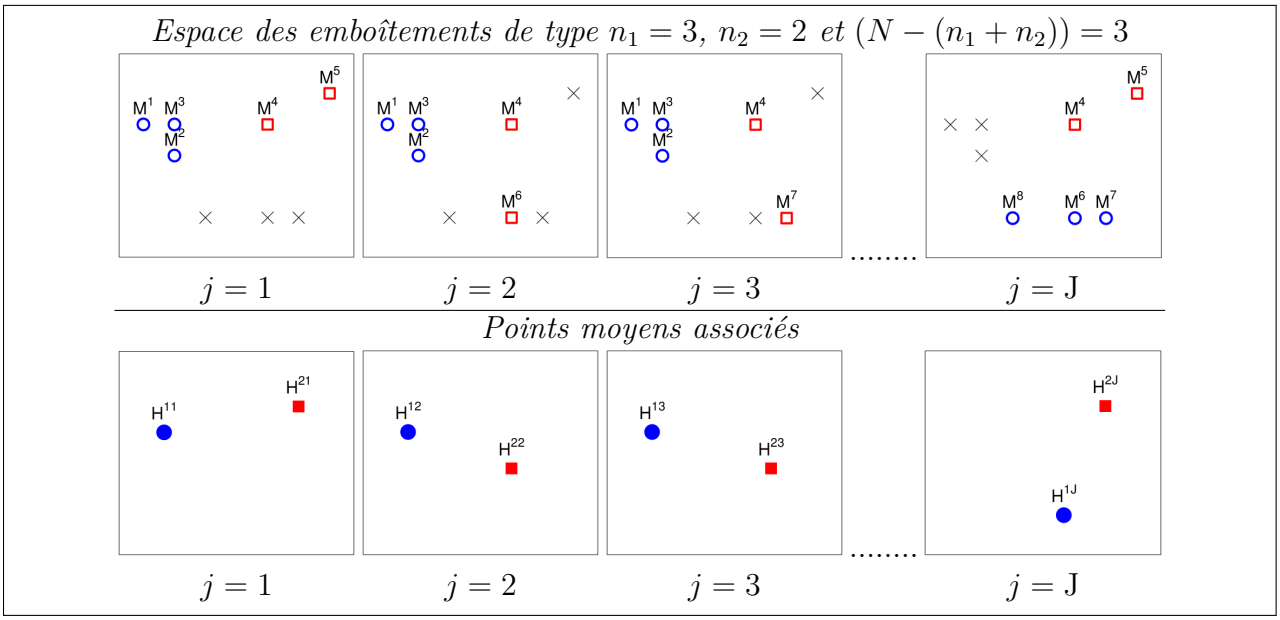


FIGURE 4.5 – *Exemple « Homogénéité ».* Cas 2 : comparaison partielle ($\mathrm{C}' = \mathrm{C}-1$). Construction de l'espace des emboîtements et des points moyens associés.

Pour l'emboîtement j, la statistique de test V_{M} est la moyenne pondérée des carrés des M–distances entre les 2 points moyens H^{1j} et H^{2j} (munis des poids n_1 et n_2) et leur point moyen O^j, la figure suivante permet d'illustrer sa construction :

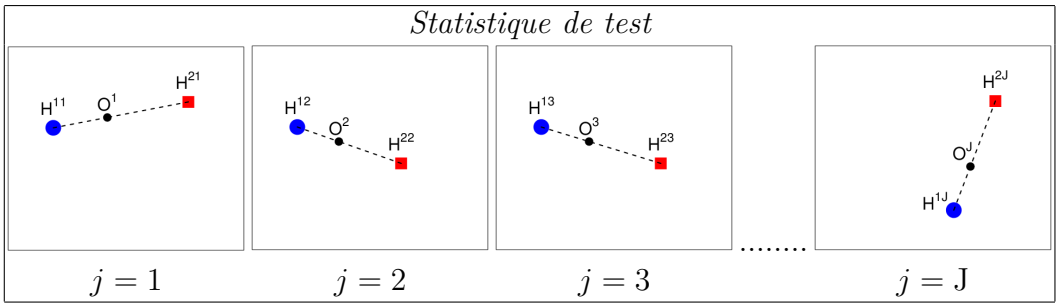


FIGURE 4.6 – *Exemple « Homogénéité ».* Cas 2 : comparaison partielle ($C' = C - 1$). Construction de la statistique de test.

Le test correspondant est mis en oeuvre dans la 2ème section de ce chapitre (p.98).

Caractéristiques des C nuages $(H^{cJ})_{c\in C}$ et de la distribution de la statistique de test V_M.

Propriété 1.1. *Pour tout $c \in C$, le point moyen du nuage H^{cJ} est le point moyen O du nuage M^I :*

$$\sum_{j\in J} H^{cj} = O$$

Démonstration.
$\forall cj \in C \times J$, posons $\varepsilon_{ic}^{j} = 1$ si l'individu i appartient au groupe $I{<}cj{>}$ (groupe d'effectif n_c du j–ème emboîtement) et $\varepsilon_{ic}^{j} = 0$, sinon.

Pour l'emboîtement j, le point moyen H^{cj} du sous–nuage $M^{I<cj>}$ de n_c points est tel que $\overrightarrow{OH^{cj}} = \frac{1}{n_c} \sum_{i\in I<cj>} \overrightarrow{OM^i} = \frac{1}{n_c} \sum_{i\in I} \varepsilon_{ic}^{j} \overrightarrow{OM^i}$ $(c \in C)$.

Considérons maintenant, pour un c donné, le nuage H^{cJ} des J points moyens $(H^{cj})_{j\in J}$. Le point moyen de ce nuage, noté $\overline{H^c}$, est tel que :

$$\begin{aligned}
\overrightarrow{O\overline{H}^c} &= \frac{1}{J} \sum_{j\in J} \overrightarrow{OH^{cj}} \\
&= \frac{1}{J} \sum_{j\in J} \frac{1}{n_c} \sum_{i\in I} \varepsilon_{ic}^{j} \overrightarrow{OM}^i \\
&= \frac{1}{J} \times \frac{1}{n_c} \sum_{i\in I} \Big(\sum_{j\in J} \varepsilon_{ic}^{j} \Big) \overrightarrow{OM}^i
\end{aligned}$$

Or, pour c et i fixés, le nombre d'emboîtements j pour lesquels l'individu i appartient au groupe $I{<}cj{>}$ est égal au nombre d'emboîtements possibles de $N-1$ individus dans $C-1$ groupes d'effectifs $n_{c'}$ $(c' \neq c)$ et un groupe d'effectif $(n_c - 1)$. D'où :

$$\sum_{j\in J} \varepsilon_{ic}^{j} = \frac{(N-1)!}{(n_c - 1)! \times \prod\limits_{\substack{c'\in C \\ c'\neq c}} n_{c'}!} = \frac{N!}{\prod\limits_{c\in C} n_c!} \times \frac{n_c}{N} = J \times \frac{n_c}{N}$$

Et on a par conséquent :

$$\begin{aligned}
\overrightarrow{O\overline{H}^c} &= \frac{1}{J} \times \frac{1}{n_c} \sum_{i\in I} (J \times \tfrac{n_c}{N}) \overrightarrow{OM}^i \\
&= \frac{1}{N} \sum_{i\in I} \overrightarrow{OM}^i = \overrightarrow{0} \quad \text{(d'après la caractérisation barycentrique du point moyen (\textit{cf.} p.13),}
\end{aligned}$$

O étant le point moyen du nuage M^I).
c.q.f.d.

Pour un emboîtement j donné, notons Som^j l'endomorphisme de covariance associé au nuage $(\mathrm{H}^{cj}, n_c)_{c\in C'}$ de C′ points. Soit $\overline{Som}$ la moyenne des J endomorphismes Som^j :

$$\overline{Som} = \frac{1}{\mathrm{J}} \sum_{j\in J} Som^j$$

Propriété 1.2. *$\overline{Som}$ ne dépend que de Som et est égal à :*

$$Som \times \frac{\mathrm{C}'-1}{N-1} \times \frac{N}{N'}$$

En particulier, lorsque $\mathrm{C}' = \mathrm{C}$ *:* $\overline{Som} = Som \times \frac{\mathrm{C}-1}{N-1}$.

Démonstration. D'après la propriété 3.2 (p.16), on a :

$$\begin{array}{rcl} \forall \overrightarrow{\alpha} \in \mathcal{L},\quad Som^j(\overrightarrow{\alpha}) & = & \sum\limits_{c\in C'} \frac{n_c}{N'} \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}} \\ & = & Som^j_{\mathrm{O}}(\overrightarrow{\alpha}) - \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} \end{array}$$

avec O^j le point moyen du nuage $(\mathrm{H}^{cj}, n_c)_{c\in C'}$ de C′ points du j–ème emboîtement.

D'où :

$$\begin{array}{rcl} \overline{Som}(\overrightarrow{\alpha}) & = & \frac{1}{\mathrm{J}} \sum\limits_{j\in J} Som^j_{\mathrm{O}}(\overrightarrow{\alpha}) - \frac{1}{\mathrm{J}} \sum\limits_{j\in J} \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} \\ & = & \frac{1}{\mathrm{J}} \sum\limits_{j\in J} \sum\limits_{c\in C'} \frac{n_c}{N'} \langle \overrightarrow{\mathrm{OH}^{cj}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OH}^{cj}} - \frac{1}{\mathrm{J}} \sum\limits_{j\in J} \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} \\ & = & \sum\limits_{c\in C'} \frac{n_c}{N'} \Big(\frac{1}{\mathrm{J}} \sum\limits_{j\in J} \langle \overrightarrow{\mathrm{OH}^{cj}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OH}^{cj}} \Big) - \frac{1}{\mathrm{J}} \sum\limits_{j\in J} \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} \\ & = & \sum\limits_{c\in C'} \frac{n_c}{N'} Som^c(\overrightarrow{\alpha}) - \frac{1}{\mathrm{J}} \sum\limits_{j\in J} \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} \end{array}$$

où Som^c est l'endomorphisme de covariance du nuage des J points $(\mathrm{H}^{cj},1)_{j\in J}$

Pour un c donné, le nombre d'emboîtements de J pour lesquels le groupe d'effectif n_c est constitué du même n_c–uplet d'individus est égal au nombre d'emboîtements possibles de $N - n_c$ individus dans $\mathrm{C}-1$ groupes d'effectifs $n_{c'}$ ($c' \neq c$). En reprenant les notations de la démonstration de la propriété 1.1 (p.90), on a :

$$\sum_{j\in J} \Big(\prod_{i\in I<cj>} \varepsilon^j_{ic} \Big) = \frac{(N-n_c)!}{\prod\limits_{\substack{c'\in C \\ c'\neq c}} n_{c'}!} = \frac{N!}{\prod\limits_{c\in C} n_c!} \times \frac{(N-n_c)!\; n_c!}{N!} = \frac{\mathrm{J}}{\binom{N}{n_c}}$$

Si on considère les J groupes d'effectif n_c des J emboîtements, le même n_c–uplet d'individus est présent $\frac{\mathrm{J}}{\binom{N}{n_c}}$ fois, il est donc équivalent d'étudier le nuage élémentaire de J points $(\mathrm{H}^{cj},1)_{j\in J}$ ou le nuage équipondéré de $\mathrm{J}' = \binom{N}{n_c}$ points $(\mathrm{H}^{cj'}, \frac{\mathrm{J}}{\binom{N}{n_c}})_{j'\in J'}$ où $J' = \{1,\ldots,j',\ldots,\binom{N}{n_c}\}$. En conséquence, Som^c est aussi l'endomorphisme de covariance associé au nuage de J′ points $(\mathrm{H}^{cj'}, \frac{\mathrm{J}}{\binom{N}{n_c}})_{j'\in J'}$, et celui des points moyens de l'échantillonnage de $\binom{N}{n_c}$ (*cf.* chapitre 3).

D'après la propriété 1.2 (p.57), $Som^c = \frac{N-n_c}{N-1} \times \frac{Som}{n_c}$ où Som est l'endomorphisme de covariance associé au nuage M^I.

De même, si on considère les J groupes d'effectif N' (réunion des C′ groupes d'effectif $(n_c)_{c\in C'}$) des J emboîtements, le même N'–uplet d'individus est présent $\frac{\mathrm{J}}{\binom{N}{N'}}$ fois. Par un raisonnement analogue, l'endomorphisme de covariance du nuage des J points O^j s'exprime en fonction de Som par :

$$\frac{1}{\mathrm{J}} \sum_{j\in J} \langle \overrightarrow{\mathrm{OO}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OO}^j} = \frac{N-N'}{N-1} \times \frac{Som(\overrightarrow{\alpha})}{N'}$$

On a donc :
$\overline{Som} = \frac{1}{N'} \sum_{c \in C'} n_c \times (\frac{N-n_c}{N-1} \times \frac{Som}{n_c}) - \frac{N-N'}{N-1} \times \frac{Som}{N'}$
$= \frac{1}{N'} \times \frac{Som}{N-1} \times \sum_{c \in C'} (N - n_c) - \frac{N-N'}{N-1} \times \frac{Som}{N'}$
$= \frac{1}{N'} \times \frac{Som}{N-1} \times (\mathrm{C}'N - N' - N + N')$
$= Som \times \frac{\mathrm{C}'-1}{N-1} \times \frac{N}{N'}$
c.q.f.d.

Faisons maintenant choix d'un repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$ de $\mathcal{M}$. Pour un emboîtement j donné, notons $\mathbf{B}^j$ la matrice de covariance du nuage $(\mathrm{H}^{cj}, n_c)_{c \in C'}$ de C′ points. Soit $\overline{\mathbf{B}}$ la moyenne des J matrices de covariance $\mathbf{B}^j$: $\overline{\mathbf{B}} = \sum_{j \in J} \mathbf{B}^j / \mathrm{J}$

Corollaire 1.1. *Dans le repère orthonormé* $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$ *de* $\mathcal{M}$, $\overline{\mathbf{B}}$ *ne dépend que de la matrice de covariance* $\mathbf{V}$ *du nuage* M^I *:*

$$\overline{\mathbf{B}} = \mathbf{V} \times \frac{\mathrm{C}' - 1}{N - 1} \times \frac{N}{N'}$$

En particulier, lorsque $\mathrm{C}' = \mathrm{C}$ *:* $\overline{\mathbf{B}} = \mathbf{V} \times \frac{\mathrm{C}-1}{N-1}$.

Propriété 1.3. *La moyenne de la statistique* M*–Variance est :*

$$\mathrm{Moy}(V_{\mathrm{M}}) = L \times \frac{\mathrm{C}' - 1}{N - 1} \times \frac{N}{N'}$$

En particulier, lorsque $\mathrm{C}' = \mathrm{C}$ *:* $\mathrm{Moy}(V_{\mathrm{M}}) = L \times \frac{\mathrm{C}-1}{N-1}$.

Démonstration. Considérons le nuage pondéré des C′ × J points $(\mathrm{H}^{cj}, 1 \times n_c)_{cj \in C' \times J}$.
A l'emboîtement canonique $(C' \times J) < C' >$ de $C' \times J$ dans C' sont associés le nuage inter (équipondéré) des J points (O^j, N') et le nuage intra $(\mathrm{H}^{CJ}_{intra}, n_c)_{cj \in C' \times J}$ défini par :

$$\overrightarrow{\mathrm{OH}^{cj}_{intra}} = \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}}$$

L'endomorphisme de covariance du nuage intra est, par définition, la moyenne des J endomorphismes de covariance Som^j des J nuages $(\mathrm{H}^{C'j})_{j \in J}$, c'est–à–dire ce que nous avons noté $\overline{Som}$ (*cf.* propriété 1.2). D'après la propriété 4.2 (chapitre 1, p.18), la variance du nuage intra calculée en fonction de la distance de Mahalanobis associée à $\overline{Som}$ est égale à L.
On a donc :

$$\sum_{j \in J} \sum_{c \in C'} \frac{n_c}{\mathrm{J} \times N'} \langle \overline{Som}^{-1} (\overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}}) | \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}} \rangle = L$$

Or, d'après la propriété 1.2, $\overline{Som} = Som \times \frac{\mathrm{C}'-1}{N-1} \times \frac{N}{N'}$ donc $\overline{Som}^{-1} = Som^{-1} \times \frac{N-1}{\mathrm{C}'-1} \times \frac{N'}{N}$
d'où $L = \sum_{j \in J} \sum_{c \in C'} \frac{n_c}{\mathrm{J} \times N'} \frac{N-1}{\mathrm{C}'-1} \times \frac{N'}{N} \langle Som^{-1} (\overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}}) | \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}} \rangle = \sum_{j \in J} \sum_{c \in C'} \frac{n_c}{\mathrm{J} \times N'} \frac{N-1}{\mathrm{C}'-1} \times \frac{N'}{N} | \overrightarrow{\mathrm{O}^j\mathrm{H}^{jc}} |^2$
et $\mathrm{Moy}(V_{\mathrm{M}}) = \frac{1}{\mathrm{J}} \sum_{j \in J} \sum_{c \in C'} \frac{n_c}{N'} | \overrightarrow{\mathrm{O}^j\mathrm{H}^{cj}} |^2 = L \times \frac{\mathrm{C}'-1}{N-1} \times \frac{N}{N'}$.
c.q.f.d.

1.1.2 Cas particulier d'un nuage unidimensionnel

Dans le cas d'un nuage unidimensionnel ($L = 1$), à I est associée la variable–coordonnée $x^I = (x^i)_{i\in I}$ de moyenne m et de variance v. Considérons la partition de I en C groupes d'effectifs n_c : les C′ groupes d'intérêt dont nous voulons étudier l'homogénéité et le groupe restant c_r. On a $\mathrm{Moy}(x^{I<c>}) = m^c$ et $N' = \sum_{c\in C'} n_c$ ($C' \subseteq C$). Si les groupes d'intérêt forment une partition de I, alors $\mathrm{C}' = \mathrm{C}$ et $N' = N$, sinon $\mathrm{C}' = \mathrm{C} - 1$.

Remarque : Le test d'homogénéité décrit au paragraphe 1.1.1 (p.85) peut être effectué dans le cas d'un nuage unidimensionnel, cependant nous en donnons ici une version spécifique envisageable uniquement dans ce cas particulier.

Comme pour le cas multidimensionnel (*cf.* p.85), nous construisons l'ensemble des $\mathrm{J} = \frac{N!}{\prod_{c\in C} n_c!}$ emboîtements possibles. Au sous–ensemble $I<cj>$ est associée la valeur moyenne $m^{cj} = \sum_{i\in I<cj>} x^i/n_c$.

Lorsque nous souhaitons étudier l'homogénéité de C′ groupes, la statistique V définie par :

$$\begin{array}{rccl} V: & \mathbb{R} & \to & \mathbb{R}^+ \\ & m^{Cj} & \mapsto & \frac{1}{2}\sum\limits_{c\in C'}\sum\limits_{c'\in C'} \frac{n_c n_{c'}}{N'^2}\,(m^{cj} - m^{c'j})^2 \end{array}$$

peut être prise comme statistique de test. Il s'agit de la variance des moyennes des groupes pondérées par n_c (variance inter), écrite en fonction des différences entre les moyennes prises deux à deux, c'est pourquoi nous l'appelons *Variance*. Sa valeur observée est :

$$v_{obs} = \frac{1}{2}\sum_{c\in C'}\sum_{c'\in C'} \frac{n_c n_{c'}}{N'^2}\,(m^c - m^{c'})^2$$

Seuil observé exact. La proportion des emboîtements pour lesquels la valeur de la statistique V est supérieure ou égale à la valeur observée v_{obs} détermine le seuil observé exact du test : $p_{obs} = p(V \geq v_{obs})$. La conclusion est donnée en termes d'hétérogénéité des groupes, pour la statistique de test *Variance*.

Caractéristiques de la distribution de la statistique *Variance*.

Propriété 1.4. *La moyenne de la statistique Variance est :*

$$\mathrm{Moy}(V) = v \times \frac{\mathrm{C}' - 1}{N - 1} \times \frac{N}{N'}$$

En particulier, lorsque $\mathrm{C}' = \mathrm{C}$ *:* $\mathrm{Moy}(V) = v \times \frac{\mathrm{C}-1}{N-1}$.

Démonstration. Cas particulier de la propriété 1.3 (p.92) avec $V_{\mathrm{M}} = V/v$ et $L = 1$.

1.2 Test approché

1.2.1 Principe du test

Soit $\mathrm{S}^2_{\mathrm{M}}$ la statistique définie par :

$$\mathrm{S}^2_{\mathrm{M}} = \frac{V_{\mathrm{M}}}{\frac{1}{N-1}\frac{N}{N'}}$$

où V_{M} est telle que définie p.85. On a $\mathrm{Moy}(\mathrm{S}^2_{\mathrm{M}}) = L(\mathrm{C}'-1)$. En ajustant la distribution de la statistique $\mathrm{S}^2_{\mathrm{M}}$ par un χ^2 à $L(\mathrm{C}'-1)$ degrés de liberté ($\chi^2_{L(\mathrm{C}'-1)}$), nous pouvons définir un seuil approché.

Seuil observé approché. Soit $s^2_{\mathrm{M}\,obs} = \frac{v_{\mathrm{M}\,obs}}{\frac{1}{N-1}\frac{N}{N'}}$, la proportion $p(\mathrm{S}^2_{\mathrm{M}} \geq s^2_{\mathrm{M}\,obs})$ est approximativement égale à $p(\chi^2_{L(\mathrm{C}'-1)} \geq s^2_{\mathrm{M}\,obs}) = \widetilde{p}_{obs}$, seuil observé approché.

Nous concluons en termes d'hétérogénéité des groupes (*cf.* p.86), pour la statistique $\mathrm{S}^2_{\mathrm{M}}$.

Le test approché est illustré par l'exemple des races canines, p.119.

1.2.2 Cas particulier d'un nuage unidimensionnel

Nous sommes dans la situation décrite au paragraphe 1.1.2 (p.93) où $L = 1$. Le test approché consiste à ajuster la distribution de la statistique $\mathrm{S}^2 = \frac{V}{\frac{v}{N-1}\frac{N}{N'}}$ (*cf.* définition de V p.93) par celle d'un χ^2 à $\mathrm{C}'-1$ degrés de liberté.

Seuil observé approché. Soit $\mathrm{s}^2_{obs} = \frac{v_{obs}}{\frac{v}{N-1}\frac{N}{N'}}$, la proportion $p(\mathrm{S}^2 \geq \mathrm{s}^2_{obs})$ est approximativement égale à $p(\chi^2_{\mathrm{C}'-1} \geq \mathrm{s}^2_{obs}) = \widetilde{p}_{obs}$, seuil observé approché. La conclusion est donnée en termes d'hétérogénéité des groupes pour la statistique S^2.

2 Homogénéité de deux groupes indépendants

Lorsque $\mathrm{C}' = 2$, nous nous intéressons à l'homogénéité des deux groupes d'intérêt [8] $c1$ et $c2$, d'effectifs respectifs n_1 et n_2. Les individus qui n'appartiennent à aucun des deux groupes d'intérêt sont rassemblés dans le groupe c_r : $C = \{c1, c2, c_r\}$. Nous considérons donc en particulier les deux sous–nuages $\mathrm{M}^{I<c1>}$ et $\mathrm{M}^{I<c2>}$ de points moyens respectifs G^1 et G^2.

Cette section est illustrée par le cas 2 de l'exemple « Homogénéité » introduit précédemment (*cf.* p.89) :

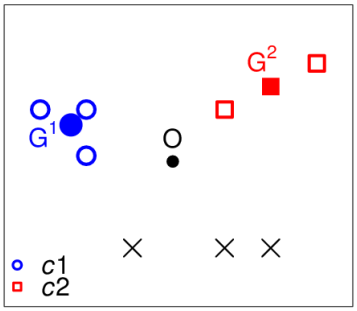


FIGURE 4.7 – *Exemple « Homogénéité ».* Cas 2. Nuage M^I, de point moyen O et sous–nuages $\mathrm{M}^{I<c1>}$ et $\mathrm{M}^{I<c2>}$, de points moyens G^1 et G^2.

8. En analyse de la variance, on dit qu'on a une comparaison à 1 degré de liberté.

2.1 Test Exact

2.1.1 Principe du test

Le principe du test reste inchangé, la construction de l'espace des emboîtements et des points moyens associés est la même. Cependant, nous choisissons ici une statistique de test plus simple (conduisant à un test équivalent) qui permet, d'une part une interprétation géométrique du seuil observé et d'autre part la construction d'une zone de compatibilité.

Statistique de test. Dans le cas de deux groupes, la statistique V_{M} peut s'écrire :

$$\begin{array}{rccl} V_{\text{M}}: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{H}^{1j}, \mathrm{H}^{2j} & \mapsto & \frac{n_1 n_2}{(n_1+n_2)^2}\ |\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}|^2 \end{array}$$

Le seuil observé exact du test est donc la proportion :

$$p_{obs} = p(\frac{n_1 n_2}{(n_1+n_2)^2}\ |\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}|^2 \geq \frac{n_1 n_2}{(n_1+n_2)^2}\ |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|^2)$$

Cette proportion est égale à la proportion

$$p_{obs} = p(|\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}| \geq |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|)$$

Nous pouvons donc utiliser, de façon équivalente, la statistique de test D_{M} suivante :

$$\begin{array}{rccl} D_{\text{M}}: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{H}^{1j}, \mathrm{H}^{2j} & \mapsto & |\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}| \end{array}$$

Comme le montre la figure 4.6 (p.90), la statistique D_{M} dépend uniquement de la M–distance entre les points $(\mathrm{H}^{1j},\ \mathrm{H}^{2j})_{j\in J}$, c'est pourquoi nous l'appelons M–*Distance*.

Construction du nuage des points–écarts. Au bipoint $(\mathrm{H}^{1j}, \mathrm{H}^{2j})_{j\in J}$, nous associons d'abord le vecteur–écart :

$$\overrightarrow{d^j} = \mathrm{H}^{2j} - \mathrm{H}^{1j}$$

Nous construisons ensuite le nuage pertinent D^J, appelé *nuage des points–écarts*, constitué des $\mathrm{J} = \frac{N!}{n_1!\times n_2!\times(N-(n_1+n_2))!}$ points D^j définis par :

$$\mathrm{D}^j = \mathrm{O} + \overrightarrow{d^j}$$

Notons $\mathrm{D}_{obs} = \mathrm{O} + \overrightarrow{d_{obs}}$ avec $\overrightarrow{d_{obs}} = \mathrm{G}^2 - \mathrm{G}^1$.

Nous montrons dans la suite que la construction du nuage des points–écarts permet une interprétation géométrique du seuil observé du test.

Remarque : Le choix de pointer les vecteurs $(\overrightarrow{d^j})_{j\in J}$ en O est arbitraire quoique naturel.

Par construction, la statistique D_{M} peut également s'écrire :

$$\begin{array}{rccl} D_{\text{M}}: & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{D}^j & \mapsto & |\overrightarrow{\mathrm{OD}^j}| \end{array}$$

et nous avons dans ce cas $d_{\mathrm{M}\,obs} = |\overrightarrow{\mathrm{OD}_{obs}}|$ $(= |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|)$ et $p_{obs} = p(D_{\mathrm{M}} \geq d_{\mathrm{M}\,obs})$.

Pour l'exemple précédent, les étapes de la construction du nuage D^J sont représentées sur la figure 4.8.

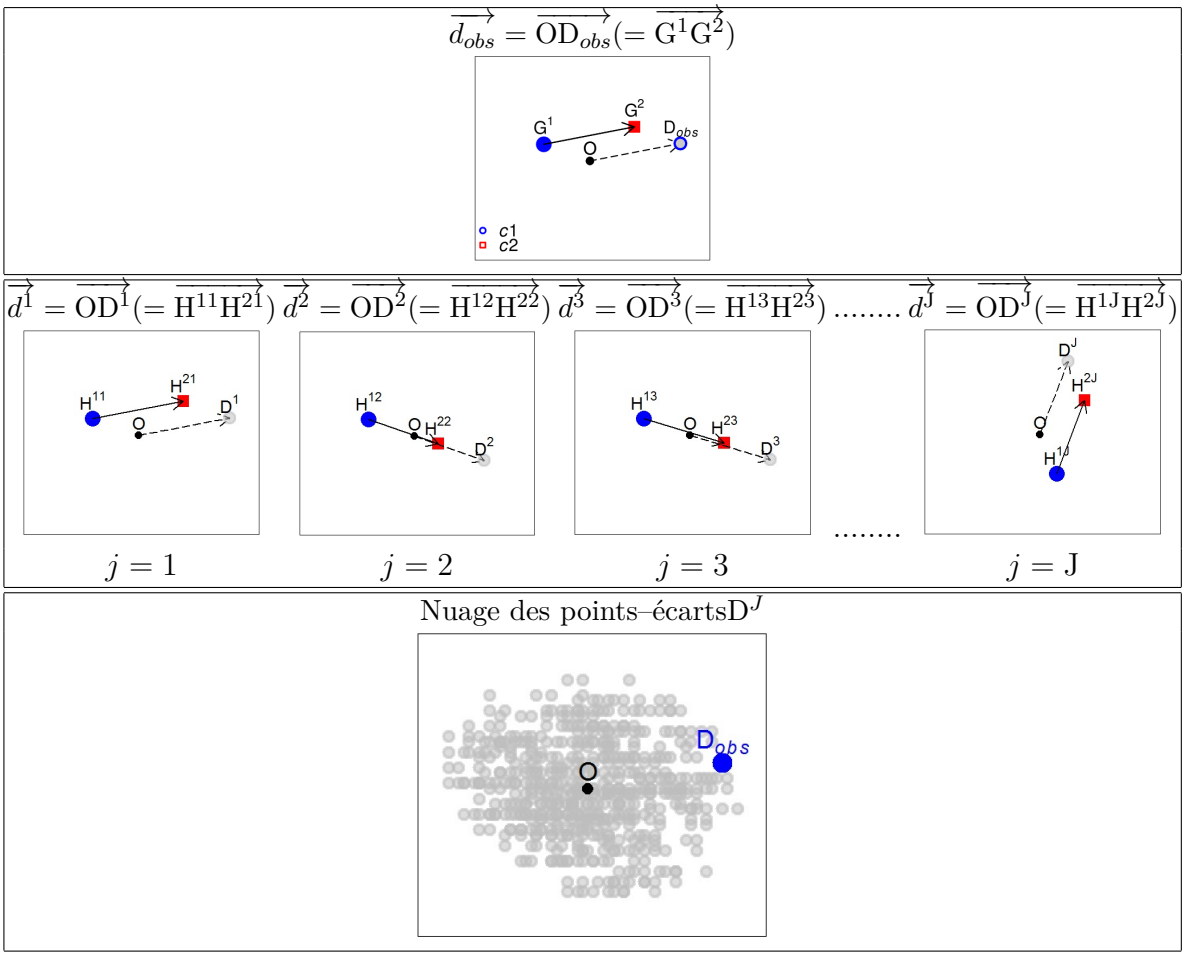


FIGURE 4.8 – *Exemple « Homogénéité »*. Cas 2. Construction du nuage des points–écarts D^J ($\mathrm{J} = \frac{N!}{n_1! \times n_2! \times (N-(n_1+n_2))!} = \frac{8!}{3! \times 2! \times 3!} = 560$ points).

Caractéristiques du nuage des points–écarts D^J et de la statistique de test D.

Propriété 2.1. *Le point moyen du nuage des points–écarts* D^J *est le point* O.

Démonstration. Le point moyen du nuage D^J, noté $\overline{\mathrm{D}}$, est tel que :

$$\begin{aligned}
\overrightarrow{\mathrm{OD}} &= \tfrac{1}{\mathrm{J}} \sum_{j \in J} \overrightarrow{\mathrm{OD}^j} \\
&= \tfrac{1}{\mathrm{J}} \sum_{j \in J} \overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}} \\
&= -\tfrac{1}{\mathrm{J}} \sum_{j \in J} \overrightarrow{\mathrm{OH}^{1j}} + \tfrac{1}{\mathrm{J}} \sum_{j \in J} \overrightarrow{\mathrm{OH}^{2j}} \\
&= \overrightarrow{0} \qquad \Big(\text{d'après la caractérisation barycentrique du point moyen (\textit{cf.} p.1),}
\end{aligned}$$

O étant le point moyen des nuages H^{1J} et H^{2J} (*cf.* propriété 1.1, p.90)$\Big)$.

c.q.f.d.

Soit Som^{D} l'endomorphisme dont les vecteurs propres déterminent les directions principales du nuage des points–écarts.

Propriété 2.2. *L'endomorphisme* Som^{D} *est proportionnel à l'endomorphisme Som associé au nuage* M^I *:*

$$Som^{\mathrm{D}} = \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times Som$$

En particulier, lorsque $\mathrm{C}' = \mathrm{C}$*, on a* $Som^{\mathrm{D}} = \frac{1}{N-1} \times \frac{1}{f_1 f_2} \times Som$*, avec* $f_1 = n_1/N$ *et* $f_2 = n_2/N$*.*

Démonstration.

$$\begin{array}{rcl} \forall \overrightarrow{\alpha} \in \mathcal{L}, \quad Som^{\mathrm{D}}(\overrightarrow{\alpha}) & = & \frac{1}{\mathrm{J}} \sum\limits_{j \in J} \langle \overrightarrow{\mathrm{OD}^j} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{OD}^j} \\ & = & \frac{1}{\mathrm{J}} \sum\limits_{j \in J} \langle \overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}} \end{array}$$

Or O^j est le point moyen des points (H^{1j}, n_1) et (H^{2j}, n_2). On a donc : $\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}} = -\frac{n_1+n_2}{n_2} \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}}$.
D'où :

$$\begin{array}{rcl} Som^{\mathrm{D}}(\overrightarrow{\alpha}) & = & \frac{1}{\mathrm{J}} \sum\limits_{j \in J} \langle -\frac{n_1+n_2}{n_2} \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle (-\frac{n_1+n_2}{n_2} \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}}) \\ & = & (\frac{n_1+n_2}{n_2})^2 \times \frac{1}{\mathrm{J}} \sum\limits_{j \in J} \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} \end{array}$$

Soit Som^j l'endomorphisme de covariance associé au nuage formé des deux points (H^{1j}, n_1) et (H^{2j}, n_2) :

$$\begin{array}{rcl} Som^j(\overrightarrow{\alpha}) & = & \frac{1}{n_1+n_2} (n_1 \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} + n_2 \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{2j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{2j}}) \\ & = & \frac{1}{n_1+n_2} (n_1 \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} + n_2 \langle -\frac{n_1}{n_2} \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle (-\frac{n_1}{n_2} \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}})) \\ & = & \frac{1}{n_1+n_2} (n_1 + \frac{n_1^2}{n_2}) \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} \\ & = & \frac{n_1}{n_2} \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} \end{array}$$

D'après la propriété 1.2 (p.91), on a ici : $\frac{1}{\mathrm{J}} \sum\limits_{j \in J} Som^j(\overrightarrow{\alpha}) = Som(\overrightarrow{\alpha}) \times \frac{1}{N-1} \times \frac{N}{n_1+n_2}$.
On a donc : $\frac{1}{\mathrm{J}} \sum\limits_{j \in J} \langle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} | \overrightarrow{\alpha} \rangle \overrightarrow{\mathrm{O}^j\mathrm{H}^{1j}} = Som(\overrightarrow{\alpha}) \times \frac{1}{N-1} \times \frac{N}{n_1+n_2} \times \frac{n_2}{n_1}$.
Et :

$$\begin{array}{rcl} Som^{\mathrm{D}} & = & (\frac{n_1+n_2}{n_2})^2 \times Som \times \frac{1}{N-1} \times \frac{N}{n_1+n_2} \times \frac{n_2}{n_1} \\ & = & \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times Som \end{array}$$

c.q.f.d.

Dans le repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$ de $\mathcal{M}$, notons $\mathbf{V}^{\mathrm{D}}$ la matrice de covariance du nuage des points–écarts D^J. Rappelons que $\mathbf{V}$ est la matrice de covariance du nuage M^I.

Corollaire 2.1. *On a :*

$$\mathbf{V}^{\mathrm{D}} = \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times \mathbf{V}$$

En effet, dans ce repère orthonormé, la matrice de l'endomorphisme *Som* est $\mathbf{V}$ et celle de l'endomorphisme Som^{D} est $\mathbf{V}^{\mathrm{D}}$, d'où la relation en appliquant la propriété 2.2.

Corollaire 2.2. *Tout ellipsoïde d'inertie du nuage* M^I *est ellipsoïde d'inertie du nuage des points–écarts* D^J*.*

Démonstration. Découle directement de la propriété 2.2.

La figure suivante illustre ces propriétés et corollaires (*Exemple « Homogénéité », cas 2*) :

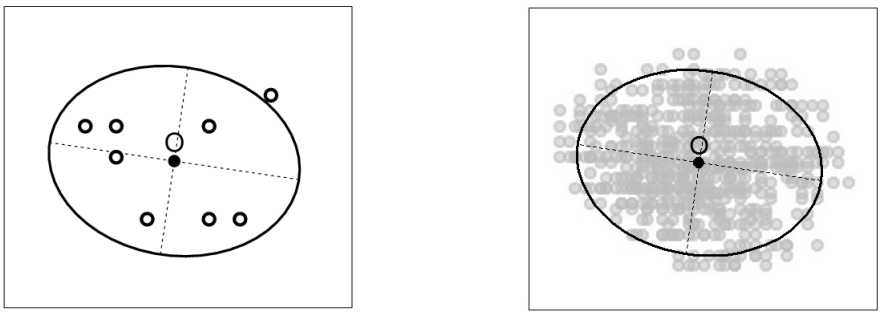


FIGURE 4.9 – *Exemple « Homogénéité »*. Cas 2. Nuage M^I (gauche) et nuage des points–écarts D^J (droite) avec leurs ellipses de concentration ($\kappa = 2$).

Propriété 2.3. *La moyenne du carré de la statistique* M*–Distance est :*

$$\mathrm{Moy}(D^2_{\mathrm{M}}) = L \times \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})$$

Démonstration. D'après la propriété 4.2 (chapitre 1, p.18), la variance du nuage D^J, calculée en fonction de la distance de Mahalanobis associée à l'endomorphisme Som^{D} est égale à L.
On a donc $\frac{1}{\mathrm{J}} \sum_{j\in J} \langle (Som^{\mathrm{D}})^{-1}(\overrightarrow{\mathrm{OD}^j}) | \overrightarrow{\mathrm{OD}^j} \rangle = L$.
Or, d'après la propriété 2.2, $Som^{\mathrm{D}} = Som \times \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})$ donc $(Som^{\mathrm{D}})^{-1} = Som^{-1} \times \frac{N-1}{N} \times \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}}$,
d'où $L = \frac{1}{\mathrm{J}} \sum_{j\in J} \frac{N-1}{N} \times \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}} \langle Som^{-1}(\overrightarrow{\mathrm{OD}^j}) | \overrightarrow{\mathrm{OD}^j} \rangle = \frac{1}{\mathrm{J}} \sum_{j\in J} \frac{N-1}{N} \times \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}} |\overrightarrow{\mathrm{OD}^j}|^2$
et donc $\frac{1}{\mathrm{J}} \sum_{j\in J} |\overrightarrow{\mathrm{OD}^j}|^2 = L \times \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) = \mathrm{Moy}(D^2_{\mathrm{M}})$.
c.q.f.d.

Interprétation géométrique du seuil observé. Dans le cas de l'homogénéité de deux groupes, le seuil observé du test exact est égal à la proportion des points D^j tels que :

$$|\overrightarrow{\mathrm{OD}^j}| \geq |\overrightarrow{\mathrm{OD}_{obs}}|$$

Il s'interprète donc géométriquement comme la proportion des points D^j situés sur ou à l'extérieur de l'ellipsoïde d'inertie du nuage M^I, de centre O et passant par D_{obs}, c'est–à–dire défini par :

$$\kappa = |\overrightarrow{\mathrm{OD}_{obs}}| \ (= |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|)$$

D'après le corollaire 2.2, cet ellipsoïde appartient également à la famille des ellipsoïdes d'inertie du nuage des points–écarts D^J.

Exemple « Homogénéité », cas 2 : Sur la figure 4.10, nous avons représenté la distribution de la statistique de test D_{M} et sa valeur observée $d_{\mathrm{M}\,obs} = 2.258$ [9] ; on a $p_{obs} = p(D_{\mathrm{M}} \geq d_{\mathrm{M}\,obs}) = 17/560 = .03035$. Nous pouvons donner une interprétation géométrique de p_{obs} : parmi les $\frac{8!}{3!\times 2!\times 3!} = 560$ points D^j du nuage des points–écarts, ceux vérifiant $|\overrightarrow{\mathrm{OD}^j}| \geq |\overrightarrow{\mathrm{OD}_{obs}}|$ sont au nombre de 17, ils sont situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I, de centre O et passant par D_{obs} (figure 4.11).

9. $d_{\mathrm{M}\,obs} = |\overrightarrow{\mathrm{OD}_{obs}}| = \sqrt{\begin{pmatrix} 6-5/3 & 9/2-11/3 \end{pmatrix} \times \begin{pmatrix} 4.109 & -0.266 \\ -0.266 & 2.359 \end{pmatrix}^{-1} \times \begin{pmatrix} 6-5/3 \\ 9/2-11/3 \end{pmatrix}} = 2.258$

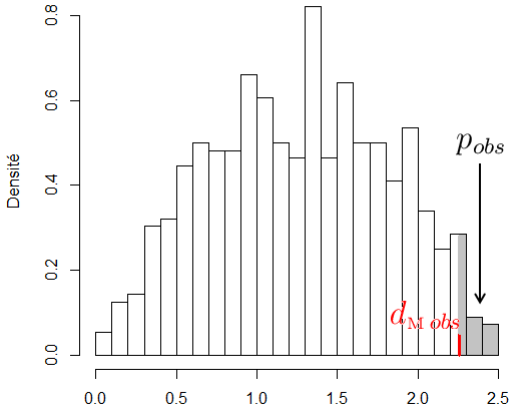


FIGURE 4.10 – *Exemple « Homogénéité »*. Cas 2. Distribution de la statistique D_{M} ($d_{\mathrm{M}\,obs} = |\overrightarrow{\mathrm{OD}_{obs}}| = 2.258$).

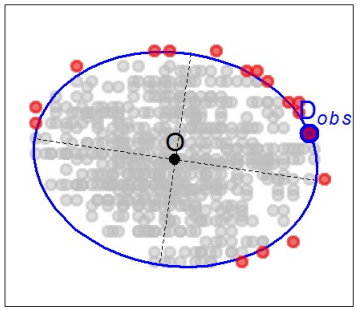


FIGURE 4.11 – *Exemple « Homogénéité »*. Cas 2. Interprétation géométrique du seuil observé exact : proportion des points D^j situés sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I, de centre O et passant par D_{obs} (points rouges).

On a $p_{obs} \leq .05$, résultat significatif au seuil .05. Pour la statistique M–*Distance*, nous pouvons dire que les deux groupes sont hétérogènes au seuil .05.

Cas particulier de la comparaison globale de deux groupes ($\mathrm{C}' = \mathrm{C} = 2$) : équivalence entre test ensembliste de typicalité et test d'homogénéité.

Propriété 2.4. *Lorsque $\mathrm{C}' = \mathrm{C} = 2$, le test d'homogénéité comparant les groupes $c1$ et $c2$ selon la statistique* M–*Distance est équivalent au test ensembliste de typicalité comparant le groupe $c1$ (ou le groupe $c2$) à la réunion des groupes $c1$ et $c2$ selon la même statistique (cf. chapitre 3, p.54).*

Démonstration. Pour comparer le groupe $c1$ (auquel est associé le nuage $\mathrm{M}^{I<c1>}$ de n_1 points et de point moyen G^1) à la réunion des groupes $c1$ et $c2$ (à laquelle est associé le nuage M^I de N points et de point moyen O) grâce au test ensembliste de typicalité selon la statistique M–*Distance* (*cf.* chapitre 3, p.54), on construit le nuage d'échantillonnage des $\binom{n_1+n_2}{n_1} = \binom{N}{n_1}$ points C^j. Le seuil observé exact est alors égal à :

$$p_{obs.Typ} = p(|\overrightarrow{\mathrm{OC}^j}| \geq |\overrightarrow{\mathrm{OG}^1}|)$$

Pour comparer les groupes $c1$ et $c2$ (auxquels sont associés les nuages $\mathrm{M}^{I<c1>}$ de n_1 points et $\mathrm{M}^{I<c2>}$ de n_2 points et admettant G^1 et G^2 pour points moyens) grâce au test d'homogénéité selon la statistique M–*Distance*, nous construisons les $\binom{n_1+n_2}{n_1} = \binom{N}{n_1}$ couples de points $(\mathrm{H}^{1j}, \mathrm{H}^{2j})$. Le seuil observé exact est alors égal à :

$$p_{obs.Hom} = p(|\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}| \geq |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|)$$

Rappelons que O est le point moyen des points G^1 et G^2 et des points H^{1j} et H^{2j} (pour j fixé), nous pouvons donc écrire :

$$\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}} = -\frac{n_1+n_2}{n_2}\overrightarrow{\mathrm{OH}^{1j}} \text{ et } \overrightarrow{\mathrm{G}^1\mathrm{G}^2} = -\frac{n_1+n_2}{n_2}\overrightarrow{\mathrm{OG}^1}$$

On a donc :

$$\begin{aligned} p_{obs.Hom} &= p(|\overrightarrow{\mathrm{H}^{1j}\mathrm{H}^{2j}}| \geq |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|) \\ &= p(|-\tfrac{n_1+n_2}{n_2}\overrightarrow{\mathrm{OH}^{1j}}| \geq |-\tfrac{n_1+n_2}{n_2}\overrightarrow{\mathrm{OG}^1}|) \\ &= p(\tfrac{n_1+n_2}{n_2}|\overrightarrow{\mathrm{OH}^{1j}}| \geq \tfrac{n_1+n_2}{n_2}|\overrightarrow{\mathrm{OG}^1}|) \\ &= p(|\overrightarrow{\mathrm{OH}^{1j}}| \geq |\overrightarrow{\mathrm{OG}^1}|) \end{aligned}$$

Or les $\binom{N}{n_1}$ points C^j et H^{1j} sont les mêmes : ce sont les points moyens des $\binom{N}{n_1}$ sous–nuages de n_1 points qu'il est possible de construire à partir du nuage M^I de N points.
D'où :

$$p_{obs.Hom} = p(|\overrightarrow{\mathrm{OH}^{1j}}| \geq |\overrightarrow{\mathrm{OG}^1}|) = p(|\overrightarrow{\mathrm{OC}^j}| \geq |\overrightarrow{\mathrm{OG}^1}|) = p_{obs.Typ}$$

c.q.f.d.

2.1.2 Zone de compatibilité

Toujours dans le cas $C = \{c1, c2, c_r\}$, nous nous intéressons à l'homogénéité de $\mathrm{C}' = 2$ groupes d'intérêt $c1$ et $c2$, d'effectifs respectifs n_1 et n_2. Rappelons que les deux sous–nuages $\mathrm{M}^{I<c1>}$ et $\mathrm{M}^{I<c2>}$ associés admettent les points G^1 et G^2 pour points moyens respectifs. Au groupe restant c_r (qui peut être vide), nous associons le sous–nuage $\mathrm{M}^{I<c_r>}$ de $N-(n_1+n_2)$ points et de point moyen G^r.

La figure 4.12 représente les trois sous–nuages $\mathrm{M}^{I<c1>}$, $\mathrm{M}^{I<c2>}$, $\mathrm{M}^{I<c_r>}$ et leurs points moyens (*Exemple « Homogénéité », cas 2*, p.89).

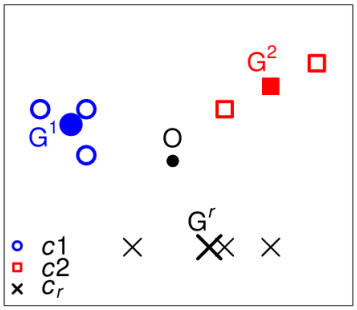


FIGURE 4.12 – *Exemple « Homogénéité »*. Cas 2. Sous–nuages $\mathrm{M}^{I<c1>}$, $\mathrm{M}^{I<c2>}$, $\mathrm{M}^{I<c_r>}$ et leurs points moyens respectifs G^1, G^2, G^r.

Pour construire la zone de compatibilité (définie ci–après), nous adoptons la démarche suivante :

1. Pour tout vecteur $\overrightarrow{u}$ de $\mathcal{L}$, nous effectuons la translation de vecteur $n_2\overrightarrow{u}$ du sous–nuage $\mathrm{M}^{I<c1>}$ et $-n_1\overrightarrow{u}$ du sous–nuage $\mathrm{M}^{I<c2>}$. Les deux sous–nuages translatés sont notés $\mathrm{M}'^{I<c1>}$ et $\mathrm{M}'^{I<c2>}$, ils admettent les points G'^1 et G'^2 pour points moyens et gardent la même structure de covariance que les sous–nuages $\mathrm{M}^{I<c1>}$ et $\mathrm{M}^{I<c2>}$ (figure 4.13) :

$$\begin{aligned} \forall i \in I<c1> \ : \ & \mathrm{M}'^{I<c1>} = \mathrm{M}^i + n_2\overrightarrow{u}, \qquad \mathrm{G}'^1 = \mathrm{G}^1 + n_2\overrightarrow{u} \\ \forall i \in I<c2> \ : \ & \mathrm{M}'^{I<c2>} = \mathrm{M}^i - n_1\overrightarrow{u}, \qquad \mathrm{G}'^2 = \mathrm{G}^2 - n_1\overrightarrow{u} \end{aligned}$$

Le nuage constitué de la réunion des trois sous–nuages $\mathrm{M}'^{I<c1>}$, $\mathrm{M}'^{I<c2>}$ et $\mathrm{M}^{I<c_r>}$ est noté M'^I.

Remarque : Par construction, le point moyen du nuage M'^I est le point moyen O du nuage M^I. En effet :

$$\begin{aligned} n_1\overrightarrow{OG'^1} + n_2\overrightarrow{OG'^2} + (N-(n_1+n_2))\overrightarrow{OG'^r} &= n_1\overrightarrow{OG^1} + n_1n_2\,\overrightarrow{u} + n_2\overrightarrow{OG^2} \\ &\quad -n_1n_2\,\overrightarrow{u} + (N-(n_1+n_2))\overrightarrow{OG'^r} \\ &= \overrightarrow{OO} \\ &= \overrightarrow{0}. \end{aligned}$$

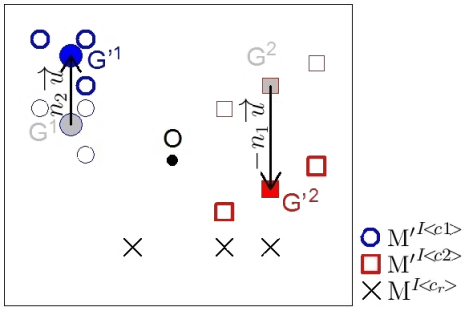


FIGURE 4.13 – *Exemple « Homogénéité »*. Cas 2. Construction des nuages $M'^{I<c1>}$ et $M'^{I<c2>}$.

2. Au bipoint (G'^1, G'^2), nous associons le vecteur–écart $\overrightarrow{G'^1G'^2}$ et nous construisons le point P tel que (figure 4.14) :

$$P = O + \overrightarrow{G'^1G'^2}$$

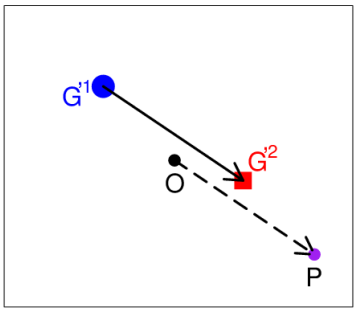


FIGURE 4.14 – *Exemple « Homogénéité »*. Cas 2. Construction du point P.

Remarque : On a :

$$\overrightarrow{OP} = \overrightarrow{G^1G^2} - (n_1+n_2)\,\overrightarrow{u}$$

En effet,

$$\begin{aligned} \overrightarrow{OP} &= \overrightarrow{G'^1G'^2} \\ &= \overrightarrow{G'^1G^1} + \overrightarrow{G^1G^2} + \overrightarrow{G^2G'^2} \\ &= -n_2\,\overrightarrow{u} + \overrightarrow{G^1G^2} - n_1\,\overrightarrow{u} \\ &= \overrightarrow{G^1G^2} - (n_1+n_2)\,\overrightarrow{u}. \end{aligned}$$

Nous étudions maintenant la compatibilité entre les points O et P (ou, de manière équivalente, entre les points G'^1 et G'^2) de la manière suivante :

1. En considérant le nuage $\mathrm{M'}^I$ de N points, nous construisons l'espace des $\mathrm{J} = \frac{N!}{n_1! \times n_2! \times (N-(n_1+n_2))!}$ emboîtements et les points moyens d'intérêt associés

$$\mathrm{H'}^{1j} = \sum_{i \in I<c1>} \mathrm{M'}^i / n_1 \text{ et } \mathrm{H'}^{2j} = \sum_{i \in I<c2>} \mathrm{M'}^i / n_2$$

2. En notant $|\overrightarrow{\mathrm{AB}}|_{\mathrm{M'}}$ la distance de Mahalanobis attachée au nuage $\mathrm{M'}^I$ entre deux points A et B de $\mathcal{M}$, nous calculons le seuil observé exact :

$$p_{obs} = p(|\overrightarrow{\mathrm{H'}^{1j}\mathrm{H'}^{2j}}|_{\mathrm{M'}} \geq |\overrightarrow{\mathrm{OP}}|_{\mathrm{M'}}) \quad \Big(= p(|\overrightarrow{\mathrm{H'}^{1j}\mathrm{H'}^{2j}}|_{\mathrm{M'}} \geq |\overrightarrow{\mathrm{G'}^1\mathrm{G'}^2}|_{\mathrm{M'}}) \Big)$$

3. Si $p_{obs} > \alpha$, les points O et P sont compatibles au seuil α ; si $p_{obs} \leq \alpha$, ils sont incompatibles (définition 2.1 suivante).

Définition 2.1 (Compatibilité/Incompatibilité)**.** *Les points* O *et* P *sont compatibles (resp. incompatibles) au seuil* α *si le seuil observé exact* $p_{obs} = p(|\overrightarrow{\mathrm{H'}^{1j}\mathrm{H'}^{2j}}|_{\mathrm{M'}} \geq |\overrightarrow{\mathrm{OP}}|_{\mathrm{M'}})$ *est strictement supérieur à* α *(resp. inférieur ou égal à* α*).*

Remarque : Comme $\overrightarrow{\mathrm{OP}} = \overrightarrow{\mathrm{G'}^1\mathrm{G'}^2}$, si les points O et P sont compatibles au seuil α alors les points $\mathrm{G'}^1$ et $\mathrm{G'}^2$ sont également compatibles au seuil α.

Définition 2.2 (Zone de compatibilité)**.** *La zone de compatibilité au seuil* $(1-\alpha)$ *est l'ensemble des points* P *compatibles avec le point* O *au seuil* α*.*

Définition 2.3 (Point–limite d'incompatibilité)**.** *Un point* P *est point–limite d'incompatibilité s'il est juste incompatible avec le point* O *au seuil* α*, c'est–à–dire si* $p(|\overrightarrow{\mathrm{H'}^{1j}\mathrm{H'}^{2j}}|_{\mathrm{M'}} \geq |\overrightarrow{\mathrm{OP}}|_{\mathrm{M'}}) = \alpha$*.*

Caractérisation de la zone de compatibilité.

Considérons la partition du nuage M^I en C sous–nuages $\mathrm{M}^{I<c1>}$, $\mathrm{M}^{I<c2>}$ et $\mathrm{M}^{I<c_r>}$ (*cf.* figure 4.12, p.100). En faisant choix d'un repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$, notons $\mathbf{W}$ la matrice de covariance du nuage intra et $\mathbf{B}$ la matrice de covariance du nuage inter. On a $\mathbf{V} = \mathbf{B} + \mathbf{W}$.

Notons également :

- G^{12} le barycentre des points G^1 et G^2, munis respectivement des pondérations n_1 et n_2,
- $\mathbf{d}$ le vecteur–colonne associé aux coordonnées de $\overrightarrow{\mathrm{G}^{12}\mathrm{G}^r}$,
- $f_1 = n_1/N$, $f_2 = n_2/N$, $f_r = (N - (n_1 - n_2))/N$,
- $\mathbf{Y}$ la matrice définie par : $\mathbf{Y} = \mathbf{W} + (f_1 + f_2) f_r \, \mathbf{d}\mathbf{d}'$.

La $\mathbf{Y}$–norme du vecteur $\overrightarrow{u}$, notée $|\overrightarrow{u}|_{\mathbf{Y}}$, est définie de la façon suivante :

$$\forall \overrightarrow{u} \in \mathcal{L}, \; |\overrightarrow{u}|_{\mathbf{Y}} = \sqrt{\mathbf{u}'\mathbf{Y}^{-1}\mathbf{u}}$$

où $\mathbf{u}$ est le vecteur–colonne associé aux coordonnées de $\overrightarrow{u}$.

Théorème 2.1. *Soit* P *un point–limite d'incompatibilité au seuil* α*. Si* $|\overrightarrow{\mathrm{OP}}|_{\mathbf{Y}} = \kappa_\alpha$*, alors tout point de l'ellipsoïde défini par l'ensemble des points* Q *tels que* $|\overrightarrow{\mathrm{OQ}}|_{\mathbf{Y}} = \kappa_\alpha$ *est aussi point–limite d'incompatibilité au seuil* α*.*

Démonstration. Considérons le nuage $\mathrm{M'}^I$ (*cf.* figure 4.13, p.101) et sa matrice de covariance $\mathbf{V'}$ ($\neq \mathbf{V}$, matrice de covariance du nuage M^I). On a :

$$\mathbf{V'} = \mathbf{W'} + \mathbf{B'}$$

où $\mathbf{W'}$ et $\mathbf{B'}$ sont les matrices de covariance intra et inter associées à la partition en C sous–nuages $\mathrm{M'}^{I<c1>}$, $\mathrm{M'}^{I<c2>}$ et $\mathrm{M}^{I<c_r>}$ du nuage $\mathrm{M'}^I$.

Par construction (translation), les nuages $\mathrm{M'}^{I<c1>}$ et $\mathrm{M}^{I<c1>}$, $\mathrm{M'}^{I<c2>}$ et $\mathrm{M}^{I<c2>}$ ont la même structure de covariance. Par conséquent, les deux nuages intra issus des partitions en C sous–nuages des nuages M^I et $\mathrm{M'}^I$ sont identiques, on a donc $\mathbf{W'} = \mathbf{W}$. D'où :

$$\begin{aligned}\mathbf{V'} &= \mathbf{W} + \mathbf{B'} \\ &= \mathbf{W} + f_1\ \mathbf{d}_1\mathbf{d'}_1 + f_2\ \mathbf{d}_2\mathbf{d'}_2 + f_r\ \mathbf{d}_r\mathbf{d'}_r\end{aligned}$$

où $\mathbf{d}_1$, $\mathbf{d}_2$ et $\mathbf{d}_r$ sont respectivement les vecteurs–colonnes des coordonnées de $\overrightarrow{\mathrm{OG'}^1}$, $\overrightarrow{\mathrm{OG'}^2}$ et $\overrightarrow{\mathrm{OG}^r}$.

D'après le premier théorème de Huyghens (*cf.* chapitre 1, p.13) on a :

$$\frac{f_1\ \mathbf{d}_1\mathbf{d'}_1}{f_1+f_2} + \frac{f_2\ \mathbf{d}_2\mathbf{d'}_2}{f_1+f_2} = \mathbf{d}_{12}\mathbf{d'}_{12} + \mathrm{Var}(\{\mathrm{G'}^1, \mathrm{G'}^2\})$$

où $\mathbf{d}_{12}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{OG}^{12}}$, les bipoints $(\mathrm{G'}^1, \mathrm{G'}^2)$ et $(\mathrm{G}^1, \mathrm{G}^2)$ admettent le même point moyen G^{12}.

D'où $\mathbf{V'} = \mathbf{W} + (f_1+f_2)\ \mathbf{d}_{12}\mathbf{d'}_{12} + (f_1+f_2)\ \mathrm{Var}(\{\mathrm{G'}^1, \mathrm{G'}^2\}) + f_r\ \mathbf{d}_r\mathbf{d'}_r$.

Or $(f_1+f_2)\ \mathbf{d}_{12}\mathbf{d'}_{12} + f_r\ \mathbf{d}_r\mathbf{d'}_r = \mathrm{Var}(\{\mathrm{G}^{12}, \mathrm{G}^r\}) = (f_1+f_2)f_r\ \mathbf{dd'}$,
et $(f_1+f_2)\ \mathrm{Var}(\{\mathrm{G'}^1, \mathrm{G'}^2\}) = (f_1+f_2)f_1f_2\ \mathbf{ee'}$ où $\mathbf{e}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{G'}^1\mathrm{G'}^2}$ (ou de $\overrightarrow{\mathrm{OP}}$).

$$\begin{aligned}\text{On a donc : } \mathbf{V'} &= \mathbf{W} + (f_1+f_2)f_r\ \mathbf{dd'} + (f_1+f_2)f_1f_2\ \mathbf{ee'} \\ &= \mathbf{Y} + (f_1+f_2)f_1f_2\ \mathbf{ee'}\end{aligned}$$

et, d'après la propriété généralisée de réciprocité (*cf.* chapitre 1, p.19) :

$$|\overrightarrow{\mathbf{e}}|^2_{\mathbf{Y}} = \frac{|\overrightarrow{\mathbf{e}}|^2_{\mathrm{M'}}}{1-(f_1+f_2)f_1f_2|\overrightarrow{\mathbf{e}}|^2_{\mathrm{M'}}}$$

Soit h_α un nombre positif tel que le point P $(= \mathrm{O} + \overrightarrow{\mathrm{G'}^1\mathrm{G'}^2})$ soit point–limite d'incompatibilité au seuil α :

$$|\overrightarrow{\mathrm{OP}}|_{\mathrm{M'}} = h_\alpha$$

et soit le κ_α–ellipsoïde de centre O passant par P défini par :

$$\kappa^2_\alpha = |\overrightarrow{\mathrm{OP}}|^2_{\mathbf{Y}} = \frac{|\overrightarrow{\mathrm{OP}}|^2_{\mathrm{M'}}}{1-(f_1+f_2)f_1f_2|\overrightarrow{\mathrm{OP}}|^2_{\mathrm{M'}}} = \frac{h^2_\alpha}{1-(f_1+f_2)f_1f_2h^2_\alpha}$$

Soit Q un point appartenant à ce κ_α–ellipsoïde :

$$|\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{Y}} = \kappa^2_\alpha$$

Pour étudier la compatibilité entre les points O et Q, construisons le nuage $\mathrm{M''}^I$ (*cf.* p.100) en choisissant le vecteur $\overrightarrow{u}$ de sorte que $|\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{Y}} = \kappa^2_\alpha$, c'est–à–dire tel que (*cf.* remarque p.101) :

$$|\overrightarrow{\mathrm{G}^1\mathrm{G}^2} - (n_1+n_2)\overrightarrow{u}|^2_{\mathbf{Y}} = \kappa^2_\alpha$$

Soit $\mathbf{V}''$ la matrice de covariance du nuage M''^{I}, on a :

$$\mathbf{V}'' = \mathbf{Y} + (f_1 + f_2) f_1 f_2 \ \mathbf{q}\mathbf{q}'$$

où $\mathbf{q}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{OQ}}$.
En notant $|\overrightarrow{\mathrm{AB}}|_{\mathrm{M}''}$ la distance de Mahalanobis attachée au nuage M''^{I} entre deux points A et B de $\mathcal{M}$, on a d'après la propriété généralisée de réciprocité :

$$|\overrightarrow{\mathbf{q}}|^2_{\mathbf{Y}} = \frac{|\overrightarrow{\mathbf{q}}|^2_{\mathrm{M}''}}{1 - (f_1 + f_2) f_1 f_2 |\overrightarrow{\mathbf{q}}|^2_{\mathrm{M}''}}$$

On peut donc écrire :

$$\begin{aligned} |\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{Y}} = \kappa^2_\alpha &\Leftrightarrow \frac{|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}}{1-(f_1+f_2)f_1f_2|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}} = \kappa^2_\alpha \\ &\Leftrightarrow \frac{|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}}{1-(f_1+f_2)f_1f_2|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}} = \frac{h^2_\alpha}{1-(f_1+f_2)f_1f_2h^2_\alpha} \\ &\Leftrightarrow |\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''} = h^2_\alpha \end{aligned}$$

Q est donc aussi un point–limite d'incompatibilité.
c.q.f.d.

Nous pouvons dire que les points-limites d'incompatibilité au seuil α appartiennent, aux tracas du discret près, au même κ_α–ellipsoïde de centre O défini par l'ensemble des points Q tels que :

$$|\overrightarrow{\mathrm{OQ}}|_{\mathbf{Y}} = \kappa_\alpha$$

La zone de compatibilité au seuil $1 - \alpha$ est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de ce κ_α–ellipsoïde.

Dans la mesure où la compatibilité est évaluée grâce au test exact (par opposition au test approché que nous exposons dans la suite), cette zone peut également être appelée *zone de compatibilité exacte.*

Exemple « Homogénéité », cas 2 : Sur la figure 4.15 est représenté en rouge l'ensemble des points P juste incompatibles avec le point O au seuil $\frac{28}{560} = .05$ [10].

L'ellipse ajustée à cet ensemble de points (en bleu sur la figure 4.15) est l'ellipse de centre O définie par l'ensemble des points Q tels que $|\overrightarrow{\mathrm{OQ}}|_{\mathbf{Y}} = 4.134$ [11]. La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

10. *Note sur le calcul :* pour illustrer le théorème 2.1, nous avons ici voulu obtenir des points–limites d'incompatibilité au seuil .05. Pour ce faire, nous avons quadrillé l'espace afin de faire varier le vecteur $\overrightarrow{u}$ (*cf.* p.100). Ici la première coordonnée de $\overrightarrow{u}$ (sur l'axe horizontal) varie -2 à 10 par pas de 0.2 tandis que la seconde (sur l'axe vertical) varie de -9 à 10 par pas de 0.2. Pour chaque vecteur $\overrightarrow{u}$ ainsi obtenu, nous avons construit les nuages $\mathrm{M}'^{I<c1>}$ et $\mathrm{M}'^{I<c2>}$ correspondants et effectué le test à partir de ces nuages translatés et du nuage $\mathrm{M}^{I<c_r>}$ restant. Les points–limites obtenus ont ensuite été ajustés par une ellipse dont le κ est donné ici.

11. *Méthode d'ajustement :* pour un point–limite d'incompatibilité P donné, nous avons ici calculé la valeur k telle que $k = |\overrightarrow{\mathrm{OP}}|_{\mathbf{Y}}$. En effectuant ce procédé pour tous les points–limites trouvés, nous obtenons un ensemble de valeurs dont la moyenne donne le κ de l'ellipse ajustée.

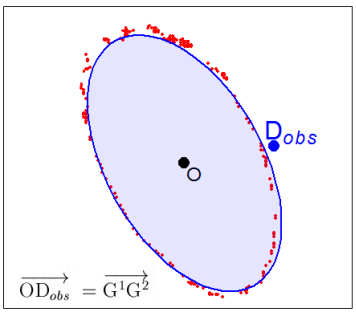


FIGURE 4.15 – *Exemple « Homogénéité »*. Cas 2. Points–limites d'incompatibilité au seuil .05 (en rouge) et zone de compatibilité au seuil .95 ($\kappa = 4.134$, en bleu).

Le point D_{obs} est situé à l'extérieur de la zone de compatibilité : les points O et D_{obs} sont incompatibles au seuil .05.

Cas particulier d'un emboîtement en deux groupes ($C' = C = 2$).

Les deux groupes d'intérêt $c1$ et $c2$ forment une partition de l'ensemble I : $C = \{c1, c2\}$. Ce cas particulier est illustré par l'exemple suivant : considérons un ensemble I de $N = 5$ individus : $I = \{i1, \ldots, i5\}$ et la partition en $C' = C = 2$ groupes suivante : $I < c1 >= \{i1, i2, i3\}$ et $I < c2 >= \{i4, i5\}$. On a donc $n_1 = 3$, et $n_2 = 2$. A l'ensemble I est associé le nuage M^I : nuage plan ($L = 2$) de 5 points, de point moyen O. Le nuage M^I et sa partition en 2 sous–nuages sont représentés sur la figure 4.16.

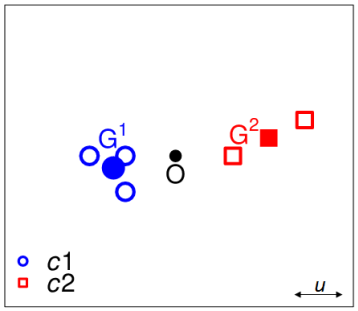


FIGURE 4.16 – Sous–nuages $M^{I<c1>}$ et $M^{I<c2>}$ de points moyens respectifs G^1 et G^2. Pour les calculs numériques, nous prenons u comme unité de longueur.

Dans le cas particulier d'un emboîtement en deux groupes, le nuage $M^{I<c_r>}$ n'existe pas, les formules se simplifient et on a $\mathbf{Y} = \mathbf{W}$. Le théorème 2.1 peut donc être reformulé ainsi :

Théorème 2.2. *Si un point* P *est point–limite d'incompatibilité au seuil* α*, alors tout point de l'ellipsoïde d'inertie du nuage intra passant par* P *est aussi point–limite d'incompatibilité au seuil* α*.*

La démonstration du théorème 2.1 (p.103) suffit à démontrer ce théorème. Cependant, une démonstration simplifiée propre à ce cas particulier est donnée ici à titre indicatif.

Démonstration. Considérons le nuage M'^I (tel que défini p.102) et sa matrice de covariance $\mathbf{V}'$ ($\neq \mathbf{V}$, matrice de covariance du nuage M^I) qui se décompose de la façon suivante :

$$\mathbf{V}' = \mathbf{W}' + \mathbf{B}'$$

où $\mathbf{W}'$ et $\mathbf{B}'$ sont les matrices de covariance intra et inter associées à la partition en deux sous–nuages du nuage M'^I.

Par construction (translation), les nuages $\mathrm{M}'^{I<c1>}$ et $\mathrm{M}^{I<c1>}$, $\mathrm{M}'^{I<c2>}$ et $\mathrm{M}^{I<c2>}$ ont la même structure de covariance. Par conséquent, les deux nuages intra issus des partitions en $\mathrm{C}' = 2$ sous–nuages des nuages M^I et M'^I sont identiques, on a donc $\mathbf{W}' = \mathbf{W}$. D'où :

$$\begin{aligned}\mathbf{V}' &= \mathbf{W} + \mathbf{B}' \\ &= \mathbf{W} + f_1 f_2 \ \mathbf{e}\mathbf{e}'\end{aligned}$$

où $\mathbf{e}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{G}'^1\mathrm{G}'^2}$.
D'après la propriété généralisée de réciprocité (*cf.* chapitre 1, p.19), on a :

$$|\overrightarrow{\mathbf{e}}|^2_{\mathbf{W}} = \frac{|\overrightarrow{\mathbf{e}}|^2_{\mathrm{M}'}}{1 - f_1 f_2 |\overrightarrow{\mathbf{e}}|^2_{\mathrm{M}'}}$$

avec $|\cdot|_{\mathbf{W}}$ la norme de Mahalanobis attachée au nuage intra.

Soit h_α un nombre positif tel que le point P $(= \mathrm{O} + \overrightarrow{\mathrm{G}'^1\mathrm{G}'^2})$ soit point–limite d'incompatibilité au seuil α :

$$|\overrightarrow{\mathrm{OP}}|_{\mathrm{M}'} = h_\alpha$$

et soit le κ_α–ellipsoïde d'inertie du nuage intra passant par P, c'est–à–dire tel que :

$$\kappa_\alpha^2 = |\overrightarrow{\mathrm{OP}}|^2_{\mathbf{W}} = \frac{|\overrightarrow{\mathrm{OP}}|^2_{\mathrm{M}'}}{1 - f_1 f_2 |\overrightarrow{\mathrm{OP}}|^2_{\mathrm{M}'}} = \frac{h_\alpha^2}{1 - f_1 f_2 h_\alpha^2}$$

Soit Q un point appartenant à ce κ_α–ellipsoïde :

$$|\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{W}} = \kappa_\alpha^2$$

Pour étudier la compatibilité entre les points O et Q, construisons le nuage M''^I (*cf.* p.100) en choisissant le vecteur $\overrightarrow{u}$ de sorte que $|\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{W}} = \kappa_\alpha^2$, c'est–à–dire tel que (*cf.* remarque p.101) :

$$|\overrightarrow{\mathrm{G}^1\mathrm{G}^2} - (n_1 + n_2)\overrightarrow{u}|^2_{\mathbf{W}} = \kappa_\alpha^2$$

Soit $\mathbf{V}''$ la matrice de covariance du nuage M''^I, on a :

$$\mathbf{V}'' = \mathbf{W} + f_1 f_2 \ \mathbf{q}\mathbf{q}'$$

où $\mathbf{q}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{OQ}}$.
En notant $|\overrightarrow{\mathrm{AB}}|_{\mathrm{M}''}$ la distance de Mahalanobis attachée au nuage M''^I entre deux points A et B de $\mathcal{M}$, on a d'après la propriété généralisée de réciprocité :

$$|\overrightarrow{\mathbf{q}}|^2_{\mathbf{W}} = \frac{|\overrightarrow{\mathbf{q}}|^2_{\mathrm{M}''}}{1 - f_1 f_2 |\overrightarrow{\mathbf{q}}|^2_{\mathrm{M}''}}$$

On peut donc écrire :

$$\begin{aligned}|\overrightarrow{\mathrm{OQ}}|^2_{\mathbf{W}} = \kappa_\alpha^2 &\Leftrightarrow \frac{|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}}{1 - f_1 f_2 |\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}} = \kappa_\alpha^2 \\ &\Leftrightarrow \frac{|\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}}{1 - f_1 f_2 |\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''}} = \frac{h_\alpha^2}{1 - f_1 f_2 h_\alpha^2} \\ &\Leftrightarrow |\overrightarrow{\mathrm{OQ}}|^2_{\mathrm{M}''} = h_\alpha^2\end{aligned}$$

Q est donc aussi un point–limite d'incompatibilité.
c.q.f.d.

Dans ce cas particulier, nous pouvons dire que les points-limites d'incompatibilité au seuil α appartiennent, aux tracas du discret près, au même κ_α–ellipsoïde d'inertie du nuage intra. La zone de compatibilité au seuil $1-\alpha$ est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de ce κ_α–ellipsoïde.

Le nuage intra et son ellipse de concentration sont représentés à gauche de la figure 4.17. A droite est représentée la zone de compatibilité au seuil $1-\alpha$ si P est point–limite d'incompatibilité au seuil α.

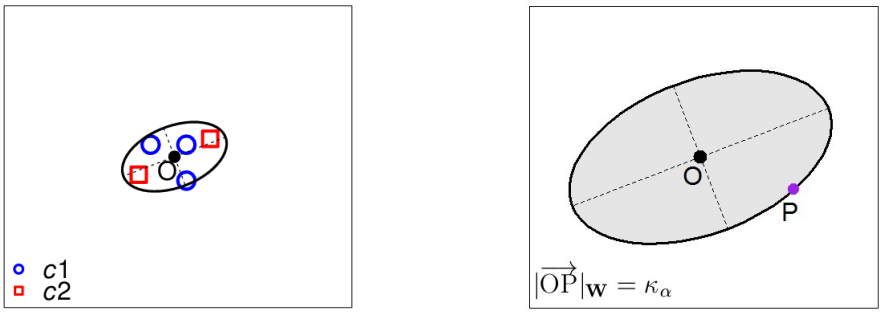


FIGURE 4.17 – Nuage intra et son ellipse de concentration (gauche), zone de compatibilité au seuil $1-\alpha$ si P est point–limite d'incompatibilité au seuil α (droite).

Décomposition d'une comparaison à $C'-1$ degrés de liberté en $C'-1$ comparaisons à 1 degré de liberté. Lorsque nous comparons C' groupes (avec $C' > 2$) et que le test d'homogénéité est significatif, c'est–à–dire que les groupes sont hétérogènes, nous sommes en mesure de nous demander où se situent les différences. Une procédure possible consiste à effectuer $C'-1$ comparaisons à 1 degré de liberté (binaires) associées à $C'-1$ contrastes orthogonaux. Par exemple, si on a un emboîtement équilibré en trois groupes $c1$, $c2$ et $c3$ (effectifs égaux), nous pouvons prendre les bases de contrastes orthogonaux $(1,-1,0)$ et $(1/2,1/2,-1)$, qui consistent respectivement à effectuer la comparaison partielle des groupes $c1$ et $c2$ et la comparaison entre le regroupement des groupes $c1$ et $c2$ et le groupe $c3$. Dans le cas non équilibré, où les groupes $c1$, $c2$ et $c3$ sont d'effectifs respectifs n_1, n_2 et n_3, nous pouvons prendre la base de contrastes n_c–orthogonaux [12] $(1,-1,0)$ et $(\frac{n_1}{n_1+n_2}, \frac{n_2}{n_1+n_2}, -1)$.

Les zones de compatibilité correspondantes, définies pour la comparaison de deux groupes, peuvent alors être construites (se référer alors à la procédure décrite au paragraphe 2.1.2, p.100).

2.1.3 Cas particulier d'un nuage unidimensionnel

Nous sommes ici dans la situation décrite au paragraphe 1.1.2 (p.93). Cependant, lorsque nous nous intéressons à l'homogénéité de deux groupes d'intérêt $c1$ et $c2$ de moyennes respectives m^1 et m^2 ($C' = 2 \leq C$), nous pouvons utiliser la statistique D suivante comme statistique de test :

$$\begin{array}{rccc} D: & \mathbb{R} & \to & \mathbb{R} \\ & m^{1j}, m^{2j} & \mapsto & m^{1j} - m^{2j} \end{array}$$

Cette statistique dépend uniquement de la différence entre les couples de valeurs $(m^{1j}, m^{2j})_{j\in J}$, c'est pourquoi nous l'appelons *Différence*. Notons $d_{obs} = m^1 - m^2$ sa valeur observée.

12. Deux contrastes u_c et v_c ($\sum_{c\in C} u_c = \sum_{c\in C} v_c = 0$) sont n_c–orthogonaux si $\sum_{c\in C} \frac{u_c v_c}{n_c} = 0$.

Choix de la statistique de test : afin de tenir compte du signe de l'effet, nous utilisons la statistique de test D qui mesure la différence entre les couples de valeurs $(m^{1j}, m^{2j})_{j\in J}$, plutôt que la statistique D_{M}, utilisée précédemment, qui mesure la valeur absolue de cette différence.

Seuil observé exact. Il s'impose d'effectuer ici un test *unilatéral* (*supérieur*, si $m^1 > m^2$ ou *inférieur*, si $m^1 < m^2$). Si $m^1 > m^2$, on calcule le seuil observé exact supérieur p_{sup}, c'est–à–dire la proportion des valeurs de D pour lesquelles $D \geq d_{obs}$; si $m^1 < m^2$, on calcule le seuil observé exact inférieur p_{inf}, c'est–à–dire la proportion des valeurs de D pour lesquelles $D \leq d_{obs}$. Le seuil observé exact unilatéral p_{unil} est par définition p_{sup} si $m^1 > m^2$ ou p_{inf} si $m^1 < m^2$.

Nous énonçons la conclusion du test, en termes d'hétérogénéité des deux groupes, en tenant compte du signe de l'effet (pour la statistique D).

Si $m^1 > m^2$ et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. Les groupes $c1$ et $c2$ sont hétérogènes au seuil unilatéral $\alpha/2$. La moyenne du groupe $c1$ est significativement supérieure à celle du groupe $c2$ au seuil unilatéral $\alpha/2$.

Si $m^1 < m^2$ et $p_{unil} \leq \alpha/2$, le test est significatif au seuil unilatéral $\alpha/2$. Les groupes $c1$ et $c2$ sont hétérogènes au seuil unilatéral $\alpha/2$. La moyenne du groupe $c1$ est significativement inférieure à celle du groupe $c2$ au seuil unilatéral $\alpha/2$.

Si $p_{unil} > \alpha/2$, le test est non significatif au seuil bilatéral α. Nous ne pouvons pas dire que les deux groupes sont hétérogènes au seuil bilatéral α.

Remarque : Comme dans le cas multidimensionnel, lorsque $\mathrm{C}' = \mathrm{C} = 2$, il y a équivalence entre le test d'homogénéité comparant les groupes $c1$ et $c2$ selon la statistique *Différence* et le test de typicalité comparant le groupe $c1$ (ou le groupe $c2$) à la réunion des groupes $c1$ et $c2$ selon la statistique *Moyenne* (*cf.* chapitre 3, p.63).

Caractéristiques de la distribution de la statistique de test D.

Propriété 2.5. *La moyenne de la statistique Différence est nulle :*

$$\mathrm{Moy}(D) = 0$$

Démonstration.
$$\begin{aligned}\mathrm{Moy}(D) &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} (m^{1j} - m^{2j}) \\ &= \tfrac{1}{\mathrm{J}} \sum_{j\in J} m^{1j} - \tfrac{1}{\mathrm{J}} \sum_{j\in J} m^{2j}\end{aligned}$$
D'après la propriété 1.1 (p.90), le point H^{cj} pouvant être remplacé par la valeur m^{cj} (ici $c = \{1, 2\}$) et le point O par la valeur m, on a :
$$\begin{aligned}\mathrm{Moy}(D) &= m - m \\ &= 0\end{aligned}$$
c.q.f.d.

Propriété 2.6. *La variance de la statistique Différence est :*

$$\mathrm{Var}\, D = \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times v$$

Démonstration. Découle directement de la propriété 2.2 (p.97).

Intervalle de compatibilité. Pour construire l'intervalle de compatibilité, nous adaptons la démarche décrite au paragraphe 2.1.2 (p.100) au cas unidimensionnel :

$$\forall u \in \mathbb{R},\ x'^{I<c1>} = x^{I<c1>} + n_2 u \text{ et } x'^{I<c2>} = x^{I<c2>} - n_1 u$$

On a $\text{Moy}(x'^{I<c1>}) = m'_1$ et $\text{Moy}(x'^{I<c2>}) = m'_2$ et on pose $m' = m'_1 - m'_2$.

Les valeurs 0 et m', ou de façon équivalente m'_1 et m'_2, sont compatibles au seuil α si le seuil observé exact associé est strictement supérieur à α (définition 2.1 p.102, adaptée au cas unidimensionnel).

La limite inférieure (resp. supérieure) de l'intervalle de compatibilité au seuil $1-\alpha$ est alors la plus petite (resp. la plus grande) valeur m' compatible avec 0 au seuil unilatéral $\alpha/2$.

2.2 Test approché

2.2.1 Principe du test

Supposons que le nuage des points–écarts D^J (*cf.* p.95) soit ajusté par une distribution gaussienne multidimensionnelle à L dimensions (L dimension de $\mathcal{M}$), de centre O et avec la structure de covariance associée à :

$$\frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times Som$$

c'est–à–dire telle que la densité soit définie par :

$$\forall\ \overrightarrow{x} \in \mathcal{L},\ f(\overrightarrow{x}) = \frac{1}{(2\pi)^{L/2} \det(Som \times \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}))^{1/2}} \exp\,(-\frac{1}{2} \times \frac{N-1}{N} \times \frac{n_1 n_2}{n_1+n_2} |\overrightarrow{x}|^2)$$

Dans ce cas, la distribution de la statistique $\text{T}^2 = \frac{N-1}{N} \times \frac{n_1 n_2}{n_1+n_2} \times D^2_{\text{M}}$ est une distribution du χ^2 à L degrés de liberté (*cf.* définition de D_{M} p.95).

Seuil observé approché. Soit $\text{t}^2_{obs} = \frac{N-1}{N} \times \frac{n_1 n_2}{n_1+n_2} \times d^2_{\text{M}\,obs}$ (avec $d^2_{\text{M}\,obs} = |\overrightarrow{\text{OD}_{obs}}|^2 = |\overrightarrow{\text{G}^1\text{G}^2}|^2$), la proportion $p(\text{T}^2 \geq \text{t}^2_{obs})$ est approximativement égale à $p(\chi^2_L \geq \text{t}^2_{obs}) = \widetilde{p}_{obs}$, seuil observé du test approché (seuil approché ou p–value approchée).

Nous concluons en termes d'hétérogénéité des deux groupes.

Le test approché est illustré par l'exemple des races canines, p.120.

2.2.2 Zone approchée de compatibilité

Rappelons que la zone de compatibilité au seuil $1-\alpha$ est définie comme étant l'ensemble des points P compatibles avec le point O au seuil α. Nous adoptons ici la même démarche que celle décrite au paragraphe 2.1.2 (p.100), la compatibilité entre les points P et O étant évidemment évaluée grâce au test approché.

Définition 2.4 (Zone–limite approchée d'incompatibilité)**.** *La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points–limites d'incompatibilité au seuil α, obtenus par le test approché.*

Théorème 2.3. *La zone–limite approchée d'incompatibilité au seuil* $1-\alpha$ *est le* κ*–ellipsoïde de centre* O *défini par l'ensemble des points* P *tels que* $|\overrightarrow{\mathrm{OP}}|_{\mathbf{Y}} = \kappa$*, avec :*

$$\kappa = \sqrt{\frac{\frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})\ \chi^2_L[\alpha]}{1 - (f_1 + f_2) f_1 f_2 \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})\ \chi^2_L[\alpha]}}$$

Démonstration. Les notations utilisées sont celles de la page 102. D'après la démonstration du théorème 2.1 (p.102), on sait que :

$$|\overrightarrow{\mathbf{e}}|^2_{\mathrm{M}'} = \frac{|\overrightarrow{\mathbf{e}}|^2_{\mathbf{Y}}}{1 + (f_1 + f_2) f_1 f_2 |\overrightarrow{\mathbf{e}}|^2_{\mathbf{Y}}}$$

où $\mathbf{e}$ est le vecteur–colonne des coordonnées de $\overrightarrow{\mathrm{G'^1 G'^2}}$.
La zone–limite approchée d'incompatibilité au seuil $1-\alpha$ est l'ensemble des points P vérifiant :
$\frac{N-1}{N} \times \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}} \times |\overrightarrow{\mathrm{OP}}|^2_{\mathrm{M}'} = \chi^2_L[\alpha] \Leftrightarrow \frac{N-1}{N} \times \frac{1}{\frac{1}{n_1}+\frac{1}{n_2}} \times \frac{|\overrightarrow{\mathrm{OP}}|^2_{\mathbf{Y}}}{1+(f_1+f_2)f_1 f_2 |\overrightarrow{\mathrm{OP}}|^2_{\mathbf{Y}}} = \chi^2_L[\alpha]$
Après calcul on obtient :

$$|\overrightarrow{\mathrm{OP}}|^2_{\mathbf{Y}} = \frac{\frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})\ \chi^2_L[\alpha]}{1 - (f_1 + f_2) f_1 f_2 \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2})\ \chi^2_L[\alpha]}$$

c.q.f.d.

La construction de la zone de compatibilité approchée est illustrée par l'exemple des races canines, p.120.

2.2.3 Cas particulier d'un nuage unidimensionnel

Dans le cas où $L = 1$ et $\mathrm{C}' = 2$, le test approché consiste à ajuster la distribution de la statistique *Différence D* (*cf.* paragraphe 2.1.3, p.107) par une distribution gaussienne, de moyenne nulle et de variance $\mathrm{Var}\, D$ $(= \frac{N}{N-1} \times (\frac{1}{n_1} + \frac{1}{n_2}) \times v)$. Nous utilisons ici la statistique *Écart Réduit* :

$$Z = \frac{D}{\sqrt{\mathrm{Var}\, D}}$$

dont la distribution est une distribution gaussienne $\mathcal{N}(0,1)$. Les deux statistiques D et Z conduisent à des tests combinatoires équivalents.

Seuil observé approché. Soit $z_{obs} = \frac{d_{obs}}{\sqrt{\frac{N}{N-1} \times (\frac{1}{n_1}+\frac{1}{n_2}) \times v}}$ la valeur observée de la statistique *Écart Réduit*. Notons z la variable dont la distribution est une loi de Gauss centrée réduite : $z \sim \mathcal{N}(0,1)$.

Si $m^1 > m^2$, on calcule le seuil observé approché supérieur $\widetilde{p}_{sup} = p(z \geq z_{obs})$; si $m^1 < m^2$, on calcule le seuil observé approché inférieur $\widetilde{p}_{inf} = p(z \leq z_{obs})$. Le seuil observé approché unilatéral $\widetilde{p}_{unil}$ est par définition $\widetilde{p}_{sup}$ si $m^1 > m^2$ ou $\widetilde{p}_{inf}$ si $m^1 < m^2$.

Remarque : On a $\widetilde{p}_{sup} = \widetilde{p}_{inf}$ car la distribution gaussienne est symétrique.

Nous énonçons la conclusion du test en termes d'hétérogénéité des groupes, pour la statistique Z, en tenant compte du signe de l'effet (*cf.* p.108).

Intervalle approché de compatibilité. Soit $z[\alpha]$ la valeur critique, au seuil bilatéral α, de la distribution gaussienne centrée réduite :

$$p(|z| > z[\alpha]) = \alpha$$

L'intervalle approché de compatibilité au seuil $1 - \alpha$ est l'ensemble des valeurs $d \in \mathbb{R}$ compatibles avec 0 au seuil bilatéral α, c'est–à–dire vérifiant :

$$\frac{|d|}{\sqrt{\text{Var } D}} < z[\alpha]$$

ou, ce qui revient au même, appartenant à l'intervalle :

$$]-z[\alpha]\sqrt{\text{Var } D}\,;\ +z[\alpha]\sqrt{\text{Var } D}[$$

Cet intervalle est centré sur 0.

3 Autres tests

Aussi bien pour les tests basés sur le modèle normal que sur le bootstrap, nous présentons dans cette section le cas d'une comparaison globale ($\mathrm{C}' = \mathrm{C}$).

3.1 Test de Hotelling pour la comparaison de deux groupes indépendants

Munissons l'espace $\mathcal{M}$ du repère orthonormé $(\mathrm{O}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$ et plaçons nous dans ce repère.

Le test de Hotelling d'écart nul, présenté au chapitre 2 (p.38) peut s'étendre à la comparaison de deux groupes indépendants $c1$ et $c2$, d'effectifs respectifs n_1 et n_2. Supposons que les nuages $\mathrm{M}^{I<c1>}$ et $\mathrm{M}^{I<c2>}$, de points moyens G^1 et G^2 et de matrices de covariance $\mathbf{V}1$ et $\mathbf{V}2$, soient deux échantillons identiquement et indépendamment distribués de deux distributions gaussiennes multidimensionnelles à L dimensions, de points moyens Γ^1 et Γ^2 (inconnus) et de même structure de covariance $\mathbf{\Sigma}$.

Le vecteur $\overrightarrow{\Gamma^1\Gamma^2}$ est estimé par le vecteur $\overrightarrow{\mathrm{G}^1\mathrm{G}^2}$. Sous le modèle multinormal et en supposant l'homogénéité de la structure de covariance des deux groupes, la matrice $\mathbf{\Sigma}$ est estimée par la matrice $\mathbf{S}$ (estimateur sans biais) :

$$\mathbf{S} = \frac{n_1\mathbf{V}1 + n_2\mathbf{V}2}{n_1 + n_2 - 2}$$

Remarque : Nous sortons ici du cadre combinatoire pour nous placer dans le cadre fréquentiste.

Notons $\overrightarrow{d}$ la variable *Vecteur–Écart* dont une réalisation est $\overrightarrow{\mathrm{G}^1\mathrm{G}^2}$, $\mathbf{d}$ la variable–vecteur (coordonnées) associée et $\mathbf{d}_\Gamma$ le vecteur des coordonnées de $\overrightarrow{\Gamma^1\Gamma^2}$. De même, notons S la variable *Covariance* dont une réalisation est $\mathbf{S}$. Les propriétés d'échantillonnage sont les suivantes (*cf.* par exemple Le Roux & Rouanet, 2004, p.326 [50] ou Saporta, 2006, p.347–348 [78]) :

– $\mathbf{d}$ et S sont statistiquement indépendants,

- $\mathbf{d} \sim \mathcal{N}(\mathbf{d}_\Gamma, \mathbf{\Sigma} \times (\frac{1}{n_1} + \frac{1}{n_2}))$: la distribution de $\mathbf{d}$ est une distribution gaussienne multidimensionnelle à L dimensions de moyenne $\mathbf{d}_\Gamma$ et de matrice de covariance $\mathbf{\Sigma} \times (\frac{1}{n_1} + \frac{1}{n_2})$,
- $(n_1 + n_2 - 2)\mathbf{S} \sim \mathcal{W}_L(n_1 + n_2 - 2, \mathbf{\Sigma})$: la distribution de $(n_1 + n_2 - 2)\mathbf{S}$ est une distribution de Wishart à L dimensions et à $n_1 + n_2 - 2$ degrés de liberté ($\mathbf{\Sigma}$ est définie positive).

Seuil observé du test. Sous le modèle multinormal–Whishart, testons l'hypothèse nulle $\mathcal{H}_0 : \Gamma^1 = \Gamma^2$ à l'aide du test de Hotelling.

Soit $D^2_{\mathbf{S}}$ la statistique $\mathbf{S}$–norme de $\overrightarrow{d}$ (au carré), c'est–à-dire la statistique définie par :

$$D^2_{\mathbf{S}} = |\overrightarrow{d}|^2_{\mathbf{S}}$$

La valeur observée de la statistique $D^2_{\mathbf{S}}$ est $d^2_{\mathbf{S}\ obs} = |\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|^2$. On a la propriété suivante :

> Sous $\mathcal{H}_0$, la distribution de la statistique $T^2 = \frac{n_1 n_2}{n_1+n_2} D^2_{\mathbf{S}}$ est celle de $\frac{L}{k} F_{L,n_1+n_2-1-L}$ avec $k = (n_1 + n_2 - 1 - L)/(n_1 + n_2 - 2)$ (F usuel de Fisher–Snedecor centré avec L et $n_1 + n_2 - 1 - L$ degrés de liberté).

Le seuil observé du test de Hotelling pour la comparaison de deux groupes est la probabilité d'échantillonnage que la statistique T^2 dépasse sa valeur observée $t^2_{obs} = \frac{n_1 n_2}{n_1+n_2} \times d^2_{\mathbf{S}\ obs}$. Soit $F[\alpha]$ la valeur critique de la distribution F_{L,n_1+n_2-1-L} au seuil supérieur α. Si $\frac{k}{L} t^2_{obs} = \frac{k}{L} \frac{n_1 n_2}{n_1+n_2} \times d^2_{\mathbf{S}\ obs} \geq F[\alpha]$, alors le résultat du test est significatif au seuil α.

Cas particulier d'un nuage unidimensionnel : test de Student. Supposons que les protocoles numériques $x^{I<c1>}$ de taille n_1 et $x^{I<c2>}$ de taille n_2, de moyennes m^1 et m^2 et de variances $v1$ et $v2$, soient deux échantillons identiquement et indépendamment distribués de deux distributions gaussiennes, de moyennes respectives μ^1 et μ^2 inconnues et de même variance σ^2. Le test de Student (deux groupes indépendants) permet de tester l'hypothèse nulle $\mathcal{H}_0 : \mu^1 = \mu^2$.

En prenant comme statistique de test le rapport :

$$t = \sqrt{n_1 + n_2 - 2} \frac{D}{\sqrt{(n_1 v1^2 + n_2 v2^2)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

où D est la statistique *Différence*, alors sous $\mathcal{H}_0$, t est distribué selon une loi $t_{n_1+n_2-2}$ (loi de Student à $n_1 + n_2 - 2$ degrés de liberté) (*cf.* par exemple Saporta, 2006, p.341 [78]).

Soit $t_{obs} = \sqrt{n_1 + n_2 - 2} \frac{m^1 - m^2}{\sqrt{(n_1 v1^2 + n_2 v2^2)(\frac{1}{n_1} + \frac{1}{n_2})}}$ la valeur observée de la statistique de test. Le seuil observé unilatéral est alors défini comme étant la probabilité, sous $\mathcal{H}_0$, que la statistique de test t soit plus extrême que la statistique observée t_{obs}.

3.2 Tests basés sur le modèle normal pour la comparaison de plus de deux groupes

Nous comparons C (> 2) groupes et nous supposons que les C nuages $(\mathrm{M}^{I<c>})_{c \in C}$ sont des échantillons identiquement et indépendamment distribués de distributions gaussiennes multidimensionnelles à L dimensions de même structure de covariance.

Plusieurs tests multidimensionnels, basés sur des statistiques de test différentes peuvent être mis en place dans ce cadre. Le plus couramment utilisé est le test du Lambda de Wilks (adaptation multidimensionnelle du test du rapport des vraisemblances usuel). Cependant, nous pouvons également citer les tests de la trace de Lawley–Hotelling, de la trace de Pillai et de la plus grande racine de Roy (*cf.* par exemple Morrison, 1967, chapitre 5 [61]).

Dans le cas unidimensionnel et sous le modèle normal, le test de Fisher (test du rapport des vraisemblances) permet de comparer les C moyennes (*cf.* par exemple Tenenhaus, 2007, p.53 [84]).

3.3 Test du bootstrap

Le principe du bootstrap est rappelé au chapitre 2 (p.41).

Supposons maintenant que les C nuages $(\mathrm{M}^{I<c>}, n_c)_{c\in C}$ de points moyens $(\mathrm{G}^c)_{c\in C}$, soient des échantillons indépendamment et identiquement distribués de C populations inconnues de points moyens $(\Gamma^c)_{c\in C}$ inconnus. Nous nous posons la question suivante : *L'hypothèse « $\Gamma^1 = \Gamma^2 = \ldots = \Gamma^{\mathrm{C}}$ » est-elle compatible avec les données, ou non ?*

Pour répondre à cette question, le test du bootstrap est construit de la façon suivante :

1. On construit les C nuages $(\mathrm{M}^{I(c)})_{c\in C}$ tels que pour un c donné et pour tout i de $I<c>$:

   $$\mathrm{M}^{i(c)} = \mathrm{M}^{i<c>} - \overrightarrow{\mathrm{OG}^c}$$

   les C nuages $(\mathrm{M}^{I<c>})_{c\in C}$ sont donc translatés de telle sorte que leurs points moyens soient le point O : définition du nuage intra pointé en O.

   Remarque : On se place ici sous l'hypothèse « $\Gamma^1 = \Gamma^2 = \ldots = \Gamma^{\mathrm{C}}$ » (hypothèse nulle).

2. On tire B échantillons bootstrap de taille $(n_c)_{c\in C}$ (tirages avec remise) à partir de chacun des C sous–nuages $(\mathrm{M}^{I(c)})_{c\in C}$. Soit $B = \{1, \ldots, b, \ldots \mathrm{B}\}$ l'ensemble indexant les B échantillons bootstrap. Notons M^{*cb}, le sous–nuage de n_c points constituant le b–ème échantillon bootstrap obtenu à partir du sous–nuage $\mathrm{M}^{I(c)}$. L'espace des C × B sous–nuages bootstrap M^{*cb} est noté $\mathcal{B}$. Pour un b donné, le nuage constitué de la réunion des C sous–nuages M^{*cb} est noté M^{*b}.

   (a) On considère l'application H^* qui au sous–nuage M^{*cb} associe son point moyen H^{*cb}.

   $$\begin{array}{rccc} H^* : & \mathcal{B} & \to & \mathcal{M} \\ & \mathrm{M}^{*cb} & \mapsto & \mathrm{H}^{*cb} \end{array}$$

   (b) On fait choix d'une statistique de test, ici la statistique V^*, définie de la façon suivante :

   $$\begin{array}{rccc} V^* : & \mathcal{M} & \to & \mathbb{R}^+ \\ & \mathrm{H}^{*Cb} & \mapsto & \sum\limits_{c\in C} \frac{n_c}{N} \, |\overrightarrow{\mathrm{OH}^{*cb}}|^2_{*\mathbf{W}b} \end{array}$$

   où $|\ .\ |_{*\mathbf{W}b}$ est la norme de Mahalanobis attachée au nuage intra construit à partir du nuage M^{*b}.

   Soit $v^*_{obs} = \sum\limits_{c\in C} \frac{n_c}{N} \, |\overrightarrow{\mathrm{OG}^c}|^2_{\mathbf{W}}$, où $|\ .\ |_{\mathbf{W}}$ est la norme de Mahalanobis attachée au intra construit à partir du nuage M^I.

3. On détermine la proportion des nuages M^{*b} pour lesquels la valeur de la statistique V^* est supérieure ou égale à v^*_{obs}. Cette proportion définit le seuil observé du test du bootstrap et est notée p^*_{obs}, avec $p^*_{obs} = p(V^* \geq v^*_{obs})$.

   Si $p^*_{obs} \leq \alpha$ (seuil α fixé), le résultat du test est significatif au seuil α. Pour la statistique V^*, on peut dire que les données sont en faveur de l'hypothèse « $\Gamma^1 = \Gamma^2 = \ldots = \Gamma^{\mathrm{C}}$ ».

   Si $p^*_{obs} > \alpha$, le résultat du test n'est pas significatif au seuil α. Pour la statistique V^*, on ne peut pas dire que les données sont en faveur de l'hypothèse « $\Gamma^1 = \Gamma^2 = \ldots = \Gamma^{\mathrm{C}}$ ».

L'inconvénient principal du test du bootstrap présenté ici est le suivant : la norme utilisée pour le calcul de la statistique de test change pour chaque échantillon bootstrap. La matrice de covariance du nuage intra construit à partir du nuage M^{*b} doit donc être inversée pour chaque échantillon bootstrap. Or, il est impossible de s'assurer de l'inversibilité de cette matrice. Il est donc fort probable, surtout pour des petites tailles de données, d'obtenir des cas de non inversibilité et de ne pas pouvoir calculer la statistique de test. Cependant, en s'assurant de la condition d'inversibilité pré–citée, le test du bootstrap présente une bonne alternative au test combinatoire lorsque le principe d'échangeabilité n'est pas satisfait (*cf.* par exemple Pesarin & Salmaso, 2010 [64] ; ter Braak, 1992 [85]).

Le test du bootstrap est illustré par l'exemple des races canines, p.121.

4 Applications : exemple des races canines

Le jeu de données utilisé est celui déjà présenté au cours du chapitre 3 (p.70) dans lequel 27 races canines sont décrites par les six variables catégorisées suivantes : « taille », « poids », « vélocité », « intelligence », « affection » et « agressivité ». Les 27 races canines sont réparties en trois groupes d'intérêt selon leur « fonction » :

- Groupe 1 : « compagnie »,
- Groupe 2 : « chasse »,
- Groupe 3 : « utilité ».

Une ACM est effectuée (*cf.* p.70), nous nous intéressons ici aux données de la table 3.2 (p.71) des coordonnées des trois groupes dans le premier plan principal.

L'ensemble I est composé des 27 races canines, nous lui associons le nuage M^I de 27 points dans le premier plan principal. Le nuage M^I, son point moyen O et son ellipse indicatrice ($\kappa = 1$) sont représentés sur la figure 4.18 suivante.

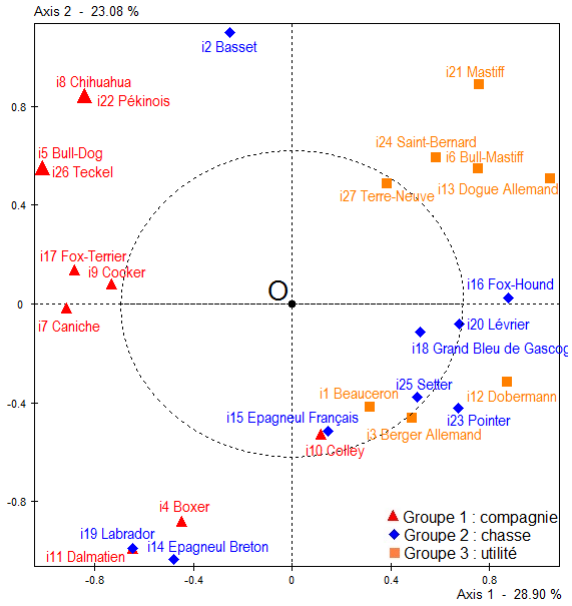


FIGURE 4.18 – *Exemple des Races canines.* Nuage M^I des 27 races canines et son ellipse indicatrice ($\kappa = 1$) dans premier plan principal.

Soient $\mathrm{M}^{I<c1>}$, $\mathrm{M}^{I<c2>}$ et $\mathrm{M}^{I<c3>}$ les sous–nuages composés des $n_1 = 10$, $n_2 = 9$ et $n_3 = 8$ points des groupes 1 (« compagnie »), 2 (« chasse ») et 3 (« utilité »), leurs points moyens respectifs sont notés G^1, G^2 et G^3 (*cf.* figure 4.19).

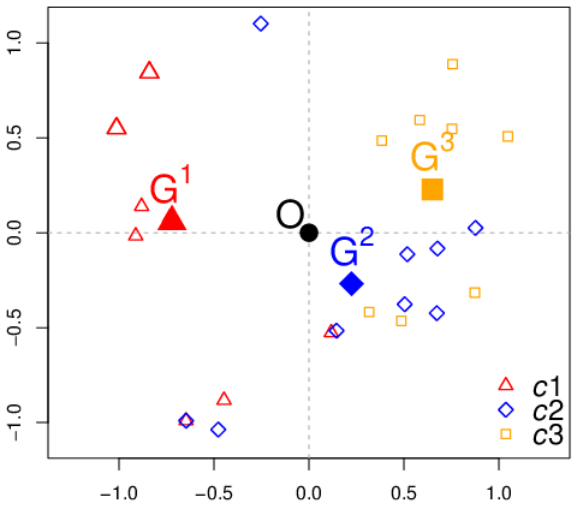


FIGURE 4.19 – *Exemple des Races canines.* Partition du nuage M^I en 3 sous–nuages et points moyens associés.

La matrice de covariance du nuage M^I, notée $\mathbf{V}$, est diagonale et a pour éléments diagonaux les variances des deux premiers axes, c'est–à–dire $\lambda_1 = 0.481$ et $\lambda_2 = 0.385$.

Le nuage intra est représenté sur la figure ci–après, la matrice $\mathbf{W} = \begin{pmatrix} 0.147 & -0.008 \\ -0.008 & 0.344 \end{pmatrix}$ est sa matrice de covariance.

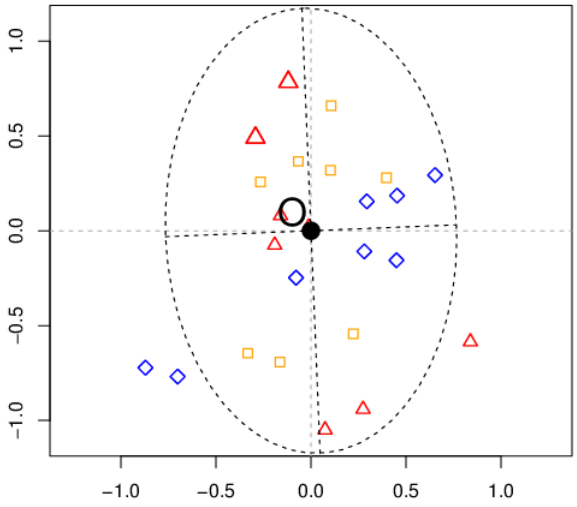


FIGURE 4.20 – *Exemple des Races canines.* Nuage intra et son ellipse de concentration ($\kappa = 2$).

Les vecteurs $\overrightarrow{\mathrm{OG}^1}$, $\overrightarrow{\mathrm{OG}^2}$ et $\overrightarrow{\mathrm{OG}^3}$ ont pour coordonnées respectives : $\begin{pmatrix}-0.721\\0.059\end{pmatrix}$, $\begin{pmatrix}0.224\\-0.268\end{pmatrix}$ et $\begin{pmatrix}0.650\\0.228\end{pmatrix}$. On a $|\overrightarrow{\mathrm{OG}^1}|^2 = \frac{(-0.721)^2}{0.481} + \frac{0.059^2}{0.385} = 1.089$, $|\overrightarrow{\mathrm{OG}^2}|^2 = 0.291$ et $|\overrightarrow{\mathrm{OG}^3}|^2 = 1.012$ (carrés des M–distances).

Les vecteurs $\overrightarrow{\mathrm{G}^1\mathrm{G}^2}$, $\overrightarrow{\mathrm{G}^2\mathrm{G}^3}$ et $\overrightarrow{\mathrm{G}^1\mathrm{G}^3}$ ont pour coordonnées respectives : $\begin{pmatrix}0.945\\-0.327\end{pmatrix}$, $\begin{pmatrix}0.426\\0.496\end{pmatrix}$ et $\begin{pmatrix}1.371\\0.169\end{pmatrix}$. Et on a $|\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|^2 = 2.133$, $|\overrightarrow{\mathrm{G}^1\mathrm{G}^3}|^2 = 3.978$ et $|\overrightarrow{\mathrm{G}^2\mathrm{G}^3}|^2 = 1.017$ (carrés des M–distances).

4.1 Test d'homogénéité exact

4.1.1 Comparaison globale

Nous nous proposons de répondre à la question suivante : *les groupes 1 (« compagnie »), 2 (« chasse ») et 3 (« utilité ») sont–ils hétérogènes ?*

Ensemble et espace des emboîtements. Pour cet exemple, le nombre combinatoire $\frac{27!}{10!\ 9!\ 8!} = 2.05 \times 10^{11}$ est trop important pour que l'ensemble des emboîtements « complet » soit construit. Nous utilisons la méthode de Monte Carlo afin d'engendrer un ensemble « restreint » d'emboîtements. Pour cet exemple, nous choisissons de considérer un ensemble de $\mathrm{J} = 50\,000$ emboîtements (50 000 affectations possibles des 27 individus dans les 3 groupes [13]). A cet ensemble « restreint », nous associons l'*espace des emboîtements* correspondant (*cf.* paragraphe 1.1.1, p.85).

Seuil observé exact. Au j–ème emboîtement de l'espace des emboîtements « restreint » sont associés les 3 points moyens H^{1j}, H^{2j} et H^{3j} à partir desquels est calculée la statistique de test V_{M} (*cf.* paragraphe 1.1.1, p.85). Sa distribution est donnée sur la figure 4.21.

13. Pour des raisons de temps de calcul, nous ne vérifions pas le fait que les affectations engendrées soient différentes. En effet, la probabilité d'obtenir plusieurs fois la même affectation est très faible.

La valeur observée $v_{\mathrm{M}\,obs}$ de la statistique V_{M}, calculée en fonction des carrés des M–distances entre les couples de points, vaut :

$$v_{\mathrm{M}\,obs} = \frac{10 \times 9}{27^2}|\overrightarrow{\mathrm{G}^1\mathrm{G}^2}|^2 + \frac{9 \times 8}{27^2}|\overrightarrow{\mathrm{G}^2\mathrm{G}^3}|^2 + \frac{10 \times 8}{27^2}|\overrightarrow{\mathrm{G}^1\mathrm{G}^3}|^2 = 0.8003$$

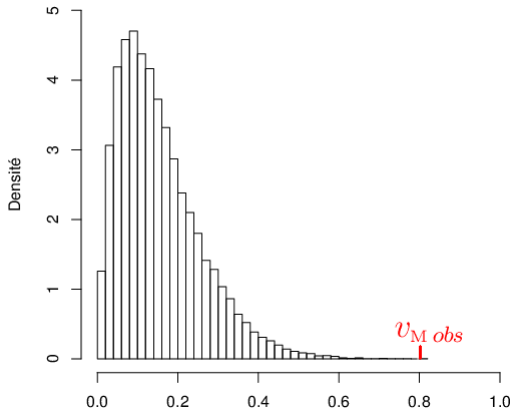


FIGURE 4.21 – *Exemple des Races canines, comparaison globale.* Distribution de la statistique V_{M} ($v_{\mathrm{M}\,obs} = 0.8003$).

Parmi les J = 50 000 emboîtements, il y en a 3 pour lesquels la valeur de la statistique V_{M} est supérieure ou égale à la valeur observée $v_{\mathrm{M}\,obs}$. D'où $p_{obs} = \frac{3}{50\,000} = 6 \times 10^{-5}$, résultat significatif au seuil .05 [14].

Conclusion : Pour la statistique M–*Variance*, nous pouvons dire que les trois groupes sont hétérogènes [15] au seuil .05.

4.1.2 Comparaisons associées à des contrastes orthogonaux

Nous souhaitons maintenant préciser où se trouvent les différences. Au vu des écarts descriptifs entre les points moyens ($|\overrightarrow{\mathrm{G}^1\mathrm{G}^2}| = 1.460$, $|\overrightarrow{\mathrm{G}^1\mathrm{G}^3}| = 1.994$, $|\overrightarrow{\mathrm{G}^2\mathrm{G}^3}| = 1.008$), les points G^2 et G^3 étant les plus proches, il est naturel d'effectuer dans un premier temps la comparaison partielle des groupes 2 et 3, c'est–à–dire la comparaison associée au contraste $(0, 1, -1)$.

Comparaison partielle des groupes 2 et 3.

Ensemble et espace des emboîtements. En raison de l'explosion combinatoire, les $\frac{27!}{10!\ 9!\ 8!}$ emboîtements ne peuvent être construits. Comme pour le test global, nous considérons donc un ensemble des emboîtements « restreint » de J = 50 000 emboîtements et l'espace des emboîtements associé (méthode de Monte Carlo).

14. Afin de s'assurer de la stabilité du seuil observé obtenu par la méthode de Monte Carlo, nous avons effectué 10 simulations indépendantes dont les seuils observés sont donnés dans le tableau suivant :

p_{obs}	2×10^{-5}	0	4×10^{-5}	0	4×10^{-5}	4×10^{-5}	2×10^{-5}	4×10^{-5}	6×10^{-5}	6×10^{-5}

15. Les seuils observés des tests du Lambda de Wilks, de la trace de Lawley–Hotelling, de la trace de Pillai et de la plus grande racine de Roy sont tous inférieurs à .0001 (tests effectués avec le logiciel SAS©).

Seuil observé exact. Au j–ème emboîtement de l'espace des emboîtements « restreint » sont associés les 2 points moyens H^{2j} et H^{3j} et au bipoint $(\mathrm{H}^{2j}, \mathrm{H}^{3j})_{j\in J}$ est associé le vecteur–écart $\overrightarrow{d^j} = \mathrm{H}^{3j} - \mathrm{H}^{2j}$. Le nuage D^J, formé des $\mathrm{J} = 50\,000$ points D^j définis par $\mathrm{D}^j = \mathrm{O} + \overrightarrow{d^j}$ est construit. Il permet, dans le cas de deux groupes, d'interpréter géométriquement le seuil observé exact (*cf.* paragraphe 2.1.1, p.95). Notons $\mathrm{D}_{obs} = \mathrm{O} + \overrightarrow{d_{obs}}$ où $\overrightarrow{d_{obs}} = \mathrm{G}^3 - \mathrm{G}^2$. Le nuage D^J et la distribution de la statistique de test D_{M} sont représentés ci–après.

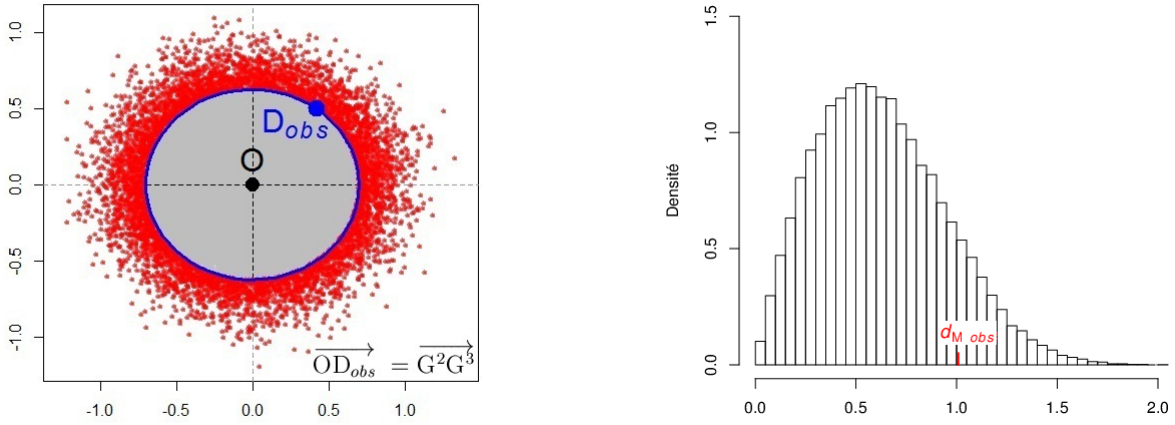


FIGURE 4.22 – *Exemple des Races canines, comparaison partielle des groupes 2 et 3.* Nuage D^J et ellipse d'inertie du nuage M^I passant par D_{obs} (gauche) et distribution de la statistique D_{M} ($d_{\mathrm{M}\,obs} = 1.008$) (droite).

Parmi les 50 000 points du nuage D^J, 6303 vérifient $|\overrightarrow{\mathrm{OD}^j}| \geq |\overrightarrow{\mathrm{OD}_{obs}}|$, ils se situent sur ou à l'extérieur de l'ellipse d'inertie du nuage M^I passant par D_{obs} (en rouge à gauche de la figure 4.22). On a donc $p_{obs} = \frac{6303}{50\,000} = .126$, résultat non significatif au seuil .05 [16].

Zone de compatibilité. En mettant en oeuvre la procédure de construction de la zone de compatibilité au seuil $1-\alpha$ décrite dans le paragraphe 2.1.2 (p.100), nous avons recherché un ensemble de points P juste incompatibles avec le point O au seuil $\frac{2\,500}{50\,000} = 0.05$. Les points obtenus sont représentés en rouge sur la figure 4.23 [17]. Sur la même figure est représentée en bleu l'ellipse ajustée à cet ensemble de points (la méthode d'ajustement est la même que celle décrite en bas de la page 104). Il s'agit de l'ellipse de centre O, définie par l'ensemble des points Q tels que $|\overrightarrow{\mathrm{OQ}}|_{\mathbf{Y}} = 1.349$ [18] (*cf.* théorème 2.1, p.102). La zone de compatibilité au seuil .95 est donc définie, aux tracas du discret près, par l'ensemble des points situés à l'intérieur de cette ellipse.

16. Afin de s'assurer de la stabilité du seuil observé obtenu par la méthode de Monte Carlo, nous avons effectué 10 simulations indépendantes dont les seuils observés sont donnés dans le tableau suivant :

p_{obs}	0.126	0.125	0.126	0.126	0.1274	0.125	0.126	0.125	0.124	0.126

17. *cf.* note sur le calcul p.104. Quadrillage horizontal : -0.8 → 1 par pas de 0.05, quadrillage vertical : -0.8 → 0.8 par pas de 0.05.

18. $\mathbf{Y}$ est telle que $\mathbf{Y} = \mathbf{W} + (\frac{9}{27} + \frac{8}{27}) \times \frac{10}{27}\ \mathbf{dd}'$ avec $\mathbf{d}$ vecteur–colonne associé aux coordonnées de $\overrightarrow{\mathrm{G}^{23}\mathrm{G}^1}$ (G^{23} barycentre des points G^2 et G^3).

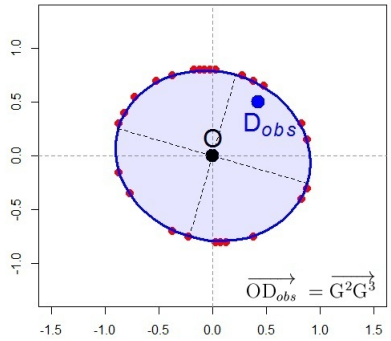


FIGURE 4.23 – *Exemple des Races canines, comparaison partielle des groupes 2 et 3.* Zone de compatibilité au seuil .95 ($\kappa = 1.349$, en bleu).

Le point D_{obs} est situé à l'intérieur de la zone de compatibilité, les points O et D_{obs} sont compatibles au seuil .05.

Conclusion : Pour la statistique M–*Distance*, nous ne pouvons pas dire que les deux groupes « chasse » et « utilité » sont hétérogènes au seuil .05.

Comparaison entre la réunion des groupes 2 et 3 et le groupe 1.

Les groupes 2 et 3 ne pouvant pas être considérés comme étant hétérogènes, il est naturel de les regrouper et d'effectuer la comparaison associée au contraste $(-1, \frac{n_2}{n_2+n_3}, \frac{n_3}{n_2+n_3})$, c'est–à–dire d'effectuer la comparaison entre le regroupement des groupes 2 et 3 et le groupe 1. Ce contraste est orthogonal au contraste $(0, 1, -1)$ étudié précédemment.

Nous obtenons le seuil : $p_{obs} = 2.5 \times 10^{-5}$, résultat significatif au seuil .05.

En conclusion, les deux comparaisons à 1 degré de liberté (binaires) associées aux contrastes orthogonaux $(0, 1, -1)$ et $(-1, \frac{n_2}{n_2+n_3}, \frac{n_3}{n_2+n_3})$ permettent de dire que la significativité du test comparant les trois groupes ($p_{obs} = 6 \times 10^{-5}$) est essentiellement due aux différences entre les groupes 1 et 2 et entre les groupes 1 et 3.

4.2 Test d'homogénéité approché

4.2.1 Comparaison globale

On a $s^2_{\mathrm{M}\,obs} = \frac{v_{\mathrm{M}\,obs}}{\frac{1}{27-1}} = 0.8003 \times 26 = 20.808$ et $L(\mathrm{C}' - 1) = 2 \times 2 = 4$ (*cf.* p.94), d'où le seuil observé approché : $\widetilde{p}_{obs} = p(\chi^2_4 \geq 20.808) = .00035$ (à comparer avec le seuil exact $p_{obs} = \frac{3}{50\,000} = 6 \times 10^{-5}$).

Le test approché conduit à la même conclusion que le test exact.

A titre illustratif, nous avons représenté sur la figure 4.24 la distribution de la statistique $\mathrm{S}^2_{\mathrm{M}}$ (ici $\mathrm{S}^2_{\mathrm{M}} = V_{\mathrm{M}} \times 26$). A cette distribution est superposée celle du χ^2 à 4 degrés de liberté (en rouge).

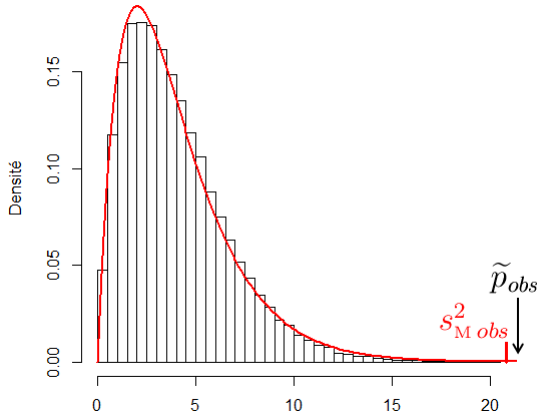


FIGURE 4.24 – *Exemple des Races canines, comparaison globale.* Distribution de la statistique $\mathrm{S}^2_{\mathrm{M}}$ ($s^2_{\mathrm{M}\,obs} = 20.808$) et du χ^2 à 4 degrés de liberté (en rouge).

4.2.2 Comparaisons associées à des contrastes orthogonaux

En utilisant le test approché, nous effectuons maintenant la comparaison partielle des groupes 2 et 3 puis la comparaison globale de la réunion des groupes 2–3 et du groupe 1 (comme au paragraphe précédent).

Comparaison partielle des groupes 2 et 3.

Seuil observé approché. On a $\mathrm{t}^2_{obs} = \frac{27-1}{27} \times \frac{1}{\frac{1}{9}+\frac{1}{8}} \times d^2_{\mathrm{M}\,obs} = \frac{26}{27} \times \frac{1}{\frac{1}{9}+\frac{1}{8}} \times 1.017 = 4.147$ et $L = 2$ (*cf.* p.109), d'où le seuil observé approché : $\widetilde{p}_{obs} = p(\chi^2_2 \geq 4.147) = .1257$ (à comparer avec le seuil observé exact $p_{obs} = \frac{6303}{50\,000} = .1261$).

Le test approché conduit à la même conclusion que le test exact.

A titre illustratif, nous avons représenté sur la figure 4.25 la distribution de la statistique T^2 (ici $\mathrm{T}^2 = \frac{27-1}{27} \times \frac{1}{\frac{1}{9}+\frac{1}{8}} \times D^2_{\mathrm{M}}$). A cette distribution est superposée celle du χ^2 à 2 degrés de liberté (en rouge).

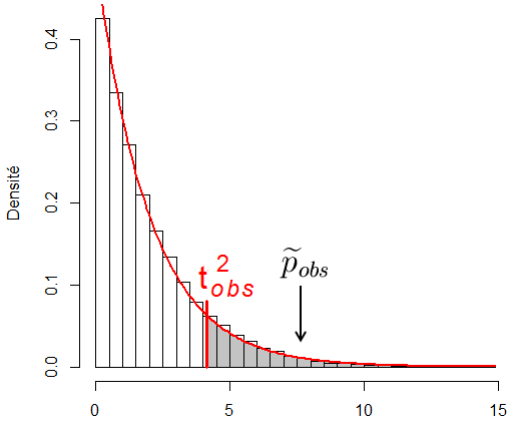


FIGURE 4.25 – *Exemple des Races canines, comparaison partielle des groupes 2 et 3.* Distributions de la statistique T^2 ($\mathrm{t}^2_{obs} = 4.147$) et du χ^2 à 2 degrés de liberté (en rouge).

Zone de compatibilité approchée. On a $\chi^2_2[.05] = 5.991$, d'après le théorème 2.3 (p.110), la zone–limite approchée de compatibilité est l'ellipse de centre O, définie par l'ensemble des points P tels que $|\overrightarrow{\mathrm{OP}}|_{\mathbf{Y}} = \sqrt{\frac{\frac{27}{27-1}\times(\frac{1}{9}+\frac{1}{8})\times 5.991}{1-(\frac{9}{27}+\frac{8}{27})\times\frac{9}{27}\times\frac{8}{27}\times\frac{27}{27-1}\times(\frac{1}{9}+\frac{1}{8})\times 5.991}} = 1.272$ (à comparer avec $\kappa = 1.349$ obtenu ci–avant).

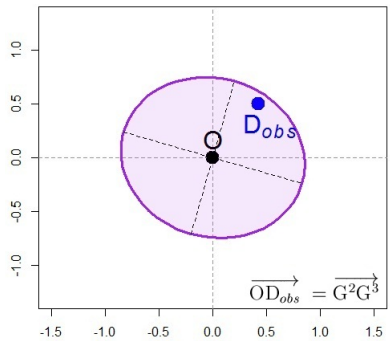


FIGURE 4.26 – *Exemple des Races canines, comparaison partielle des groupes 2 et 3.* Zone de compatibilité approchée au seuil .95 ($\kappa = 1.272$, en violet).

Le point D_{obs} est situé à l'intérieur de la zone de compatibilité, les points O et D_{obs} sont compatibles au seuil .05.

Comparaison entre la réunion des groupes 2 et 3 et le groupe 1.

Le seuil du test approché comparant la réunion des groupes 2 et 3 et le groupe 1 est égal à $\widetilde{p}_{obs} = 2.4 \times 10^{-4}$, résultat significatif au seuil .05.

En conclusion, les deux comparaisons à 1 degré de liberté (binaires) associées aux contrastes orthogonaux $(0, 1, -1)$ et $(-1, \frac{n_2}{n_2+n_3}, \frac{n_3}{n_2+n_3})$ permettent de dire que la significativité du test approché comparant les trois groupes ($\widetilde{p}_{obs} = .00035$) est essentiellement due aux différences entre les groupes 1 et 2 et entre les groupes 1 et 3.

4.3 Test du bootstrap

Nous effectuons ici le test du bootstrap dans le cas de la comparaison globale.

Nous choisissons de tirer $\mathrm{B} = 1\,000$ échantillons bootstrap. Parmi les 1 000 nuages M^{*b} construits (*cf.* p.113), il y en a 1 pour lequel la valeur de la statistique V^* est supérieure ou égale à la valeur observée $v^*_{obs} = \frac{10}{27}\,|\overrightarrow{\mathrm{OG}^1}|^2_{\mathbf{W}} + \frac{9}{27}\,|\overrightarrow{\mathrm{OG}^2}|^2_{\mathbf{W}} + \frac{8}{27}\,|\overrightarrow{\mathrm{OG}^3}|^2_{\mathbf{W}} = 2.398$ (la distribution de la statistique de test V^* est donnée sur la figure 4.27). D'où le seuil observé du test du bootstrap [19] : $p^*_{obs} = \frac{1}{1\,000} = 0.001$.

Le test du bootstrap conduit à la même conclusion que le test exact pour lequel $p_{obs} = 6 \times 10^{-5}$.

19. Afin de s'assurer de la stabilité des résultats obtenus par bootstrap, nous avons effectué 10 tests indépendants dont les seuils observés sont donnés dans le tableau suivant :

p_{obs}	0.001	0.002	0.000	0.001	0.003	0.000	0.002	0.001	0.001	0.000

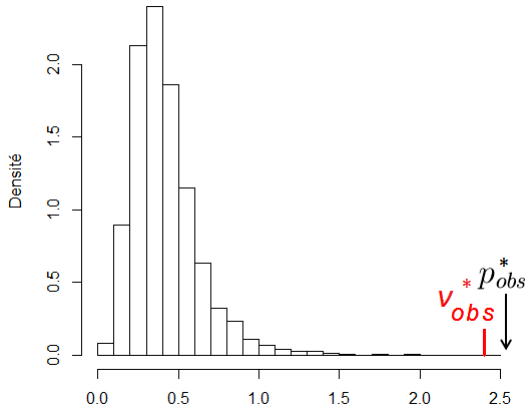


FIGURE 4.27 – *Exemple des Races canines, comparaison globale.* Distribution de la statistique $V*$ ($v^*_{obs} = 2.398$).

Chapitre 5

L'absentéisme dans les IEG de 1995 à 2011 : étude statistique de la cohorte EPIEG

Introduction

Rappelons que cette thèse a été effectuée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE), mise en place entre le Centre de Recherche en Mathématiques de la Décision (CEREMADE) et le Service Général de Médecine de Contrôle (SGMC) des Industries Electriques et Gazières (IEG).

Les salariés des Industries Electriques et Gazières bénéficient d'un « statut » particulier qui leur confère notamment l'assurance d'un régime spécial de sécurité sociale [1]. Ce dernier leur permet de continuer à percevoir un plein salaire dès le premier jour de l'arrêt médical, ceci pouvant se prolonger pendant 5 ans en cas de longue maladie. Les médecins conseils du SGMC sont les garants médico–administratifs de ces prestations, ils ont pour mission principale de vérifier systématiquement le bien fondé de l'absentéisme médical, de décider le placement en longue maladie (LM) ou en invalidité, d'évaluer les séquelles en vue de la réparation des accidents du travail (AT) et des maladies professionnelles (MP) et de défendre l'intérêt des entreprises et des agents dans les accidents avec tiers. Parallèlement à cette activité « métier », le SGMC dispose également d'une cellule santé publique visant à promouvoir des actions favorables à la santé et d'une cellule épidémiologie ayant pour principale mission d'analyser la santé des salariés.

Ce doctorat a été effectué au sein de la cellule épidémiologie. L'activité du SGMC en épidémiologie a été initialement motivée par la singularité du positionnement du SGMC : un « service de sécurité sociale intégré en entreprise ». L'observation des événements de santé entraînant une absence médicale, une indemnisation ou une pension est étudiée pendant toute la période d'emploi de l'agent. La possibilité de mesurer correctement l'incidence des pathologies vient, d'une part, de la connaissance de la population étudiée, et d'autre part, de l'enregistrement de la survenue des pathologies et de leur codage par les médecins conseils. Une base de données épidémiologique est exploitée à la cellule épidémiologie depuis 1978 (*cf.* Goldberg & al., 1980 [40]). Cette dernière permet de décrire la santé de la population sala-

1. En 2012, la branche des IEG regroupe 156 entreprises de production, de transport, de distribution et de commercialisation de l'énergie en France. Les documents relatifs au statut des IEG (décret n° 46–1541 du 22 juin 1946) sont téléchargeables à l'adresse : http ://sgeieg.fr.

riée des IEG dès les années 1980, au moyen d'indicateurs d'absence médicale (*cf.* Chevalier & al., 1987 [14] [15] ; Godard, 1987 [37]) ou d'indicateurs divers relatifs aux accidents du travail (*cf.* Pouplet & al., 1987 [67] ; Luce & al., 1997 [56]), à la mise en longue maladie ou en invalidité (*cf.* Savelli & al., 1994 [80]), aux maladies professionnelles (*cf.* Coulondre & al., 1999 [19]), à la mortalité en activité (*cf.* Poncet & al., 2003 [66]), à l'incidence des cancers (*cf.* Chevalier & al., 1996 [16]) et des cardiopathies ischémiques (*cf.* Chevalier & al., 2001 [18]). Ces études ont pour première vocation de surveiller la santé, elles conduisent parfois à orienter la prévention (*cf.* Godard & al., 2007 [39]). La base de données alimente également les données d'absence relatives aux 20 000 volontaires de la cohorte professionnelle GAZEL depuis 1989 (*cf.* Goldberg & al., 1991 [41]).

Jusqu'en 2009, les études temporelles fournies par la cellule épidémiologie étaient essentiellement des études transversales répétées, effectuées sur données agrégées (*cf.* Godard (mémoire), 2009 [38] et Gault (thèse), 2009 [34]). Les effectifs étaient calculés à partir des statistiques du personnel (aggrégées), éditées chaque année par les services de ressources humaines. En effet, la cellule épidémiologie ne disposait pas de fichiers individuels administratifs harmonisés dans le temps. Pour pallier ce problème, la cellule épidémiologie entreprend en 2009 la création d'une cohorte, baptisée cohorte EPIEG (EPidémiologie dans les Industries Electriques et Gazières). Elle permet de contrôler précisément la population salariale au cours du temps et de suivre les évolutions socio–démographiques de chaque salarié.

Dans la suite de ce chapitre, nous décrivons d'abord la cohorte EPIEG en détaillant notamment les étapes de sa construction. Puis, nous exploitons ces données dans le but d'étudier la santé de la population salariale depuis 1995. Nous appliquons en particulier des méthodes d'analyse des données multidimensionnelles à divers tableaux de données issus de la cohorte EPIEG, que nous construisons spécialement pour étudier l'absentéisme de courte durée et de longue maladie. L'ensemble des analyses permet notamment d'identifier des groupes d'agents *sensibles* et des pathologies *émergentes*.

1 La cohorte EPIEG

A la suite d'un protocole de mise à jour des données, initié courant 2008, la cellule épidémiologie dispose, depuis décembre 2009, de fichiers annuels regroupant l'ensemble des données socio-démographiques des salariés présents en décembre des années 1995 à 2009, statutaires, et bénéficiant par conséquent du régime particulier de sécurité sociale des IEG [2]. Les fichiers regroupant l'ensemble des salariés ayant quittés les entreprises ainsi que le motif et la date de leurs départs ont également été fournis. C'est à partir de ces données que la cellule épidémiologie entreprend en 2009 la création de la cohorte EPIEG [3].

Un travail de gestion de données très conséquent fût nécessaire afin d'harmoniser les données au cours du temps, d'identifier les doublons, de corriger les valeurs erronées/aberrantes/manquantes, *etc* [4].

Les données des années 2010 et 2011 ont ensuite été ajoutées, elles proviennent des fichiers administratifs des entreprises et du nouveau système d'information du SGMC : Aramis. L'ensemble des sorties de l'emploi de 1995 à 2011 a également été contrôlé rigoureusement [5].

2. Contrôle de la sélection des salariés effectué en liaison avec la Direction Informatique et Telecom.

3. Les données de la cohorte EPIEG sont strictement confidentielles.

4. Construction de la base de données effectuée sur une période d'un an par Solène Bienaise et Catherine Godard.

5. Grâce à des fichiers complémentaires fournis par la CNIEG (Caisse Nationale des IEG).

Des études de cohorte peuvent dès lors être fournies : elles consistent à suivre dans le temps une population définie (la cohorte) et à étudier l'incidence d'un événement, par exemple la survenue d'un problème de santé donné.

Pour cette étude, nous considérons les données de la cohorte EPIEG relatives aux années 1995 à 2011 [6].

1.1 Déploiement vertical et concept de personnes/temps

L'étude de la survenue des événements de santé nécessite de connaître précisément la population à risque, c'est–à–dire l'ensemble des individus susceptibles de contracter l'événement de santé étudié sur une période donnée. Par conséquent, il est nécessaire d'introduire ici le concept de *personnes/temps* (*cf.* par exemple Rumeau–Rouquette & al., 1993 [77]) : il tient compte non seulement du nombre de personnes exposées au risque de contracter l'événement de santé étudié, mais aussi de la durée pendant laquelle elles sont présentes et exposées à ce risque. Le calcul exact de la durée de suivi d'une population consiste à sommer la durée de suivi pour chaque individu. Par exemple, si 50 personnes à risque sont suivies pendant 2 ans, la durée totale de suivi est de 100 personnes/années. Dans ce même exemple, s'il y a 5 nouveaux cas de l'événement de santé étudié, le taux d'incidence est de 5 cas par 100 personnes/années, ou plus simplement de 2.5 cas par 50 personnes/années (ou encore 0.05 cas par personne/année).

Pour le fichier de données de la cohorte EPIEG, comme il est classique de le faire pour les cohortes épidémiologiques, les années sont déployées verticalement. Chacune des lignes du fichier représente donc une personne présente, au moins un jour, pour une année donnée. Par exemple, les données d'un salarié présent durant les 17 années occupent 17 lignes du fichier (une ligne par année, déploiement vertical). La durée de suivi pour chaque individu est ensuite calculée. Elle est de 17 personnes/années pour le salarié de l'exemple précédent ; les données d'un salarié présent, lui, du 1er janvier 1995 au 30 juin 1998 occupent 4 lignes mais sa durée de suivi est de 3.5 personnes/années.

1.2 Population étudiée

La quasi–totalité des entreprises des IEG fournissent mensuellement au SGMC des fichiers comprenant les données administratives de leurs salariés. La population étudiée est l'ensemble des salariés ayant été statutaires des ces entreprises entre le 1er janvier 1995 et le 31 décembre 2011, soit 94% de l'effectif total des IEG entre ces deux dates.

Pour évaluer les effectifs de la cohorte d'une année donnée, nous pouvons considérer :
- soit le nombre de salariés statutaires présents au moins une fois durant l'année considérée,
- soit le nombre de personnes/années tenant compte du temps de présence dans l'année de chaque agent.

Pour les années 1995 à 2011, la cohorte EPIEG comprend 215 550 salariés présents au moins 1 jour de 1995 à 2011 (appelés par la suite cohortistes), soit 2 373 941.5 personnes/années (en tenant compte des entrées et sorties de l'emploi de chaque année). Pour l'année 1995, 145 310 salariés (comptant pour 141 370 personnes/années) sont inclus

6. La cohorte étant alimentée chaque année des nouvelles données socio–démographiques des salariés, la période d'étude pourra être élargie.

dans la cohorte. Ces effectifs ont tendance a augmenter jusqu'en 2000 (153 138 cohortistes / 147 243.08 personnes/années) pour finalement ne pas cesser de diminuer jusqu'en 2010, année pour laquelle la cohorte comprend 136 568 salariés (130 938 personnes/années). Nous pouvons toutefois observer une légère hausse des effectifs en 2011 : 137 144 salariés (131 625.08 personnes/années) (figure 5.1).

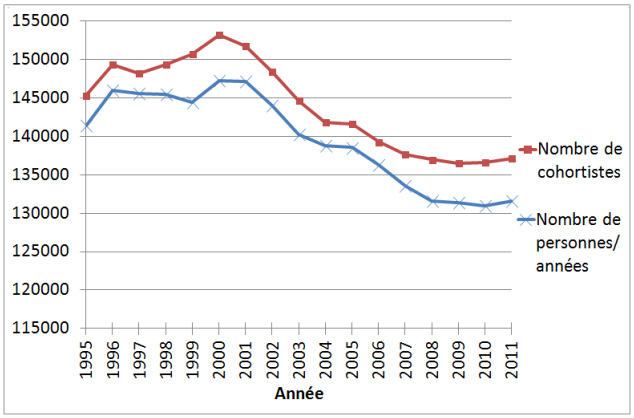


FIGURE 5.1 – Evolution du nombre de cohortistes et de personnes/années de 1995 à 2011.

Le nombre d'embauches et de sorties de l'emploi par an fluctue beaucoup de 1995 à 2011 (figure 5.2, gauche), nous observons notamment un grand nombre d'embauches en 1999 et 2000. Cette augmentation peut s'expliquer par le passage aux 35 heures mis en place en 2000 : des accords « temps de travail », visant à favoriser l'embauche, sont passés entre les organisations syndicales et les entreprises afin de compenser la réduction du temps de travail. L'effet de cette politique d'embauche s'estompe à partir de 2001. Nous constatons cependant une nouvelle augmentation à partir de 2008. Sauf pour les années 1999, 2000 et 2011, le nombre de départs est supérieur au nombre d'embauches ; ceci explique la diminution progressive des effectifs (*cf.* figure 5.1).

La répartition des motifs de sorties de l'emploi est, quant à elle, assez stable de 1995 à 2011 (figure 5.2, droite) : la majorité des agents quittant les IEG partent à la retraite ou sont mis en invalidité (entre 85% et 92%, selon les années). Les ruptures de contrat de travail et les décès en activité sont assez rares (respectivement entre 3% et 7% et entre 3% et 8%, selon les années).

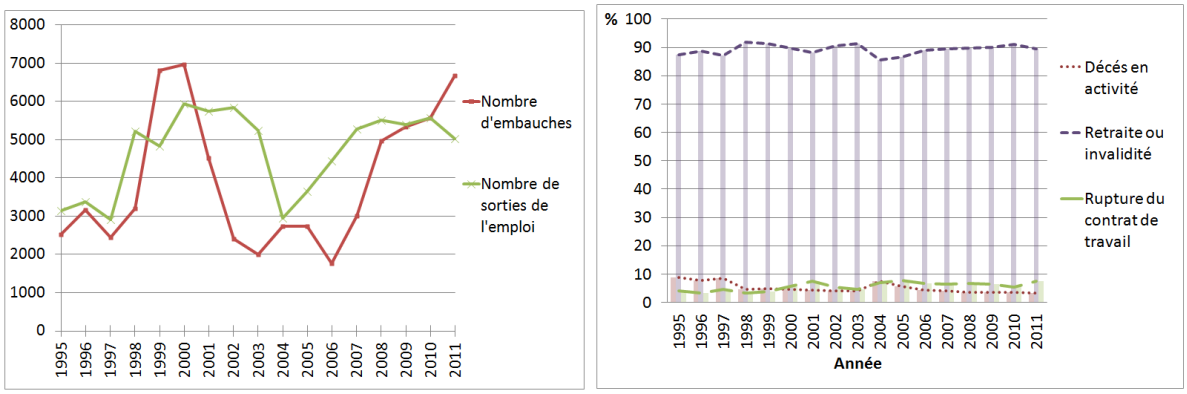


FIGURE 5.2 – A gauche : évolution du nombre d'embauches et de sorties de l'emploi de 1995 à 2011 ; à droite : évolution des motifs de sortie de l'emploi de 1995 à 2011 (en % des agents ayant quitté les IEG).

1.3 Evolution de la structure de la population

Les variables dont la fiabilité est vérifiée et qui comprennent moins de 5% de valeurs manquantes (de 1995 à 2011) sont retenues pour la cohorte EPIEG [7]. La totalité de la base est décrite en annexe (p.172). Nous distinguons :

- les variables « non évolutives », qui pour un salarié donné, ne sont pas susceptibles d'évoluer (ex : sexe, niveau de diplôme à l'embauche, *etc.*),
- les variables « évolutives », qui sont susceptibles d'évoluer (ex : situation familiale, classe d'âge, *etc.*).

Tris à plat. Plusieurs approches sont possibles pour présenter les statistiques élémentaires des données non évolutives. Nous pouvons :

- soit considérer l'ensemble des cohortistes (présents au moins 1 jour de 1995 à 2011) et ne pas tenir compte de l'année,
- soit considérer les cohortistes de chaque année (présents au moins 1 jour dans l'année considérée) et, dans ce cas, regarder des évolutions temporelles.

Intéressons nous par exemple à la répartition hommes/femmes ; nous pouvons présenter les deux graphiques suivants :

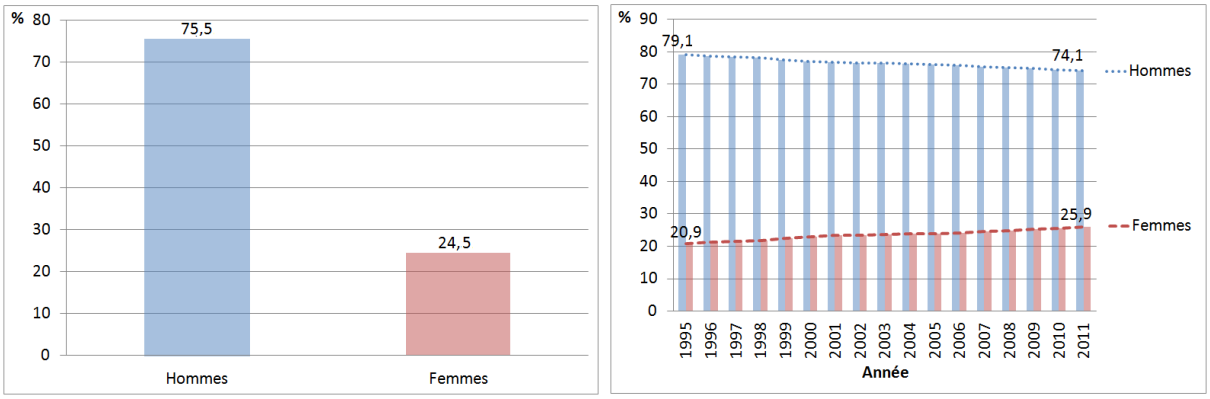


FIGURE 5.3 – A gauche : répartition hommes/femmes de l'ensemble des cohortistes ; à droite : évolution de la répartition hommes/femmes de 1995 à 2011.

La question ne se pose pas pour les données évolutives qui, par définition, doivent être étudiées année par année [8].

• *Sexe.* 75.5% des 215 550 cohortistes sont des hommes (figure 5.3, gauche), la population étudiée est donc majoritairement masculine. Cependant, nous constatons une tendance à la féminisation au cours du temps : en 1995 les femmes représentaient 20.9% des agents, ce chiffre s'élève à 25.9% en 2011 (figure 5.3, droite).

• *Age.* L'âge moyen des cohortistes tend à augmenter, que cela soit pour les hommes ou pour les femmes (figure 5.4). En effet, les départs n'étant pas totalement compensés par les embauches (*cf.* figure 5.2, gauche), la population connaît une tendance vieillissante. Le changement d'allure des courbes observé à partir de 1999 est dû au nombre important

7. Nous avons par exemple éliminé les variables « commune de naissance », « nombre de personnes vivant au foyer » et « PCS INSEE ».

8. Pour chaque agent, nous considérons les données du dernier mois d'activité de l'année : décembre s'il n'a pas quitté son emploi, mois de sortie sinon.

d'embauches effectuées en 1999 et 2000 (*cf.* figure 5.2, gauche). Nous observons également à cette période une séparation des deux courbes « hommes » et « femmes », en effet l'âge moyen des femmes diminue alors que celui des hommes continue d'augmenter, ceci s'explique par le fait qu'un grand nombre de femmes jeunes aient été embauchées ces années là. Cet écart entre l'âge moyen des hommes et l'âge moyen des femmes, observé à partir de 1999, perdure depuis et s'élève à environ deux ans.

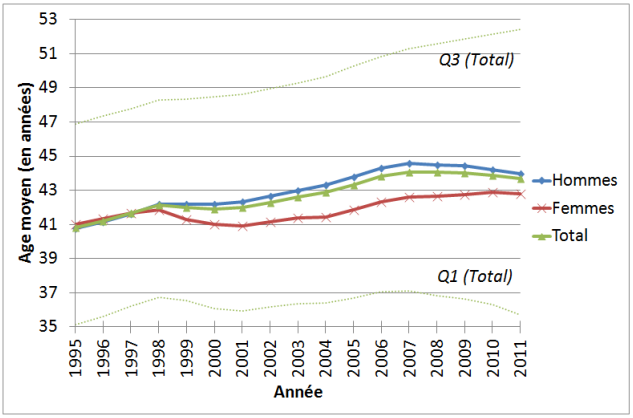


FIGURE 5.4 – Age moyen des cohortistes et ventilation par sexe.

• *Age à l'embauche et à la sortie de l'emploi.* L'âge à l'embauche et à la sortie de l'emploi est stable depuis 1995. Selon les années, il se situe en moyenne entre 26 et 29 ans pour l'embauche et entre 53 et 55 ans pour la sortie de l'emploi (figure 5.5). L'âge de départ relativement bas s'explique par :

- la particularité de certains métiers des IEG qui comprennent une part importante de « services actifs » (travail de nuit, astreinte, conditions climatiques difficiles, *etc.*) et qui permettent donc un départ anticipé,
- des éléments inhérents au statut des IEG (par exemple la possibilité de départ anticipé pour les femmes ayant au moins trois enfants).

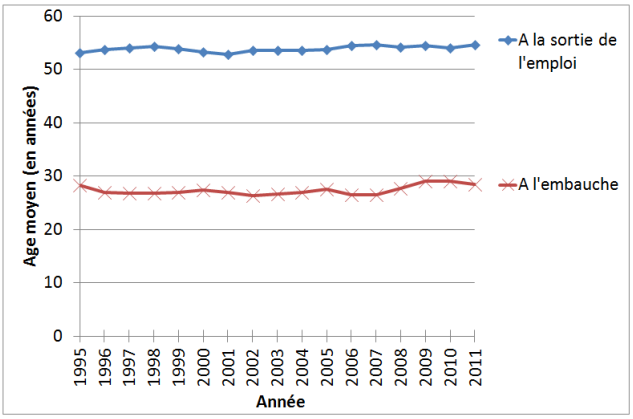


FIGURE 5.5 – Evolution de l'âge à l'embauche et à la sortie de l'emploi de 1995 à 2011.

• *Collège à l'embauche.* Le collège est une notion qu'utilisent les IEG pour caractériser le type d'emploi des salariés. Elle est divisée en trois catégories :

- agents d'exécution (correspondent aux ouvriers et employés de la nomenclature INSEE

utilisée pour le codage des PCS [9]),
– agents de maîtrise (correspondent aux employés et professions intermédiaires),
– cadres (correspondent aux cadres et professions intellectuelles supérieures).

64.3% des 215 550 cohortistes ont été embauchés en tant qu'agents d'exécution, 22.0% en tant qu'agents de maîtrise et 13.1% en tant que cadres, la donnée est manquante pour 0.6% d'entre eux (figure 5.6, gauche). Considérons maintenant la répartition du collège à l'embauche des cohortistes des années 1995 à 2011. Nous pouvons remarquer une tendance croissante concernant la proportion d'agents ayant été embauchés en tant que cadres ou agents de maîtrise. *A contrario*, la proportion d'agents ayant été embauchés en tant qu'agents d'exécution diminue (figure 5.6, droite). Ces tendances s'expliquent peut–être par une évolution des métiers au sein des IEG : à l'ère de l'informatique et des nouvelles technologies, certains métiers, qui demandaient initialement une qualification faible, nécessitent aujourd'hui un niveau supérieur, d'autres ayant même disparu.

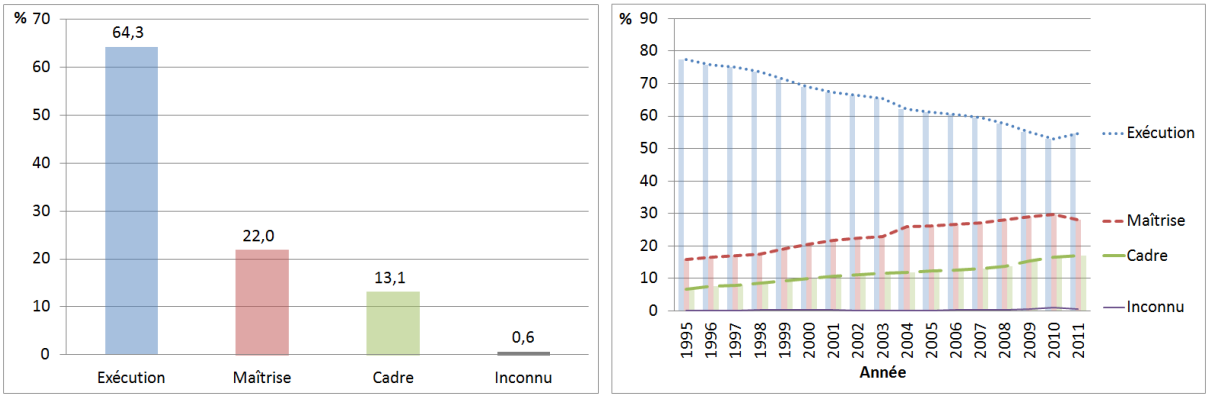


FIGURE 5.6 – A gauche : répartition du collège à l'embauche de l'ensemble des cohortistes ; à droite : répartition du collège à l'embauche des cohortistes des années 1995 à 2011.

• *Niveau de diplôme.* La répartition des niveaux de diplôme des 215 550 cohortistes est représentée sur la partie gauche de la figure 5.7. Nous constatons un taux élevé de niveaux de diplôme relativement bas, en effet 62.7% des cohortistes ont un niveau équivalent ou inférieur au Baccalauréat. Pour cette variable, 1.3% des données sont manquantes. Considérons également la répartition du niveau de diplôme des cohortistes des années 1995 à 2011 (figure 5.7, droite). Nous pouvons constater une tendance à la hausse concernant la proportion d'agents ayant un niveau de diplôme supérieur ou équivalent au Baccalauréat, au détriment des niveaux inférieurs. Comme mentionné précédemment, ces tendances sont dues en partie à l'évolution des métiers des IEG qui nécessitent une formation de plus en plus solide, mais aussi à l'évolution de la société au sein de laquelle l'enseignement a tendance à se démocratiser (*cf.* par exemple Merle, 2009 [59]).

9. Niveau 1 de la nomenclature INSEE des professions et catégories socioprofessionnelles (divisée en 8 catégories), téléchargeable à l'adresse : http ://www.insee.fr/fr/methodes/ ?page=nomenclatures/pcs2003/pcs2003.htm.

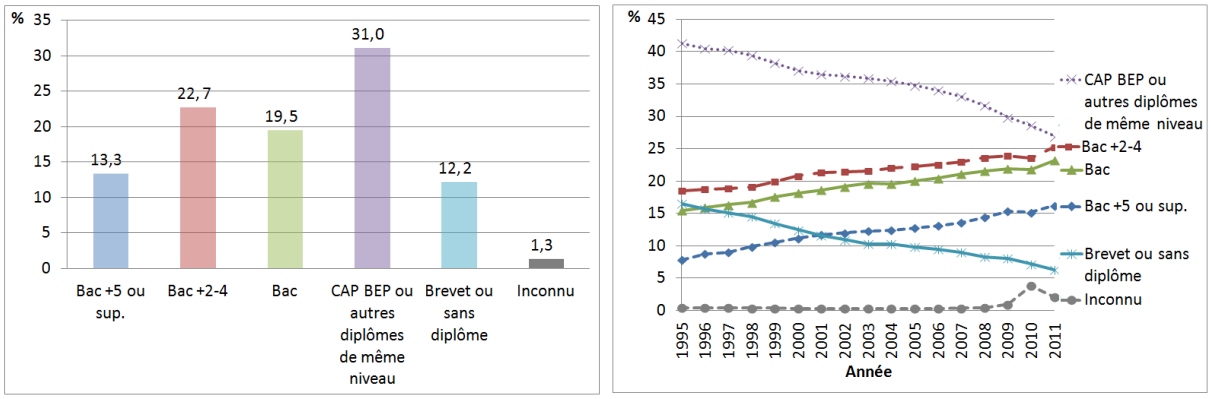


FIGURE 5.7 – A gauche : répartition du niveau de diplôme à l'embauche de l'ensemble des cohortistes ; à droite : répartition du niveau de diplôme des cohortistes des années 1995 à 2011.

• *Collège.* De 1995 à 2011, le pourcentage d'agents de maîtrise au sein des IEG reste quasiment constant et oscille autour de 52%. En revanche, les pourcentages de cadres et d'agents d'exécution augmentent et baissent progressivement : en 1995 les IEG comptent 17% de cadres et 31% d'agents d'exécution, en 2004 les cadres deviennent plus nombreux que les agents d'exécution et en 2011 les deux taux sont inversés : les IEG comptent 17% d'agents d'exécution et 32% de cadres. Cette inversion exécution/cadre est probablement la résultante de l'essor de l'informatique entraînant une évolution des métiers des IEG, de la massification scolaire et de politiques menées dans les entreprises favorisant la promotion interne et par conséquent l'« ascension sociale ».

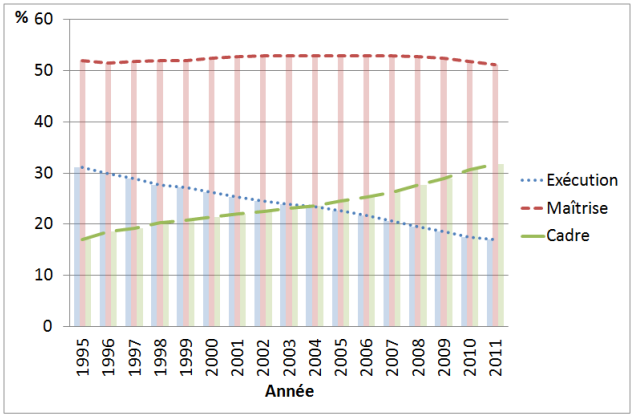


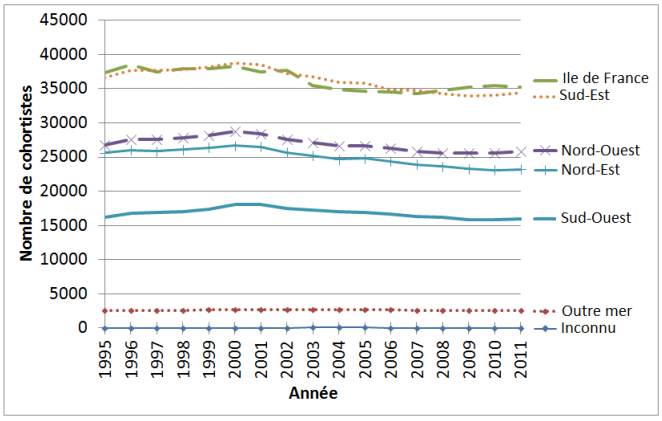
FIGURE 5.8 – Evolution de la répartition exécution/maîtrise/cadres de 1995 à 2011 (calculée sur l'ensemble des salariés).

• *Région.* Nous regroupons ici les différentes régions administratives françaises en 6 grandes régions (*cf.* table A.1, p.171, en annexe) :

- Ile de France
- Nord–Ouest
- Nord–Est
- Sud–Est (dont la Corse)
- Sud–Ouest
- Outre mer

La figure suivante représente l'évolution des effectifs selon le lieu de travail réparti dans les 6 grandes régions de 1995 à 2011. Excepté pour l'outre mer, les effectifs de chaque grande

région diminuent de façon quasiment parallèle de 1995 à 2011. L'Ile de France et le Sud–Est sont les régions les plus importantes en terme de nombre d'agents (respectivement 25.7% et 25.1% de l'effectif en 2011).



Dans la suite, nous nous intéressons en particulier aux variables *sexe, collège* et *âge*.

2 Les épisodes d'arrêt

Un salarié bénéficiaire du régime spécial mis en arrêt de travail par son médecin traitant (à l'extérieur des entreprises) continue de recevoir son plein salaire, versé par l'employeur, dès son premier jour d'arrêt. Ce n'est pas le cas pour le régime général où un délai de carence est appliqué. En revanche, le salarié concerné doit se rendre systématiquement à la consultation de son médecin conseil qui valide le bien–fondé de son absence médicale. A cette occasion, le médecin conseil instruit le dossier médical informatisé du salarié dans l'application Aramis en renseignant notamment les caractéristiques de l'*épisode d'arrêt*[10].

Il est important de distinguer les différents types d'absence (ou types de *risque*) pouvant être associés à un épisode d'arrêt :

1. épisode de « courte durée » (CD), peut être motivé par :
   - une « maladie de courte durée » (MCD) (pathologies courantes),
   - un « accident du travail » (AT),
   - une « maladie professionnelle » (MP),
2. épisode de « longue maladie » (LM), correspond à une mesure spécifique du régime spécial : lorsque le médecin conseil juge que l'absence va « durer », il place le salarié en LM (en concertation avec le médecin traitant). Le salarié continue de recevoir son plein salaire pendant 5 ans, sauf amélioration de son état et reprise du travail.

Lorsque le salarié ne peut reprendre son travail après 5 ans de LM, ou suite à un AT ou une MP, il est placé en « invalidité » jusqu'à sa retraite.

Outre le type d'absence, le médecin conseil renseigne aussi le code diagnostic de la pathologie associée à l'épisode d'arrêt. La nomenclature utilisée pour le codage des pathologies au SGMC (*cf.* [24] et table A.3, p.173, en annexe) a été définie par la cellule épidémiologie à partir du codage alphanumérique de la CIM–10 ($10^{\text{ème}}$ révision de la Classification Internationale des Maladies et des problèmes de santé connexes, *cf.* table A.4, p.174, en annexe). Elle a cependant été adaptée aux besoins inhérents au métier des médecins conseils.

10. Un épisode d'arrêt est constitué d'un arrêt initial et des éventuelles prolongations de même diagnostic.

Le médecin conseil renseigne également la date de début de l'épisode d'arrêt et sa durée en jours calendaires. D'autres informations, de nature plus administrative, sont aussi renseignées mais nous ne nous y intéressons pas ici.

Des fichiers d'épisodes d'arrêt peuvent être exportés du système Aramis par la cellule épidémiologie. Les données relatives aux épisodes d'arrêts initiés entre 1995 et 2011 ont été harmonisées à la cellule épidémiologie et consolidées en décembre 2012.

Dans ce chapitre, nous étudions l'absentéisme ventilé selon les grandes familles de pathologies qui constituent les chapitres de la nomenclature du SGMC (*cf.* table A.3, p.173, en annexe) :

- chapitre 1 : « Infectieux » (libellé court : inf),
- chapitre 2 : « Tumeurs » (tum),
- chapitre 3 : « Sang » (sang),
- chapitre 4 : « Endocrinien » (endo),
- chapitre 5 : « Psychiatrie » (psy),
- chapitre 6 : « Système nerveux » (sn),
- chapitres 7 et 8 : « Organes des sens » (sens),
- chapitre 9 : « Circulatoire » (circ),
- chapitre 10 : « Respiratoire » (resp),
- chapitre 11 : « Digestif » (dig),
- chapitre 12 : « Peau » (peau),
- chapitre 13 : « Ostéo–articulaire » (osart),
- chapitre 14 : « Génito–urinaire » (gen),
- chapitre 15 : « Grossesse, accouchement » (acmt),
- chapitre 17 : « Congénital » (cong),
- chapitre 18 : « Signes fonctionnel » (sf),
- chapitre 19 : « Chapitre 19 » (ch19),
- chapitre 20 : « Chapitre 20 » (trauma).

Croisement. A ce stade de l'exposé, nous devons introduire la notion d'épisode *calendaire* : un épisode débutant l'année n et se prolongeant l'année $n+1$ peut être coupé au 31 décembre de l'année n et transformé en deux épisodes calendaires. Par exemple, un épisode d'arrêt débutant le 1er décembre de l'année 2010 et se terminant le 20 janvier de l'année 2011 est divisé en deux épisodes *calendaires* : l'un de 31 jours relatif à l'année 2010 (du 1er au 31 décembre 2010) et l'autre de 20 jours relatif à l'année 2011 (du 1er au 20 janvier 2011). Dans la suite, nous ne considérons que des épisodes calendaires, c'est–à–dire bornés par le 1er janvier et le 31 décembre.

La cohorte EPIEG et les épisodes d'arrêt calendaires exportés peuvent être croisés (*cf.* figure 5.9), au sein de la cellule épidémiologie uniquement, à l'aide d'une procédure complexe utilisant l'identifiant crypté dédié aux statisticiens de la cellule épidémiologie. Cette procédure de cryptage fait l'objet d'une déclaration CNIL (Commission Nationale de l'Informatique et des Libertés, autorisation n° 116 95 73).

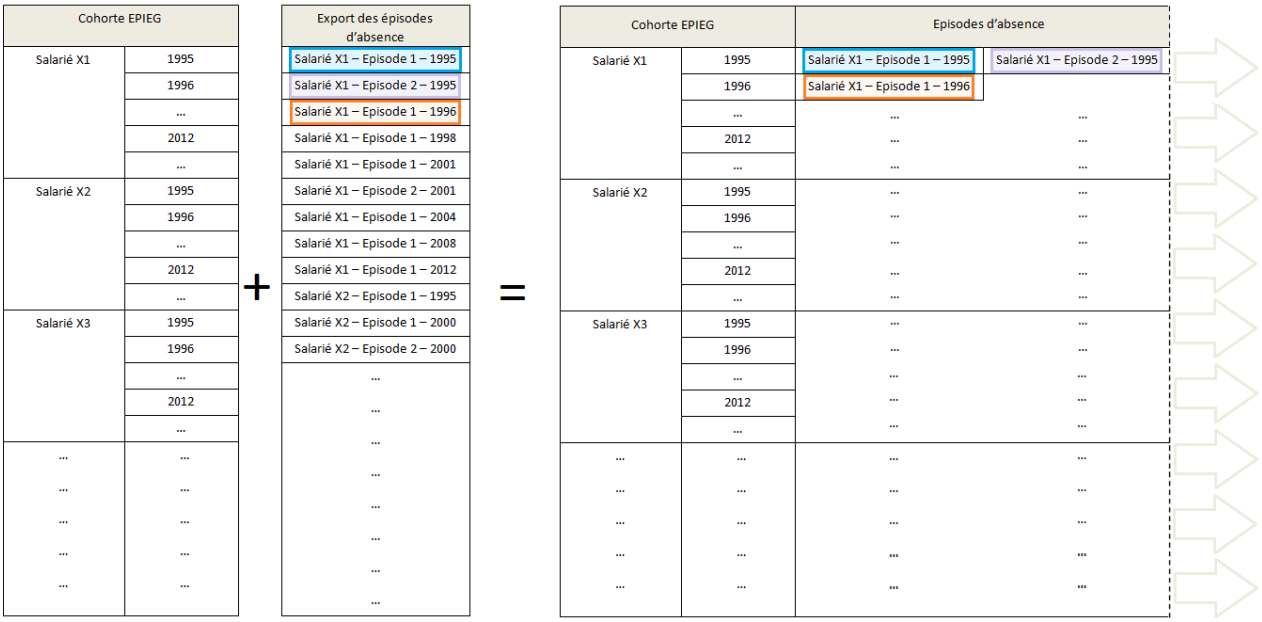


FIGURE 5.9 – Croisement de la cohorte EPIEG et des épisodes d'arrêt par Identifiant–Année.

Les bases de données issues de ces croisements ont pour finalité :

- de réunir, de façon permanente, des données de base pour l'épidémiologie dans les IEG concernant l'état de santé et la population des salariés ;
- de permettre l'analyse descriptive des données enregistrées et la surveillance épidémiologique de certains problèmes de santé émergents ;
- de faciliter la réalisation d'enquêtes ad hoc ;
- de participer à des études d'intérêt scientifique, en fournissant des fichiers anonymes, à des interlocuteurs d'études autorisées par le responsable des traitements du SGMC, puis par la CNIL (par exemple à la cohorte GAZEL) ;
- de fournir, régulièrement, des tableaux de bords agrégés d'indicateurs de santé et des analyses descriptives de la santé de la population aux managers et aux responsables de prévention ;
- de contribuer à l'évaluation d'actions de prévention.

C'est sur ces bases de données que portent les travaux appliqués de cette thèse.

3 Evolution des principaux indices d'absence

Nous présentons ici les évolutions de 1995 à 2011 des deux indices d'absence suivants :

1. Taux d'absentéisme (en %), qui se calcule de la façon suivante :

$$\frac{\text{Nombre de jours d'absence dans l'année}}{365 \times \text{ Nombre de personnes/années}^{\,a}} \times 100$$

a. Tenant compte du temps de présence dans l'année de chaque agent.

Il s'agit du nombre de jours d'absence moyen par agent, sur 100 jours calendaires de l'année considérée, c'est–à–dire week–end et jours fériés compris.

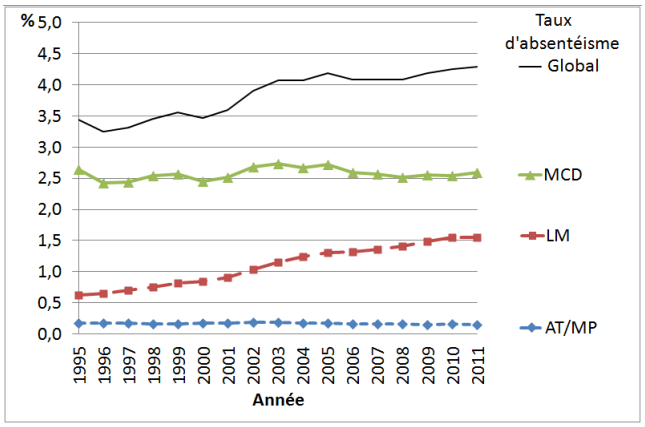


FIGURE 5.10 – Taux d'absentéisme tous types d'absence confondus (Global) et ventilé selon le type d'absence (Global = MCD + LM + AT/MP), de 1995 à 2011.

De 1995 à 2011, le taux d'absentéisme global (tous types d'basence confondus) passe de 3.44% à 4.29%. Cette augmentation semble être la conséquence de l'augmentation constante du taux d'absentéisme LM qui est multiplié par 2.5 de 1995 à 2011 : il passe de 0.62% en 1995 à 1.55% en 2011. Les taux d'absentéisme MCD et AT/MP sont eux assez stables : entre 2.43% et 2.72% pour la MCD et entre 0.14% et 0.19% pour les AT/MP.

2. Pourcentage d'agents arrêtés :

$$\frac{\text{Nombre d'agents arrêtés au moins une fois dans l'année}}{\text{Nombre d'agents présents au moins un jour dans l'année}} \times 100$$

Le graphique suivant présente l'évolution du taux d'agents arrêtés au moins une fois tous types d'absence confondus et du taux d'agents arrêtés au moins une fois pour une MCD, pour une LM et pour un/e AT/MP. Contrairement au taux d'absentéisme, le taux d'agents arrêtés « global » n'est pas égal à la somme des taux d'agents arrêtés MCD + LM + AT/MP. En effet un agent peut s'être absenté plusieurs fois au cours de la même année, pour des types d'absence différents.

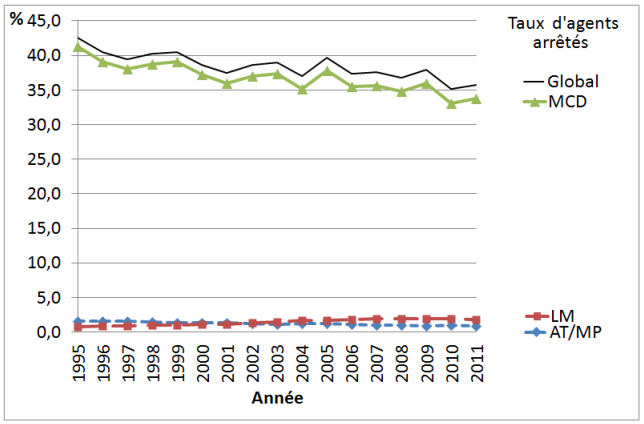


FIGURE 5.11 – Taux d'agents arrêtés tous types d'absence confondus (Global) et selon le type d'absence (MCD–AT/MP–LM), de 1995 à 2011.

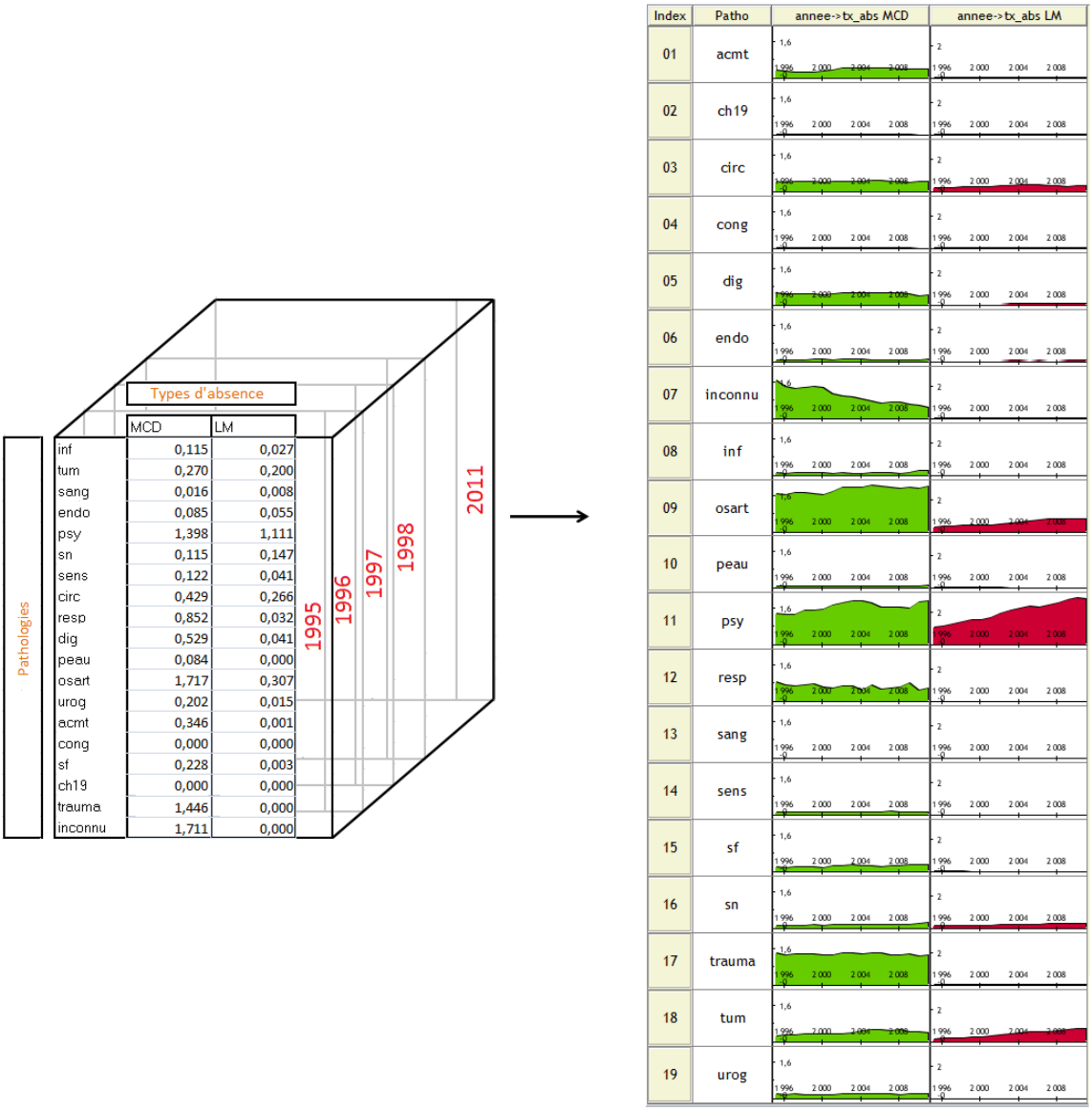
La proximité des courbes Global et MCD indique que la majorité des agents s'arrêtent pour une MCD. La diminution du taux d'agents arrêtés tous types d'absence confondus, qui passe de 42.6% en 1995 à 35.6% en 2011, s'explique donc par la diminution du taux d'agents arrêtés pour une MCD : 41.3% en 1995, 33.8% en 2011. Le taux d'agents arrêtés pour LM est lui en constante augmentation (masqué par un effet d'échelle sur la figure 5.11), il passe de 0.79% en 1995 à 1.81% en 2011. Concernant les AT/MP, le taux agents arrêtés pour ces types d'absence reste très faible : autour de 0.16%.

La diminution du taux d'agents arrêtés pour une MCD, conjuguée avec la stabilité du taux d'absentéisme MCD, met en évidence un changement au niveau de l'absentéisme de 1995 à 2011 : les arrêts sont moins fréquents mais plus longs. Les augmentations observées au niveau de la LM, aussi bien pour le taux d'absentéisme que pour le taux d'agents arrêtés peuvent cacher l'émergence de pathologies lourdes. Dans la suite, nous prolongeons ces analyses en étudiant séparément l'absentéisme de courte durée (MCD) et l'absentéisme de longue durée (LM).

Pathologies en cause. Nous voulons dès à présent fournir des premières explications concernant ces évolutions, notamment en identifiant les familles de pathologies qui semblent être sous–jacentes aux tendances observées précédemment.

Il s'agit ici d'une première étape de visualisation et d'exploration qui peut être longue et fastidieuse à mettre en oeuvre. Le logiciel DeltaMetric$_{©}$, développé avec l'aide du CEREMADE, possède un outil graphique très performant qui grâce à la visualisation qu'il permet rend cette première phase très aisée ; il constitue par conséquent une aide à la décision considérable [11]. Pour chaque année, nous disposons d'un tableau croisé Pathologie × Type d'absence (MCD ou LM) dans lequel chaque cellule contient le taux d'absentéisme par famille de pathologies en cause et par type d'absence, pour l'année considérée. Nous avons donc une succession de 17 tableaux (de 1995 à 2011) de même type, c'est–à–dire un tableau à trois dimensions Pathologie × Type d'absence × Temps (figure 5.12, gauche). A partir de ce tableau multiple, le logiciel DeltaMetric$_{©}$ construit le tableau croisé Pathologie × Type d'absence (MCD ou LM) où chaque cellule contient désormais une donnée complexe, ici une courbe représentant des évolutions temporelles (figure 5.12, droite). Nous sommes alors plongés dans le cadre de l'*Analyse de Données Fonctionnelle* avec comme troisième entrée le temps (*cf.* Rice, 2004 [69] et Zhao, Marron & Wells, 2004 [90]).

11. Pour des exemples d'applications, se référer aux travaux de Gettler–Summa & al., 2008 [36] et 2006 [35].



	MCD	LM
inf	0,115	0,027
tum	0,270	0,200
sang	0,016	0,008
endo	0,085	0,055
psy	1,398	1,111
sn	0,115	0,147
sens	0,122	0,041
circ	0,429	0,266
resp	0,852	0,032
dig	0,529	0,041
peau	0,084	0,000
osart	1,717	0,307
urog	0,202	0,015
acmt	0,346	0,001
cong	0,000	0,000
sf	0,228	0,003
ch19	0,000	0,000
trauma	1,446	0,000
inconnu	1,711	0,000

Index	Patho
01	acmt
02	ch19
03	circ
04	cong
05	dig
06	endo
07	inconnu
08	inf
09	osart
10	peau
11	psy
12	resp
13	sang
14	sens
15	sf
16	sn
17	trauma
18	tum
19	urog

FIGURE 5.12 – Tableau à trois dimensions Pathologie × Type d'absence × Année (à gauche) à partir duquel est construit le tableau de courbes représentant les évolutions du taux d'absentéisme par famille de pathologies et par type d'absence de 1995 à 2011 (à droite). Par souci de lisibilité, les échelles verticales des deux colonnes de courbes sont différentes (option de DeltaMetric©).

Outre les possibilités de gestion de données (filtrage, recodage, opérations mathématiques entre colonnes, *etc.*), le logiciel offre, dans le cas de données de type courbes, des possibilités de lissages et d'approximations polynomiales et fournit un ensemble d'indicateurs qui peuvent s'avérer utiles dans une phase exploratoire : coefficient de variation, pente de la droite des moindres carrés, abscisse extrême, *etc.* Il permet également le tri d'une colonne de courbes selon un des critères suivants : intégrale des valeurs (aire sous la courbe), intégrale des carrés des valeurs, maximum.

Les deux tableaux suivants représentent les courbes triées selon l'intégrale des valeurs pour la MCD d'une part et pour la LM d'autre part. Pour chaque courbe, nous avons calculé le taux de variation et la corrélation des rangs de Spearman entre l'année et le taux d'absentéisme (les formules sont données ci–après). La significativité de cette corrélation est testée à l'aide du test non paramétrique de Spearman associé. Une corrélation positive et un test significatif vont dans le sens d'un taux d'absentéisme à tendance croissante.

Formules :

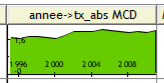
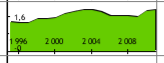
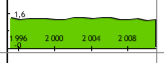
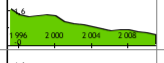
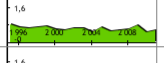
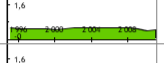
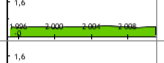
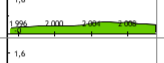
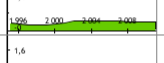
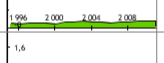
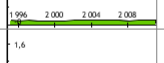
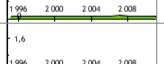
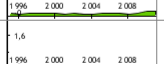
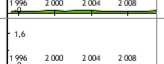
• Taux de variation entre 1995 et 2011 :

$$\frac{\text{Valeur en 2011} - \text{Valeur en 1995}}{\text{Valeur en 1995}} \times 100$$

• Coefficient de corrélation de Spearman.
Nous disposons ici de $n = 17$ valeurs du taux d'absentéisme, une par année de 1995 à 2011. A chacune de ces valeurs, nous associons le rang correspondant selon le classement par ordre croissant. De même, à chaque année, nous associons le rang correspondant selon le classement par ordre croissant. Le coefficient de corrélation de Spearman, noté r_s, est donné par la formule suivante :

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

où $(d_i)_{i=1,\dots,n}$ est ici la différence entre le rang de l'année et celui du taux d'absentéisme.

Patho	annee->tx_abs MCD	MCD_tx_variation	Spearman MCD
osart		17,92	0.75 ***
psy		38,31	0.66 **
trauma		-6,8	NS
inconnu		-71,04	-0.97 ***
resp		-31	-0.52*
dig		-17,86	NS
circ		8,97	NS
tum		56,82	0.80 ***
acmt		12,47	0.59 *
sf		40,86	0.82 ***
urog		1,17	NS
sn		108,11	0.85 ***
sens		15,16	0.51 *
inf		85,34	NS
endo		25,98	NS
peau		19,21	0.51 *
sang		66,94	NS
ch19		n/a	0.69 **
cong		n/a	NS

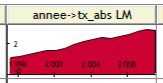
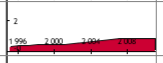
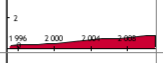
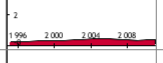
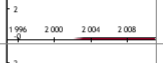
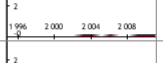
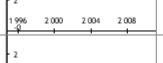
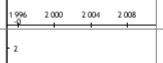
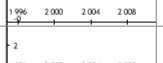
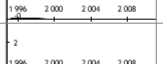
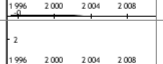
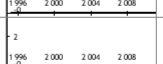
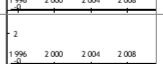
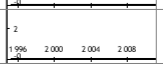
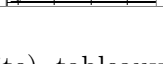
Patho	annee->tx_abs LM	LM_taux_variation	Spearman LM
psy		156,25	0.99 ***
osart		168,95	0.99 ***
tum		319,67	1***
circ		37,96	0.70 **
sn		110,21	0.95 ***
dig		151,85	0.95 ***
endo		78,92	0.64 **
resp		107,62	0.74 ***
sens		63,5	0.73 ***
inf		-5,5	NS
urog		131,99	0.86 ***
sang		271,45	0.88 ***
sf		442,19	0.73 ***
peau		n/a	NS
ch19		n/a	NS
acmt		-100	NS
cong		n/a	NS
inconnu		n/a	NS
trauma		n/a	NS

FIGURE 5.13 – Pour la MCD (à gauche) et la LM (à droite), tableaux constitués :
- des courbes représentant les évolutions du taux d'absentéisme par famille de pathologies, de 1995 à 2011, triées de façon décroissante selon l'aire sous la courbe,
- des taux de variation correspondants,
- des coefficients de corrélation de Spearman correspondants. Pour chaque coefficient, nous avons ajouté la significativité du test associé (* : $0.01 < p \leq 0.05$, ** : $0.001 < p \leq 0.01$, *** : $p \leq 0.001$, où p est le seuil observé du test).

L'étude de ces tableaux permet d'énoncer les résultats descriptifs suivants :

– Les trois familles de pathologies les plus représentées en MCD sont l'ostéo–articulaire, la psychiatrie et la traumatologie. Les taux d'absentéisme dus à des pathologies ostéo–articulaires et psychiatriques ont tendance à augmenter depuis 1995 : taux de variation respectifs de 18%, 39%. En revanche, la traumatologie reste stable. Nous pouvons également noter la tendance croissante de l'absentéisme dû à des pathologies tumorales : taux de variation de 57% ; et la tendance décroissante de l'absentéisme dû à des pathologies respiratoires et digestives : taux de variation respectifs de -31% et -18%.

Remarque : Nous constatons une diminution importante des pathologies non codées (code « inconnu »). Elle s'explique par la politique interne du SGMC encourageant les médecins conseils à renseigner les codes diagnostics de façon systématique.

– Les trois familles de pathologies prépondérantes en LM sont la psychiatrie, l'ostéo–articulaire et les tumeurs. Le taux d'absentéisme dû à chacune de ces pathologies est en nette augmentation : taux de variation respectifs de 156%, 169% et 320%.

Ces résultats laissent penser que l'augmentation du taux d'absentéisme LM (figure 5.10, p.134) est surtout due à la hausse des LM motivées par des pathologies liées à la psychiatrie et, dans une moindre mesure, aux pathologies ostéo–articulaires et tumorales. Concernant la MCD, les tendances croissantes et décroissantes observées pour certaines familles de pathologie se compensent et n'influent pas sur le taux d'absentéisme MCD, toutes pathologies confondues, qui reste stable (figure 5.10, p.134).

***Qui s'absente ?* Ajustement âge, sexe, collège.** De nombreuses études (*cf.* Chevalier & al., 1987 [15] ; 1992 [17]) montrent que l'absence est notamment liée à l'âge, au sexe et au collège. Comme nous l'avons vu au paragraphe 1.3 (p.127), les IEG ont une structure de population en constante évolution depuis 1995 : population vieillissante, comportant de plus en plus de femmes et de plus en plus de cadres (au détriment des agents d'exécution). Il est donc primordial de pouvoir déterminer si les tendances observées (augmentation du taux d'absentéisme LM notamment) ne sont pas dues à l'évolution de la structure de la population. Pour ce faire, une première étape consiste à étudier l'évolution des taux d'absentéisme MCD et LM par classe d'âge, par sexe et par collège (figure 5.14).

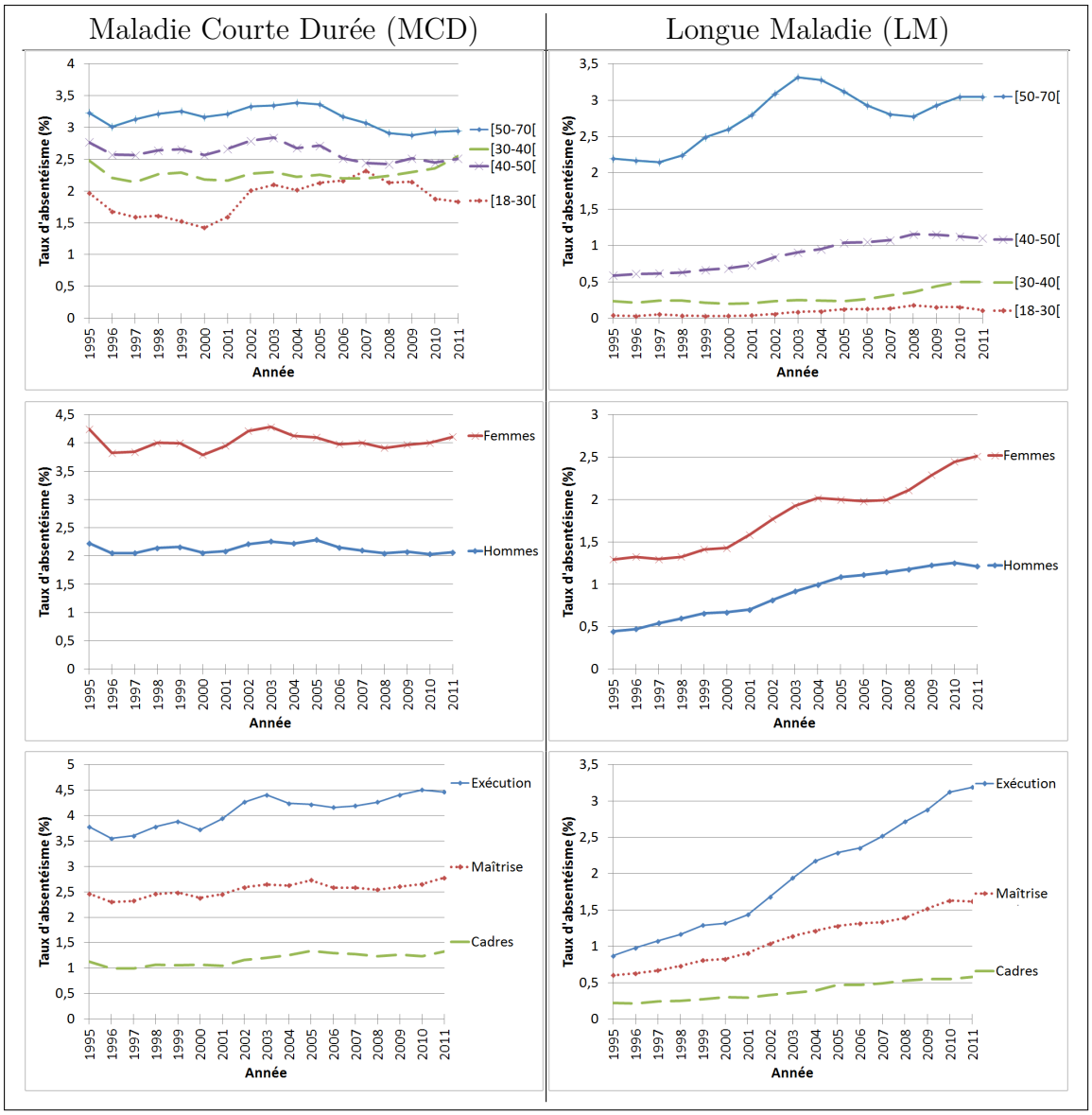


FIGURE 5.14 – Taux d'absentéisme MCD et LM par classe d'âge, par sexe et par collège, de 1995 à 2011.

Nous pouvons énoncer les résultats suivants (figure 5.14) :

- Aussi bien pour la MCD que pour la LM, l'absence augmente avec l'âge et les femmes s'absentent plus que les hommes. De plus, le taux d'absentéisme des agents d'exécution est supérieur à celui des agents de maîtrise qui lui même est supérieur à celui des cadres.
- Quelle que soit la catégorie populationnelle étudiée, les taux d'absentéisme LM sont en forte augmentation.

Afin de compléter ces premiers résultats, étudions maintenant l'évolution de ces taux pour les groupes d'agents déterminés par le croisement Age × Sexe × Collège. DeltaMetric© permet d'obtenir facilement les deux tableaux suivants [12] :

12. Colonnes « Spearman » obtenues avec le logiciel SAS et importées dans DeltaMetric©.

Sexe	Collège	Classe d'âge	annee->tx_abs MCD	tx_variation	Spearman
Femme	Exécution	[50-70[		-8,27	NS
Femme	Exécution	[40-50[		30,3	0,92 ***
Femme	Exécution	[30-40[		53,37	0,93 ***
Femme	Exécution	[18-30[		59,09	0,86 ***
Homme	Exécution	[50-70[		2,69	0,53 *
Femme	Maîtrise	[50-70[		-1,12	NS
Homme	Exécution	[40-50[		1,08	NS
Femme	Maîtrise	[40-50[		9,89	NS
Femme	Maîtrise	[30-40[		15,08	0,77 **
Femme	Maîtrise	[18-30[		-6,92	NS
Homme	Maîtrise	[50-70[		3,13	NS
Femme	Cadres	[50-70[		-9,26	NS
Homme	Exécution	[30-40[		-3,49	NS
Homme	Maîtrise	[40-50[		-6,35	NS
Femme	Cadres	[30-40[		3,91	0,49 *
Homme	Exécution	[18-30[		11,96	0,68 *
Femme	Cadres	[40-50[		0,14	NS
Homme	Maîtrise	[30-40[		-0,15	NS
Homme	Cadres	[50-70[		7,14	NS
Femme	Cadres	[18-30[		-33,6	NS
Homme	Maîtrise	[18-30[		-8,57	0,49 *
Homme	Cadres	[40-50[		-3,62	0,51 *
Homme	Cadres	[30-40[		-0,79	NS
Homme	Cadres	[18-30[		-23,49	NS

Sexe	Collège	Classe d'âge	annee->tx_abs LM	tx_variation	Spearman
Femme	Exécution	[50-70[		142,92	0,92 ***
Homme	Exécution	[50-70[		100,78	0,96 ***
Femme	Maîtrise	[50-70[		28,18	0,69 **
Femme	Exécution	[40-50[		195,79	0,95 ***
Homme	Maîtrise	[50-70[		47,26	0,69 **
Homme	Exécution	[40-50[		176,69	0,98 ***
Femme	Cadres	[50-70[		8,73	NS
Femme	Maîtrise	[40-50[		133,26	0,97 ***
Femme	Exécution	[30-40[		241,24	0,92 ***
Homme	Cadres	[50-70[		85,29	0,88 ***
Femme	Cadres	[40-50[		59,68	0,80 ***
Homme	Maîtrise	[40-50[		63,63	0,93 ***
Femme	Maîtrise	[30-40[		51,52	0,69 **
Homme	Exécution	[30-40[		74,52	0,73 ***
Femme	Exécution	[18-30[		586,8	0,74 ***
Homme	Cadres	[40-50[		50,82	0,83 ***
Femme	Maîtrise	[18-30[		n/a	0,90 ***
Femme	Cadres	[30-40[		162	NS
Homme	Maîtrise	[30-40[		284,4	0,68 **
Homme	Exécution	[18-30[		95,73	0,82***
Femme	Cadres	[18-30[		n/a	NS
Homme	Maîtrise	[18-30[		3 166,67	0,60 *
Homme	Cadres	[30-40[		321,72	NS
Homme	Cadres	[18-30[		n/a	NS

FIGURE 5.15 – Tableaux constitués des évolutions du taux d'absentéisme par groupes de salariés de 1995 à 2011, triés selon l'aire sous la courbe (à gauche : MCD, à droite : LM)

Grâce à cette phase exploratoire, nous pouvons dire que les tendances obtenues précédemment (croissance de la LM et stabilité de la MCD) ne sont pas dues aux évolutions de la structure de la population des IEG selon la classe d'âge, le sexe et le collège. En effet, nous constatons que pour la quasi–totalité des groupes d'agents Age × Sexe × Collège, le taux d'absentéisme LM est croissant tandis que le taux d'absentéisme MCD est souvent stable.

Une régression linéaire expliquant le taux d'absentéisme par les variables « sexe », « classe d'âge », « collège » et « année » a été effectuée à la cellule épidémiologie à partir du même tableau et confirme ces résultats.

Ce tableau permet également d'identifier, d'une part pour la MCD, et d'autre part pour la LM, des groupes d'agents *sensibles*, c'est–à–dire pour lesquels le taux d'absentéisme relatif à l'année 2011 est soit élevé [13], soit en forte augmentation :

13. Nous considérons que les taux d'absentéisme MCD et LM des groupes sont élevés s'ils sont respectivement supérieurs à 2.59% et 1.55%, taux d'absentéisme globaux moyens pour l'année 2011.

- Pour la MCD, les femmes agents d'exécution de chaque classe d'âge ont les plus forts taux d'absentéisme (en moyenne), ces taux tendent même à augmenter au cours du temps (sauf pour la classe d'âge [50–70[). Les hommes agents d'exécution appartenant aux classes d'âge [40–50[et [50–70[et les femmes agents de maîtrise de chaque classe d'âge ont aussi un fort taux d'absentéisme MCD (en moyenne). Chez les cadres (pour chaque sexe et chaque classe d'âge), ce taux est en moyenne assez faible, excepté pour les femmes de plus de 50 ans.
- Concernant la LM, les agents d'exécution appartenant aux classes d'âge [40–50[et [50–70[et les agents de maîtrise âgés de plus de 50 ans ont un taux d'absentéisme LM très élevé et en augmentation depuis 1995. Bien qu'étant plus faible, le taux d'absentéisme des cadres de plus de 50 ans, hommes et femmes, a tendance a augmenter. Les autres groupes de salariés, c'est–à–dire les cadres de moins de 50 ans et les agents d'exécution et de maîtrise de moins de 40 sont moins concernés par la LM.

  Pour chaque groupe d'agents, les taux d'absentéisme MCD et LM de 2010 et 2011 sont donnés en annexe, p.175.

Dans cette section, nous avons répondu à la question : « *qui s'absente ?* ». Dans la suite, nous tentons de répondre à la question : « *qui s'absente pour quoi ?* », c'est–à–dire de mettre en évidence les liaisons qui peuvent exister entre groupes de salariés et absentéisme spécifique de certaines familles de pathologies. Pour ce faire, nous mettons en oeuvre deux analyses géométriques des données distinctes (analyses des correspondances), l'une pour la MCD et l'autre pour la LM. Grâce à ces analyses, nous constatons que ces liaisons ont considérablement évolué depuis 1995.

4 Analyses des correspondances

Dans cette section, nous utilisons l'analyse des correspondances afin de mettre en évidence les liaisons qui existent entre groupes d'agents et familles de pathologies en cause dans l'absentéisme LM d'une part et dans l'absentéisme MCD d'autre part.

Un aperçu des différentes méthodes d'analyses géométriques des données et de la théorie sous–jacente est donné en annexe, p.177.

4.1 Longue Maladie

4.1.1 Le tableau de données

A partir des données de la cohorte EPIEG fusionnées avec les épisodes d'arrêt LM des salariés, nous construisons le tableau soumis à l'analyse des correspondances de la façon suivante :

- en lignes : les groupes d'agents, déterminés par le croisement des variables Classe d'âge (8 modalités [14]), Sexe (2 modalités), Collège (3 modalités [15]) et dupliqués de 1995 à 2011 ;
- en colonnes : les grandes familles de pathologies (*cf.* p.132) ;

14. < 25, [25 ;30[, [30 ;35[, [35 ;40[, [40 ;45[, [45 ;50[, [50 ;55[, >=55 ans.
15. Exécution, Maîtrise, Cadres.

– dans chaque cellule : le nombre d'individus appartenant au groupe d'agents ligne et s'étant absentés pour une LM motivée par la famille de pathologies colonne, dans l'année considérée.

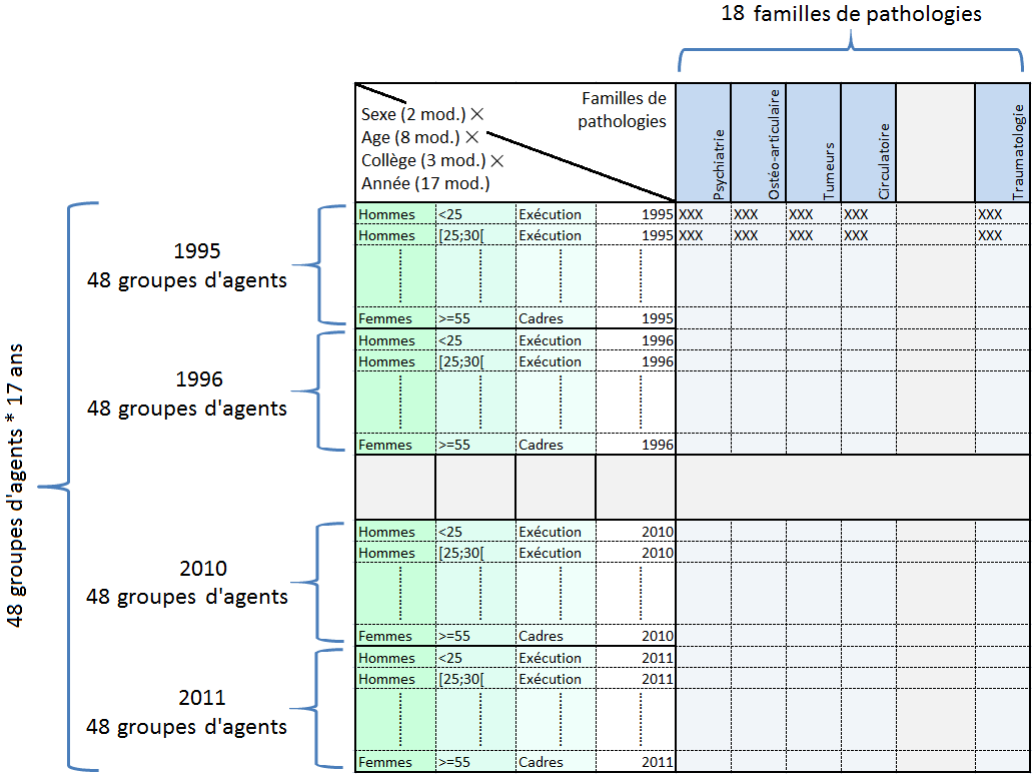


FIGURE 5.16 – Tableau soumis à l'analyse des correspondances.

4.1.2 Analyse des correspondances (année 2010)

L'analyse est effectuée avec le logiciel Coheris SPAD©.

Il s'agit ici d'identifier des liaisons récentes. Pour ce faire, nous effectuons l'analyse des correspondances du tableau relatif à l'année 2010. L'année 2010 est donc prise comme référence.

Remarque : Nous préférons travailler sur les données de l'année 2010, celles de l'année 2011 n'étant pas complètement consolidées au moment de la constitution du fichier de données (été 2012).

Les huit familles de pathologies suivantes, prépondérantes en LM, sont les colonnes retenues pour l'analyse :

– psychiatrie (libellé court : psy),
– ostéo–articulaire (osart),
– tumeurs (tum),
– circulatoire (circ),
– système nerveux (sn),
– digestif (dig),
– endocrinien (endo),
– respiratoire (resp).

Concernant les groupes d'agents, nous choisissons de ne retenir pour l'analyse que ceux pour lesquels au moins 5 des 8 familles de pathologies retenues ont motivé des épisodes d'arrêt durant l'année 2010, c'est–à–dire pour lesquels la ligne correspondante dans le tableau de données comprend au moins 5 éléments non nuls sur les 8 colonnes (pathologies) retenues. Finalement, 34 des 48 groupes d'agents de l'année 2010 ont été retenus pour l'analyse [16].

Les groupes d'agents des années 1995 à 2009 et 2011 seront par la suite projetés sur les plans principaux en utilisant la technique des éléments supplémentaires (*cf.* Cazes, 1982 [12]).

L'étude des variances des axes (tableau des valeurs propres ci–après) montre que les quatre premiers axes principaux expliquent une part importante de la variance totale (80.18%). De plus, l'écart entre la 4ème et la 5ème valeur propre est supérieur à l'écart entre la 5ème et la 6ème valeur propre. Nous interprétons donc les quatre premiers axes.

Tableau des valeurs propres
Trace de la matrice: 0.12465

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0,0359	28,83	28,83
2	0,0287	23,00	51,82
3	0,0192	15,37	67,19
4	0,0162	12,98	80,18
5	0,0112	8,97	89,15
6	0,0078	6,28	95,43
7	0,0057	4,57	100,00

La figure 5.17 suivante montre la représentation simultanée des groupes d'agents et des familles de pathologies dans les plans 1–2 et 3–4 de l'analyse. Les pathologies (représentées par un triangle) sont identifiées par leur libellé court et les groupes d'agents par la concaténation de la première lettre du sexe (H : Homme, F : Femme), de la première lettre du collège (E : Exécution, M : Maîtrise, C : Cadre) et de la classe d'âge. Les groupes hommes sont représentés en bleu et les groupes femmes en rouge.

16. La LM concernant moins les personnes jeunes que les plus âgées, les 14 groupes restants sont quasiment tous caractérisés par des classes d'âge jeunes : hommes, agents d'exécution de moins de 25 ans (1 groupe) ; hommes, agents de maîtrise de moins de 35 ans (3 groupes) ; hommes, cadres, de moins de 45 ans (5 groupes) ; femmes, agents d'exécution et de maîtrise de moins de 25 ans (2 groupes) ; femmes, cadres de moins de 35 ans (3 groupes).

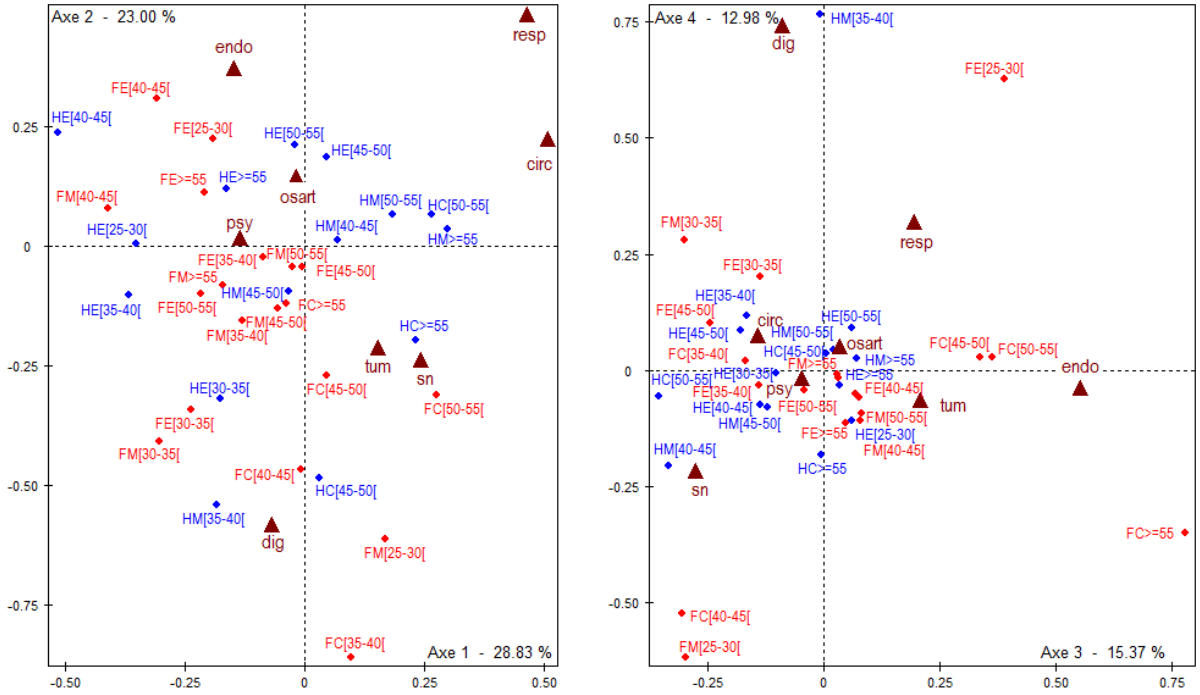


FIGURE 5.17 – Analyse des correspondances : plans 1–2 (gauche) et 3–4 (droite).

Interprétation des axes : liaisons Groupes d'agents/Familles de pathologies. Interprétons maintenant chacun des quatre premiers axes principaux à l'aide de la méthode des *contributions des points et des écarts* (*cf.* Le Roux & Rouanet, 1998 [49]).

Pour chaque axe, pour les points « groupes d'agents » d'un côté et les points « familles de pathologies » de l'autre, cette méthode consiste à :

- rechercher les points ou groupes de points dont la contribution à la variance de l'axe est supérieure à la contribution moyenne, soit $100/34 = 2.94\%$ pour les points « groupes d'agents » et $100/8 = 12.5\%$ pour les points « familles de pathologies » ;
- distinguer, parmi les points retenus, ceux à coordonnées positives de ceux à coordonnées négatives et calculer la contribution de l'écart entre les points moyens des points « négatifs » et « positifs » afin d'apprécier le résumé ainsi obtenu.

La figure suivante représente les familles de pathologies et les groupes d'agents les plus contributifs aux variances des 4 premiers axes, les proximités entre les points traduisent leurs degrés de liaison.

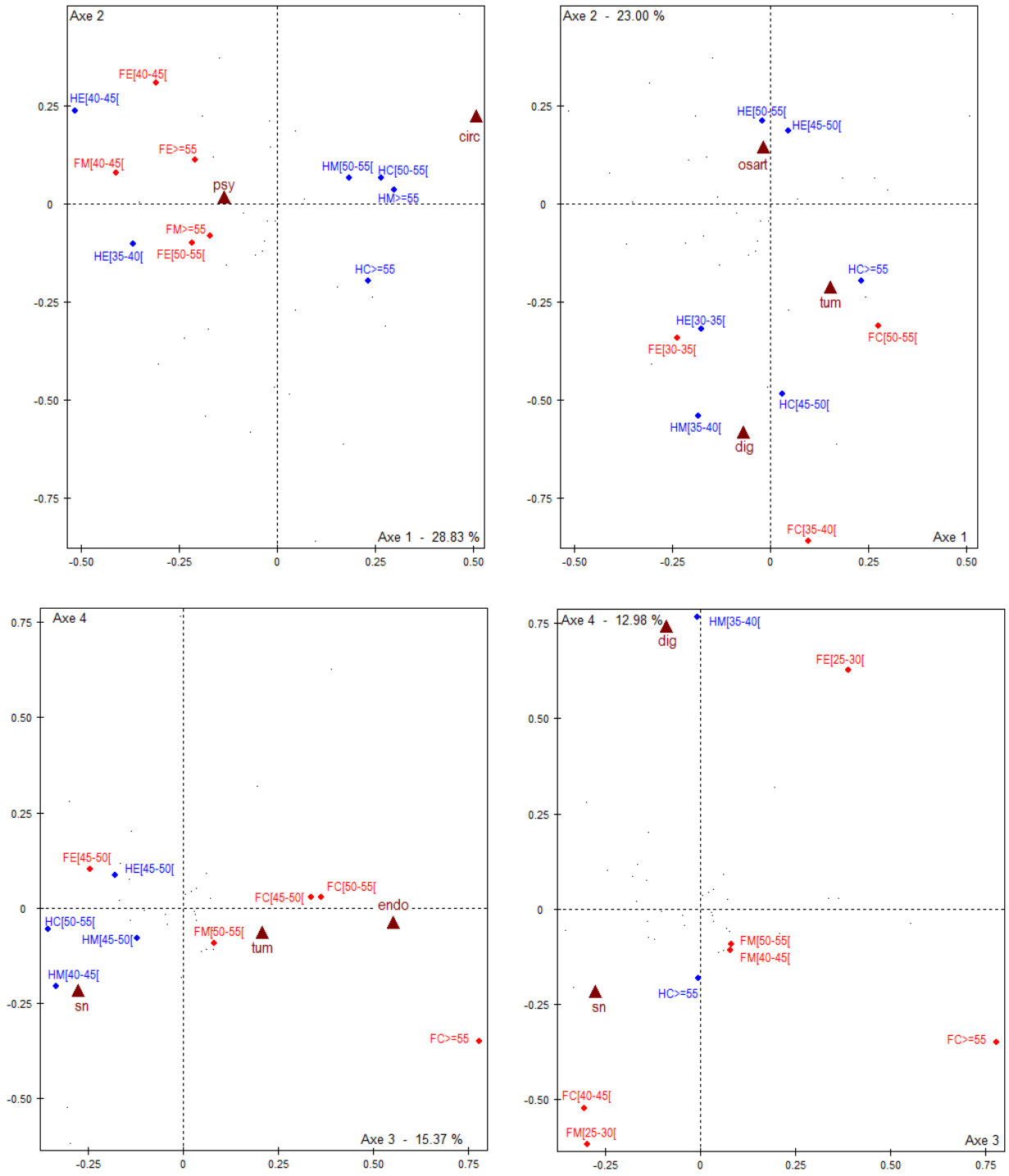


FIGURE 5.18 – Groupes d'agents et familles de pathologies les plus contributifs aux variances des 4 premiers axes : axe 1 (haut–gauche), axe 2 (haut–droit), axe 3 (bas–gauche) et axe 4 (bas–droit).

Interprétation de l'axe 1 (figure 5.18, quadrant haut–gauche).

Du côté négatif :	*Du coté positif :*
▶ Au niveau des pathologies :	▶ Au niveau des pathologies :
- Psychiatrie (contribution : 25.94%)	- Circulatoire (46.55%)
▶ Au niveau des groupes d'agents :	▶ Au niveau des groupes d'agents :
- Hommes, exécution, 40–45 ans (10.94%)	- Hommes, maîtrise, +55 ans (19.44%)
- Femmes, maîtrise, 40–45 ans (9.46%)	- Hommes, maîtrise, 50–55 ans (14.70%)
- Femmes, maîtrise, +55 ans (5.48%)	- Hommes, cadres, 50–55 ans (5.80%)
- Femmes, exécution, 50–55 ans (4.20%)	- Hommes, cadres, +55 ans (5.10%)
- Femmes, exécution, 40–45 ans (3.99%)	
- Femmes, exécution, +55 ans (3.72%)	
- Hommes, exécution, 35–40 ans (3.36%)	

Au niveau des familles de pathologies, le 1er axe oppose les pathologies psychiatriques (à gauche) aux pathologies circulatoires (à droite). Ces deux familles de pathologies contribuent ensemble à 72.49% de la variance de l'axe 1 tandis que leur écart contribue à 65.78% de cette même variance.

Au niveau des groupes d'agents, le 1er axe oppose 7 groupes d'agents (à gauche) : plutôt des femmes agents de maîtrise et d'exécution de plus de 40 ans, à 4 groupes d'agents (à droite) : plutôt des hommes agents de maîtrise et cadres de plus de 50 ans. Ces 11 groupes d'agents contribuent ensemble à 86.19% de la variance de l'axe 1 tandis que leur écart contribue à 75.82% de cette même variance.

En résumé, l'axe 1 montre une opposition entre les femmes agents de maîtrise et d'exécution de plus de 40 ans associées aux pathologies psychiatriques (à gauche) et les hommes agents de maîtrise et cadres de plus de 50 ans associés aux pathologies circulatoires (à droite).

Interprétation de l'axe 2 (figure 5.18, quadrant haut–droit).

Du côté négatif :	*Du coté positif :*
▶ Au niveau des pathologies :	▶ Au niveau des pathologies :
- Tumeurs (25.62%)	- Ostéo–articulaire (11.32%)
- Digestif (23.38%)	
▶ Au niveau des groupes d'agents :	▶ Au niveau des groupes d'agents :
- Femmes, cadres, 35–40 ans (11.09%)	- Hommes, exécution, 50–55 ans (16.99%)
- Hommes, maîtrise, 35–40 ans (10.79%)	- Hommes, exécution, 45–50 ans (5.14%)
- Hommes, cadres, 45–50 ans (5.75%)	
- Femmes, cadres, 50–55 ans (4.62%)	
- Hommes, cadres, +55 ans (4.61%)	
- Femmes, exécution, 30–35 ans (4.46%)	
- Hommes, exécution, 30–35 ans (3.17%)	

Au niveau des familles de pathologies, le 2ème axe oppose les pathologies tumorales et digestives (en bas) aux pathologies ostéo–articulaires (en haut). Ces trois familles de pathologies contribuent ensemble à 60.32% de la variance de l'axe 2 tandis que leur écart contribue à 46.37% de cette même variance.

Au niveau des groupes d'agents, le 2ème axe oppose 7 groupes d'agents (en bas) : plutôt des cadres et agents de maîtrise d'âge moyen, à 2 groupes d'agents (en haut) : plutôt des

hommes agents d'exécution. Ces 9 groupes d'agents contribuent ensemble à 66.62% de la variance de l'axe 2 tandis que leur écart contribue à 58.04% de cette même variance.

En résumé, l'axe 2 montre une opposition entre les cadres et agents de maîtrise d'âge moyen associés aux pathologies tumorales et digestives (en bas) et les hommes agents d'exécution associés aux pathologies ostéo–articulaires (en haut).

Interprétation de l'axe 3 (figure 5.18, quadrant bas–gauche).

Du côté négatif :	*Du coté positif :*
► Au niveau des pathologies :	► Au niveau des pathologies :
- Système nerveux (20.68%)	- Tumeurs (36.13%)
	- Endocrinien (26.05%)
► Au niveau des groupes d'agents :	► Au niveau des groupes d'agents :
- Hommes, cadres, 50–55 ans (19.59%)	- Femmes, cadres, +55 ans (18.38%)
- Hommes, exécution, 45–50 ans (7.17%)	- Femmes, cadres, 50–55 ans (9.29%)
- Hommes, maîtrise, 40–45 ans (6.59%)	- Femmes, cadres, 45–50 ans (5.53%)
- Femmes, exécution, 45–50 ans (6.14%)	- Femmes, maîtrise, 50–55 ans (3.05%)
- Hommes, maîtrise, 45–50 ans (3.50%)	

Au niveau des familles de pathologies, le 3ème axe oppose les pathologies du système nerveux (à gauche) aux pathologies tumorales et endocriniennes (à droite). Ces trois familles de pathologies contribuent ensemble à 82.86% de la variance de l'axe 3 tandis que leur écart contribue à 55.71% de cette même variance.

Au niveau des groupes d'agents, le 3ème axe oppose 5 groupes d'agents (à gauche) : plutôt des hommes d'âge moyen, à 4 groupes d'agents (à droite) : plutôt des femmes cadres plus âgées. Ces 9 groupes d'agents contribuent ensemble à 79.24% de la variance de l'axe 3 tandis que leur écart contribue à 50.85% de cette même variance.

En résumé, l'axe 3 montre une opposition entre les hommes d'âge moyen associés aux pathologies du système nerveux (à gauche) et les femmes cadres plus âgées associées aux pathologies tumorales et endocriniennes (à droite).

Interprétation de l'axe 4 (figure 5.18, quadrant bas–droit)

Du côté négatif :	*Du coté positif :*
► Au niveau des pathologies :	► Au niveau des pathologies :
- Système nerveux (14.95%)	- Digestif (67.53%)
► Au niveau des groupes d'agents :	► Au niveau des groupes d'agents :
- Hommes, cadres, +55 ans (7.04%)	- Hommes, maîtrise, 35–40 ans (37.98%)
- Femmes, maîtrise, 50–55 ans (4.81%)	- Femmes, exécution, 25–30 ans (11.32%)
- Femmes, cadres, +55 ans (4.46%)	
- Femmes, maîtrise, 25–30 ans (3.70%)	
- Femmes, cadres, 40–45 ans (3.29%)	
- Hommes, maîtrise, 40–45 ans (2.97%)	

Au niveau des familles de pathologies, le 4ème axe oppose les pathologies du système nerveux (en bas) aux pathologies digestives (en haut). Ces deux familles de pathologies contribuent ensemble à 82.48% de la variance de l'axe 4 tandis que leur écart contribue à 81.83% de cette même variance.

Au niveau des groupes d'agents, le 4ème axe oppose 6 groupes d'agents (en bas) : plutôt des cadres et agents de maîtrise, à 2 groupes d'agents (en haut) : plutôt des agents d'exécution et de maîtrise relativement jeunes. Ces 8 groupes d'agents contribuent ensemble à 75.57% de la variance de l'axe 4 tandis que leur écart contribue à 63.77% de cette même variance.

En résumé, l'axe 4 montre une opposition entre les cadres et agents de maîtrise associés aux pathologies du système nerveux (en bas) et les agents d'exécution et de maîtrise relativement jeunes associés aux pathologies digestives (en haut).

Les principales liaisons entre les groupes d'agents et les familles pathologies que nous pouvons déduire de l'interprétation des 4 premiers axes principaux sont mises en évidence (entourées) sur la figure 5.19.

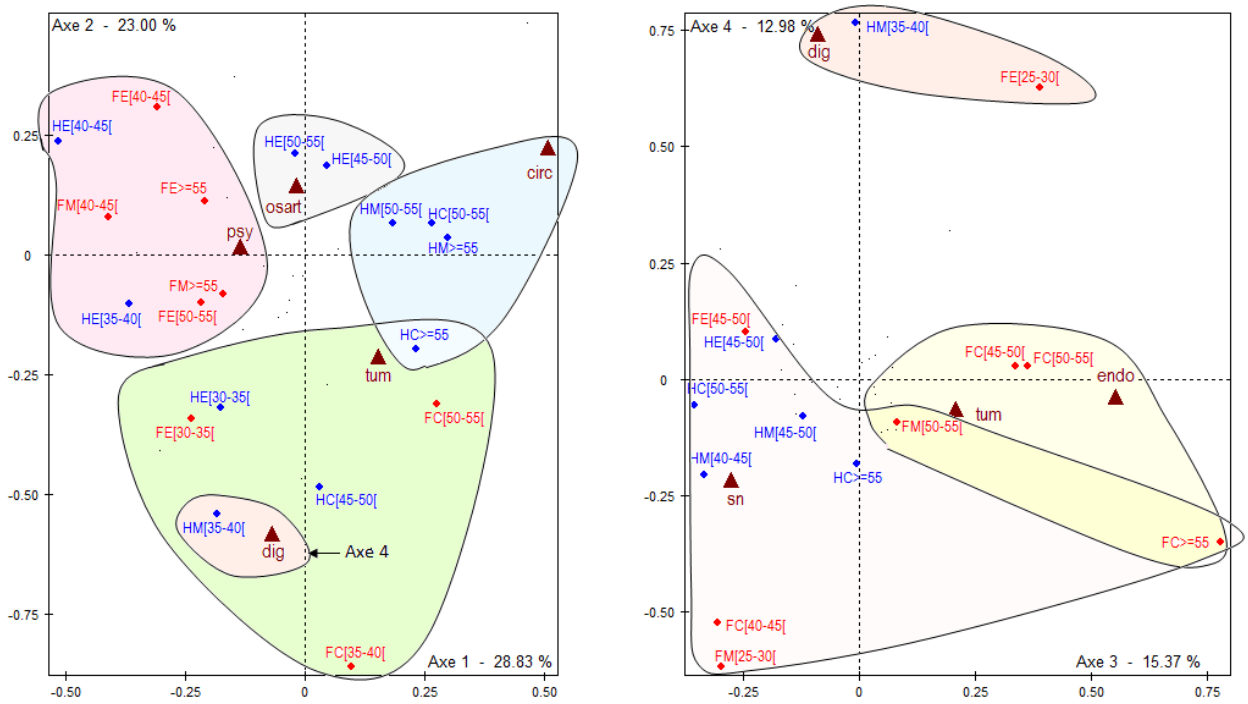


FIGURE 5.19 – Absentéisme LM : principales liaisons identifiées par l'analyse des correspondances.

Etude des groupes d'agents : différences selon le sexe, le collège et la classe d'âge. Sur le plan gauche de la figure 5.20, nous avons relié chaque groupe d'agents « hommes » au point moyen des groupes d'agents « hommes » et chaque groupe d'agents « femmes » au point moyen des groupes d'agents « femmes » (représentation « en étoile »). Nous avons fait de même pour chaque collège (au centre) et pour chaque classe d'âge (à droite). En complément, la table 5.1 contient la double décomposition de la variance pour les groupes déterminés par les variables *sexe*, *collège* et *classe d'âge* pour chacun des quatre premiers axes.

La variance inter des groupes déterminés par le sexe (hommes et femmes) et par le collège (exécution, maîtrise et cadres) est la plus importante sur les deux premiers axes principaux. Nous pouvons donc dire que les hommes se situent plutôt dans le quadrant droit–haut du premier plan principal (c'est–à–dire du côté positif de l'axe 1 et du côté positif de l'axe 2) tandis que les femmes sont plutôt dans le quadrant gauche–bas ; et qu'il existe un gradient exécution/maîtrise/cadres le long des axes 1 et 2 (exécution : en haut à gauche, maîtrise :

au centre, cadres : en bas à droite). En revanche, nous ne pouvons rien dire concernant la position des classes d'âge : les variances inter relatives à chaque axe sont assez proches.

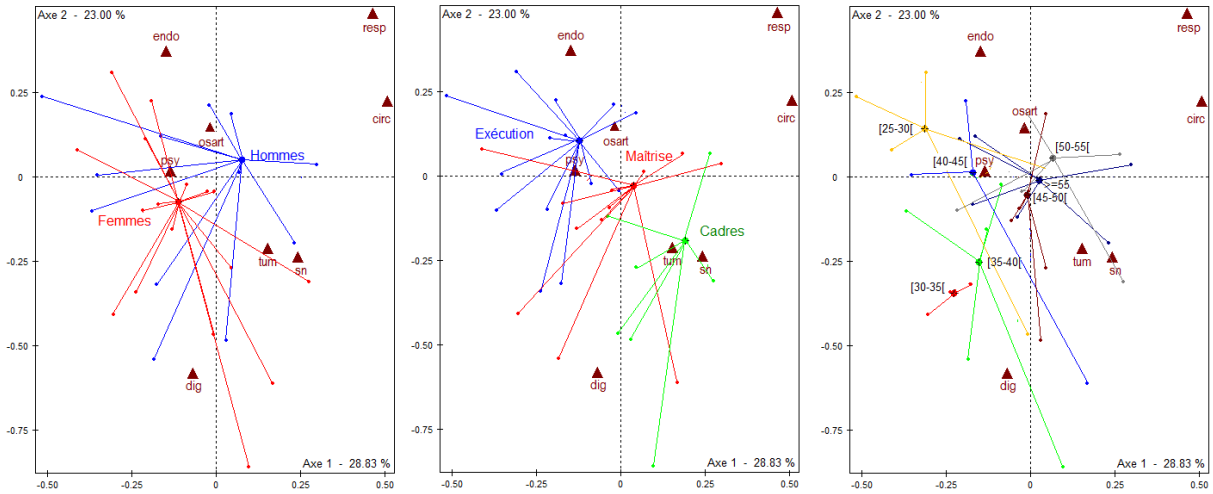


FIGURE 5.20 – Représentation « en étoile » des groupes hommes et femmes (à gauche), exécution, maîtrise, cadres (au centre) et de toutes les classes d'âge (à droite).

Sexe	Axe 1	Axe 2	Axe 3	Axe 4
Variance inter	0.0086	0.0037	0.0013	0.0008
Variance intra	0.0273	0.0250	0.0179	0.0154
η^2	24%	13%	7%	5%
Collège	Axe 1	Axe 2	Axe 3	Axe 4
Variance inter	0.0099	0.0080	0.0004	0.0014
Variance intra	0.0260	0.0206	0.0188	0.0148
η^2	28%	28%	2%	9%
Classe d'âge	Axe 1	Axe 2	Axe 3	Axe 4
Variance inter	0.0110	0.0092	0.0033	0.0035
Variance intra	0.0250	0.0194	0.0159	0.0127
η^2	30%	32%	17%	22%

TABLE 5.1 – Tableau de double décomposition de la variance selon chacun des quatre premiers axes, pour les groupes déterminés par le sexe, le collège et la classe d'âge.

Ces éléments mettent en évidence des différences entre les hommes et les femmes et entre les agents d'exécution, les agents de maîtrise et les cadres au niveau de l'absentéisme LM. Les classes d'âge, elles, sont moins discriminantes.

Commentaire :

En résumé, pour l'absentéisme de longue maladie et pour l'année 2010, les plans 1–2 et 3–4 de l'analyse des correspondances suggèrent les principales liaisons groupes d'agents/familles de pathologies suivantes :

1. Plan 1–2 :
   - Les LM motivées par des pathologies psychiatriques sont sur–représentées chez les hommes agents d'exécution de 35 à 45 ans, chez les femmes agents d'exécution de 40 à 45 ans et de plus de 50 ans, ainsi que chez les femmes agents de maîtrise de 40 à 45 ans et de plus de 55 ans. Pour l'ensemble des individus de ces 7 groupes, 68% des épisodes LM motivés par des pathologies psychiatriques sont dus à des troubles dépressifs et anxieux (ce taux est de 61% dans l'ensemble de la population).
   - Les LM motivées par des pathologies circulatoires sont sur–représentées chez les hommes agents de maîtrise et cadres de plus de 50 ans. Pour l'ensemble des individus de ces 4 groupes, 44% des épisodes LM motivés par des pathologies circulatoires sont dus à des cardiopathies ischémiques et 9% à des pathologies artérielles (ces taux sont respectivement de 31% et 5% dans l'ensemble de la population). Il s'agit d'affections classiquement rapportées au tabac.
   - Les LM dues à des pathologies ostéo–articulaires sont sur–représentées chez les hommes agents d'exécution de 45 ans à 55 ans. Pour les individus de ces 2 groupes d'agents, 51% des épisodes LM motivés par des pathologies ostéo–articulaires sont dus à des dorsopathies (ce taux est de 46% dans l'ensemble de la population).
   - Les LM motivées par des pathologies tumorales sont sur–représentées chez les femmes cadres de 35 à 40 ans et de 50 à 55 ans, mais aussi chez les hommes cadres de 45 à 50 ans et de plus de 55 ans et, dans une moindre mesure, chez les agents d'exécution âgés de 30 à 35 ans (hommes et femmes). Les tumeurs en cause sont la plupart du temps malignes, c'est–à–dire cancéreuses.
2. Plan 3–4 :
   - Les LM motivées par des pathologies tumorales sont également sur–représentées chez les femmes cadres de plus de 45 ans et chez les femmes maîtrise de 50 à 55 ans. Pour les femmes de ces 4 groupes, 49% des épisodes LM motivés par des pathologies tumorales sont dus à des cancers du sein (ce taux est de 24% dans l'ensemble de la population).
   - Les LM dues à des pathologies affectant le système nerveux sont sur–représentées chez les agents d'exécution de 45 à 50 ans (hommes et femmes), chez les femmes agents de maîtrise de 25 à 30 ans et de 50 à 55 ans, chez les femmes cadres de 40 à 45 ans et de plus de 55 ans, chez les hommes agents de maîtrise de 40 à 50 ans et chez les hommes cadres de plus de 55 ans. Pour ces 10 groupes d'agents, la pathologie du système nerveux la plus fréquente est la sclérose en plaque (37% des épisodes).
   - Les LM dues à des pathologies endocriniennes sont sur–représentées chez les femmes cadres de plus de 45 ans et chez les femmes agents de maîtrise de 50 à 55 ans. Pour les individus de ces 4 groupes d'agents, 38% des épisodes LM motivés par des pathologies endocriniennes sont dus à des pathologies surrénaliennes (ce taux est de 15% dans l'ensemble de la population). De même, pour les individus de ces 4 groupes, le diabète est la cause de 53% des épisodes LM motivés par des pathologies endocriniennes (ce taux est inférieur à celui observé dans l'ensemble de la population qui culmine à 73%).

– Les LM motivées par des pathologies digestives sont sur–représentées chez les hommes agents de maîtrise de 35 à 40 ans et les femmes agents d'exécution de 25 à 30 ans. Pour les individus de ces 2 groupes d'agents, 73% des épisodes LM motivés par des pathologies digestives sont dus à des pathologies intestinales (maladie de Crohn, rectocolite hémorragique ou autres) (ce taux est de 43% dans l'ensemble de la population).

Retour sur les groupes d'agents sensibles. L'analyse du taux d'absentéisme LM par groupes d'agents (p.142) a permis d'identifier les groupes suivants comme étant sensibles :

1. agents d'exécution, hommes et femmes, de plus de 40 ans,
2. agents de maîtrise, hommes et femmes, de plus de 50 ans,
3. cadres, hommes et femmes, de plus de 50 ans.

Grâce aux résultats de l'analyse des correspondances, nous pouvons identifier les principales familles de pathologies en cause dans l'absentéisme LM de ces groupes d'agents sensibles :

1. – femmes, agents d'exécution de plus de 40 ans : pathologies psychiatriques et affectant le système nerveux ;
   – hommes, agents d'exécution de plus de 40 ans : pathologies psychiatriques pour les plus jeunes d'entre eux, pathologies ostéo–articulaires pour les plus âgés ;
2. – femmes, agents de maîtrise de plus de 50 ans : pathologies psychiatriques et affectant le système nerveux ;
   – hommes, agents de maîtrise de plus de 50 ans : pathologies circulatoires ;
3. – femmes, cadres de plus de 50 ans : tumeurs et pathologies endocriniennes ;
   – hommes, cadres de plus de 50 ans : pathologies circulatoires et affectant le système nerveux.

Nous avons ici répondu à la question « *Qui s'absente pour quoi ?* » en longue maladie. L'identification des liaisons entre groupes d'agents et familles de pathologies en cause dans l'absentéisme LM pourrait permettre à l'entreprise d'orienter la prévention en mettant en place des campagnes spécifiques et ciblées.

4.1.3 Evolutions des liaisons

Il s'agit ici d'étudier l'évolution des principales liaisons entre groupes d'agents sensibles et familles de pathologies en cause dans l'absentéisme LM de 1995 à 2011. Pour ce faire, nous projetons les groupes d'agents des années 1995 à 2009 et 2011 (*cf.* figure 5.16, p.143), en tant qu'éléments supplémentaires, sur les axes principaux de l'analyse des correspondances précédente. Des trajectoires de groupes d'agents peuvent alors être considérées, nous les étudions dans le premier plan principal (optimisant la qualité de l'ajustement).

Remarque : Le tableau 5.16 (p.143) est en fait un tableau multiple à trois entrées (Groupes d'agents × Familles de pathologies × Année), que nous avons représenté en superposant les tableaux Groupes d'agents × Familles de pathologies de chaque année. Nous avons ici choisi d'effectuer l'analyse du tableau relatif à une année de référence (2010) et de considérer les

autres tableaux comme éléments supplémentaires. De nombreuses autres méthodes d'analyses de recherche des axes principaux existent pour le traitement particulier des tableaux multiples, nous en donnons un bref aperçu en annexe, page 180.

Etudions dans un premier temps les trajectoires moyennes des groupes hommes et femmes dans le premier plan principal [17].

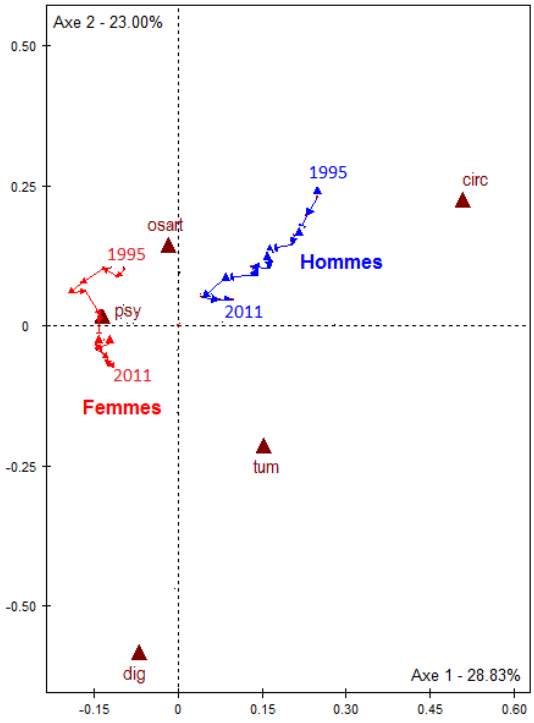


FIGURE 5.21 – Trajectoires des groupes hommes et femmes sur le premier plan principal (seules les familles de pathologies les plus contributives aux variances des deux premiers axes sont représentées).

Les hommes semblent se déplacer de la droite vers la gauche de l'axe 1 et du haut vers le bas de l'axe 2 ; les femmes semblent elles se déplacer du haut vers le bas de l'axe 2. Pour plus de précision et dans le but d'identifier des évolutions notables, nous mettons en oeuvre la procédure suivante, axe par axe :

1. évaluation de l'alignement des années selon l'axe considéré grâce à la corrélation des rangs de Spearman et au test de signification associé (corrélation entre les coordonnées des points et les années, *cf.* p.137) ;
2. évaluation de l'écart observé entre les deux années extrêmes grâce au calcul de l'écart calibré. L'écart calibré (d_{cal}) est un indice de l'importance de l'écart observé (d_{obs}), il est donné ici par le rapport entre la différence observée et l'écart–type, c'est–à–dire la racine carrée de la valeur propre associée à l'axe considéré. Nous considérons que l'écart observé est [18] :
   - négligeable si $d_{cal} < 0.4$,
   - notable si $0.4 \leq d_{cal} < 0.8$,

17. Le point « hommes–1995 » est le point moyen des groupes d'agents hommes de 1995 projetés sur le premier plan principal. Il en est de même pour tous les points de la trajectoire.

18. Ces seuils sont donnés dans Le Roux & Rouanet, 2004 [50].

– important si $d_{cal} \geq 0.8$.

Un test de Spearman significatif et un écart observé notable ou important entre les deux années extrêmes traduisent une réelle évolution.

Pour les trajectoires des hommes et des femmes, les résultats sont donnés dans les tableaux suivants :

	Hommes	Femmes
Axe 1	-0.96 S***	0.05 NS
Axe 2	-0.99 S***	-0.98 S***

TABLE 5.2 – Corrélation des rangs de Spearman (entre année et coordonnée) et significativité du test associé.

	Hommes			Femmes		
	d_{obs}	d_{cal}	Ecart	d_{obs}	d_{cal}	Ecart
Axe 1	-0.15	-0.80	important	-0.03	-0.17	négligeable
Axe 2	-0.20	-1.17	important	-0.16	-0.97	important

TABLE 5.3 – Écarts observés et calibrés entre 1995 et 2011.

De 1995 à 2011, les points moyens des hommes se situent tous dans le quadrant droit–gauche du premier plan principal. Pour toutes les années, nous pouvons donc dire qu'il existe une liaison entre l'absentéisme LM des hommes et les pathologies circulatoires et ostéo–articulaires. Cependant, l'analyse de la trajectoire des hommes montre qu'ils ont de plus en plus tendance à s'absenter également pour des pathologies psychiatriques et tumorales. En effet, en 1995, 47% des épisodes LM des hommes sont motivés par des pathologies psychiatriques et 7% par des pathologies tumorales, contre respectivement 51% et 15% en 2010.

Concernant les femmes, les points moyens des années 1995 à 2011 se situent tous du côté gauche du premier plan principal. Pour toutes les années, nous pouvons donc dire qu'il existe une liaison entre l'absentéisme LM des femmes et les pathologies psychiatriques. Cependant, l'analyse de la trajectoire des femmes met en évidence qu'elles ont de plus en plus tendance à s'absenter pour des pathologies tumorales (la trajectoire des femmes franchit l'axe 1). En effet, 11% des épisodes LM des femmes en 1995 sont motivés par des pathologies tumorales, contre 18% en 2010.

Etudions maintenant les trajectoires moyennes des groupes d'agents considérés comme étant sensibles. Les trajectoires pour lesquelles le test de Spearman est significatif et l'écart calibré est notable ou important selon l'axe 1 et/ou selon l'axe 2, sont représentées sur la figure 5.22. Les résultats sont résumés dans les tables 5.4 et 5.5.

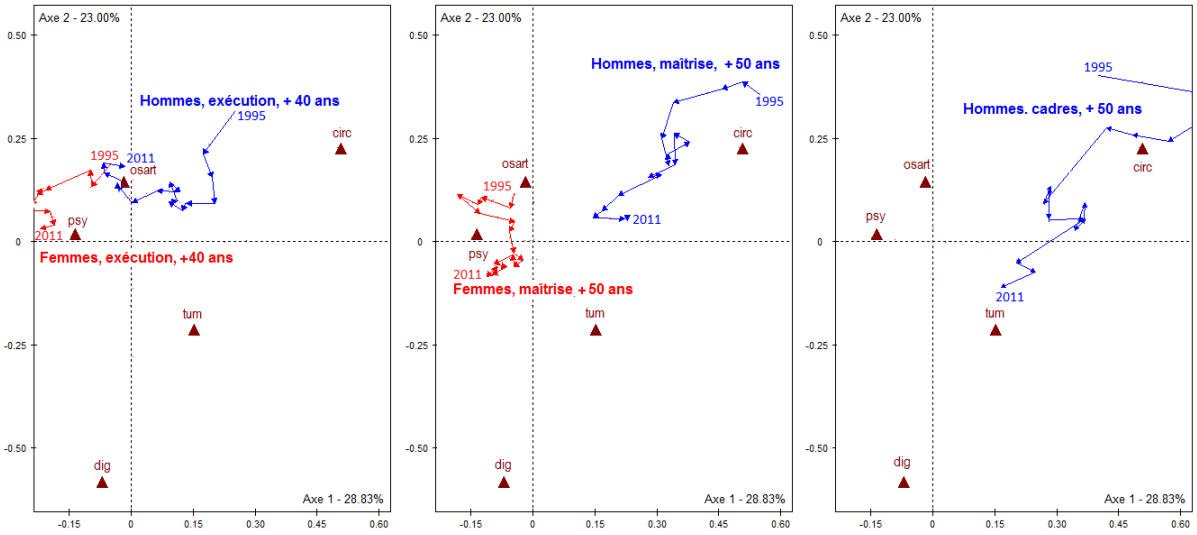


FIGURE 5.22 – Trajectoires des groupes sensibles sur le premier plan principal.

	Hommes Exécution + 40 ans	Femmes Exécution + 40 ans	Hommes Maîtrise + 50 ans	Femmes Maîtrise + 50 ans	Hommes Cadres + 50 ans	Femmes Cadres + 50 ans
Axe 1	-0.94 S***	-0.41 NS	-0.85 S***	0.04 NS	-0.77 S***	0.37 NS
Axe 2	0.14 NS	-0.90 S***	-0.94 S***	-0.95 S***	-0.94 S***	-0.12 NS

TABLE 5.4 – Corrélation des rangs de Spearman (entre année et coordonnée) et significativité du test associé.

	Hommes, Exécution, + 40 ans			Femmes, Exécution, + 40 ans		
	d_{obs}	d_{cal}	Ecart	d_{obs}	d_{cal}	Ecart
Axe 1	-0.27	-1.40	important	-0.17	-0.90	important
Axe 2	-0.13	-0.79	notable	-0.15	-0.91	important

	Hommes, Maîtrise, + 50 ans			Femmes, Maîtrise, + 50 ans		
	d_{obs}	d_{cal}	Ecart	d_{obs}	d_{cal}	Ecart
Axe 1	-0.31	-1.66	important	-0.07	-0.35	négligeable
Axe 2	-0.29	-1.72	important	-0.20	-1.18	important.

	Hommes, Cadres, + 50 ans			Femmes, Cadres, + 50 ans		
	d_{obs}	d_{cal}	Ecart	d_{obs}	d_{cal}	Ecart
Axe 1	-0.24	-1.25	important	0.56	2.97	important
Axe 2	-0.51	-3.03	important	-0.08	-0.47	négligeable

TABLE 5.5 – Écarts observés et calibrés entre 1995 et 2011.

Nous pouvons identifier les évolutions suivantes :

– Bien qu'étant principalement motivées par des pathologies psychiatriques, les LM des femmes, agents d'exécution de plus de 40 ans et agents de maîtrise de plus de

50 ans ont de plus en plus tendance a être motivées par des pathologies tumorales et digestives.
- Bien qu'étant principalement motivées par des pathologies ostéo–articulaires, les LM des hommes, agents d'exécution de plus de 40 ans sont de plus en plus motivées par des pathologies psychiatriques.
- Les LM des hommes, agents de maîtrise et cadres de plus de 50 ans ont de plus en plus tendance a être motivées par des pathologies psychiatriques et tumorales.

Toutes ces tendances doivent être prises en compte par l'entreprise dans le but d'adapter au mieux la prévention.

4.2 Maladie Courte durée

4.2.1 Le tableau de données

A partir des données de la cohorte EPIEG fusionnées avec les épisodes d'arrêt MCD des salariés, nous construisons le tableau soumis à l'analyse des correspondances de la façon suivante :

- en lignes : les groupes d'agents, déterminés par le croisement des variables Classe d'âge (8 modalités), Sexe (2 modalités), Collège (3 modalités) et dupliqués de 1995 à 2011 ;
- en colonnes : les grandes familles de pathologies (*cf.* p.132) ;
- dans chaque cellule : le nombre d'épisodes d'absence en MCD chez les agents du groupe ligne et motivés par la famille de pathologies colonne, dans l'année considérée (*cf.* tableau 5.16, p.143).

Remarque : Pour l'étude de la longue maladie, nous avions dans chaque cellule le nombre d'agents s'étant absentés et non le nombre d'épisodes. Cependant, comme aucun agent n'a plusieurs épisodes LM dans la même année, il était équivalent, pour la LM, de considérer l'une ou l'autre de ces quantités.

4.2.2 Analyse des correspondances (année 2010)

L'analyse est effectuée avec le logiciel Coheris SPAD©.

Il s'agit ici d'identifier des liaisons récentes. Pour ce faire, nous effectuons l'analyse des correspondances du tableau relatif à l'année 2010. L'année 2010 est donc prise comme référence.

Les huit familles de pathologies suivantes, prépondérantes en MCD, sont les colonnes retenues pour l'analyse :

- ostéo–articulaire (libellé court : osart),
- psychiatrie (psy),
- traumatologie (trauma),
- respiratoire (resp),
- tumeurs (tum),
- circulatoire (circ),
- digestif (dig),
- signes fonctionnels (sf).

Concernant les lignes, nous choisissons de retenir tous les groupes d'agents.

Remarque : Une sélection sur les lignes avait été faite pour l'analyse de la LM. Cependant, l'absentéisme MCD étant beaucoup plus fréquent que l'absentéisme LM, il n'est pas nécessaire ici d'effectuer cette sélection.

L'étude des variances des axes (tableau des valeurs propres ci–après) montre que les deux premiers axes principaux expliquent une part importante de la variance totale (77.93%). De plus, l'écart entre la 2ème et la 3ème valeur propre est supérieur à l'écart entre la 3ème et la 4ème valeur propre. Nous interprétons donc les deux premiers axes.

Tableau des valeurs propres
Trace de la matrice: 0.10609

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0,0463	43,66	43,66
2	0,0364	34,27	77,93
3	0,0118	11,09	89,02
4	0,0072	6,83	95,85
5	0,0017	1,60	97,46
6	0,0015	1,41	98,87
7	0,0012	1,13	100,00

La figure 5.23 suivante montre la représentation simultanée des groupes d'agents et des familles de pathologies dans le premier plan principal de l'analyse. Les pathologies (représentées par un triangle) sont identifiées par leur libellé court et les groupes d'agents par la concaténation de la première lettre du sexe (H : Homme, F : Femme), de la première lettre du collège (E : Exécution, M : Maîtrise, C : Cadre) et de la classe d'âge. Les groupes hommes sont représentés en bleu et les groupes femmes en rouge.

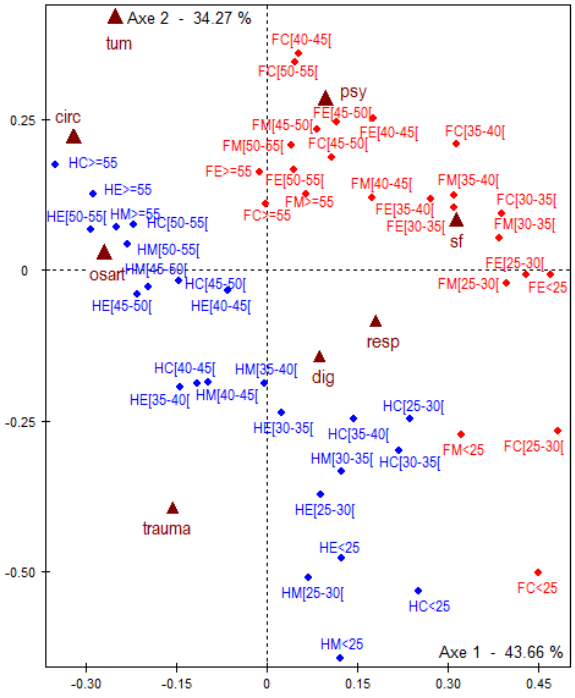


FIGURE 5.23 – Analyse des correspondances : plan 1–2.

Interprétation des axes : liaisons Groupes d'agents/Familles de pathologies. Interprétons maintenant chacun des deux premiers axes principaux à l'aide de la méthode des *contributions des points et des écarts* (*cf.* p.145). Ici nous ne retenons pour l'interprétation que les points dont la contribution est supérieure à $100/48 = 2.08\%$ pour les groupes d'agents et à $100/8 = 12.5\%$ pour les familles de pathologies (contributions supérieures à la contribution moyenne). La figure suivante représente les familles de pathologies et les groupes d'agents les plus contributifs aux variances des 2 premiers axes.

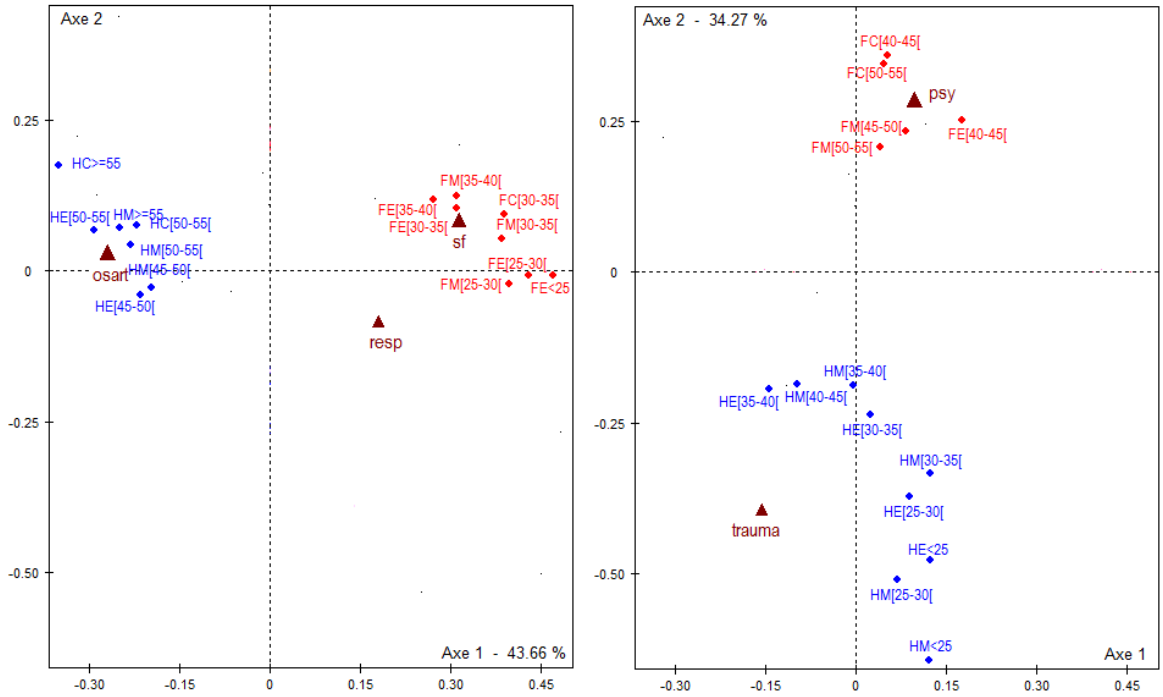


FIGURE 5.24 – Groupes d'agents et familles de pathologies les plus contributifs aux variances des 2 premiers axes : axe 1 (à gauche), axe 2 (à droite).

Interprétation de l'axe 1 (figure 5.24, gauche)

Du côté négatif :	*Du coté positif :*
► Au niveau des pathologies :	► Au niveau des pathologies :
- Ostéo–articulaire (contribution : 39.82%)	- Signes fonctionnels (22,03%)
	- Respiratoire (17.62%)
► Au niveau des groupes d'agents :	► Au niveau des groupes d'agents :
- Hommes, maîtrise, 50–55 ans (12.13%)	- Femmes, exécution, 25–30 ans (9.47%)
- Hommes, exécution, 50–55 ans (9.05%)	- Femmes, maîtrise, 30–35 ans (7.34%)
- Hommes, maîtrise, 45–50 ans (5.21%)	- Femmes, maîtrise, 35–40 ans (7.34%)
- Hommes, cadres, +55 ans (3.98%)	- Femmes, exécution, 30–35 ans (5.29%)
- Hommes, maîtrise, +55 ans (3.51%)	- Femmes, exécution, -25 ans (4.15%)
- Hommes, exécution, 45–50 ans (3.23%)	- Femmes, maîtrise, 25–30 ans (3.64%)
- Hommes, cadres, 50–55 ans (3.06%)	- Femmes, exécution, 35–40 ans (2.89%)
	- Femmes, cadres, 30-35 ans (2.21%)

Au niveau des familles de pathologies, le 1er axe oppose les pathologies ostéo–articulaires (à gauche) aux pathologies respiratoires et des signes fonctionnels (à droite). Ces trois familles

de pathologies contribuent ensemble à 79.47% de la variance de l'axe 1 tandis que leur écart contribue à 76.35% de cette même variance.

Au niveau des groupes d'agents, le 1er axe oppose 7 groupes d'agents (à gauche) : plutôt des hommes de plus de 45 ans, à 8 groupes d'agents (à droite) : plutôt des femmes de moins de 40 ans. Ces 15 groupes d'agents contribuent ensemble à 82.5% de la variance de l'axe 1 tandis que leur écart contribue à 78.02% de cette même variance.

En résumé, l'axe 1 montre une opposition entre les hommes de plus de 45 ans associés aux pathologies ostéo–articulaires (à gauche) et les femmes de moins de 40 ans associées aux pathologies respiratoires et des signes fonctionnels (à droite).

Interprétation de l'axe 2 (figure 5.24, droite).

Du côté négatif :	*Du coté positif :*
▶ Au niveau des pathologies :	▶ Au niveau des pathologies :
- Traumatologie (39.71%)	- Psychiatrie (31.35%)
▶ Au niveau des groupes d'agents :	▶ Au niveau des groupes d'agents :
- Hommes, maîtrise, 25–30 ans (9.98%)	- Femmes, maîtrise, 45–50 ans (6.46%)
- Hommes, exécution, 25–30 ans (9.53%)	- Femmes, maîtrise, 50–55 ans (6.33%)
- Hommes, maîtrise, -25 ans (8.94%)	- Femmes, cadres, 50–55 ans (3.11%)
- Hommes, exécution, -25 ans (8.04%)	- Femmes, exécution, 40–45 ans (2.65%)
- Hommes, maîtrise, 30–35 ans (7.09%)	- Femmes, cadres, 40–45 ans (2.37%)
- Hommes, exécution, 30–35 ans (4.15%)	
- Hommes, maîtrise, 40–45 ans (3.77%)	
- Hommes, maîtrise, 35–40 ans (3.14%)	
- Hommes, exécution, 35–40 ans (2.59%)	

Au niveau des familles de pathologies, le 2ème axe oppose les pathologies traumatologiques (en bas) aux pathologies psychiatriques (en haut). Ces deux familles de pathologies contribuent ensemble à 71.06% de la variance de l'axe 2 tandis que leur écart contribue à 70.92% de cette même variance.

Au niveau des groupes d'agents, le 2ème axe oppose 9 groupes d'agents (en bas) : plutôt des hommes de moins de 45 ans, à 5 groupes d'agents (en haut) : plutôt des femmes de plus de 45 ans. Ces 14 groupes d'agents contribuent ensemble à 78.15% de la variance de l'axe 2 tandis que leur écart contribue à 60.97% de cette même variance.

En résumé, l'axe 2 montre une opposition entre les hommes de moins de 45 ans associés aux pathologies traumatologiques (en bas) et les femmes de plus de 45 ans associées aux pathologies psychiatriques (en haut).

Les principales liaisons entre les groupes d'agents et les familles de pathologies que nous pouvons déduire de l'interprétation des 2 premiers axes principaux sont mises en évidence (entourées) sur la figure 5.25.

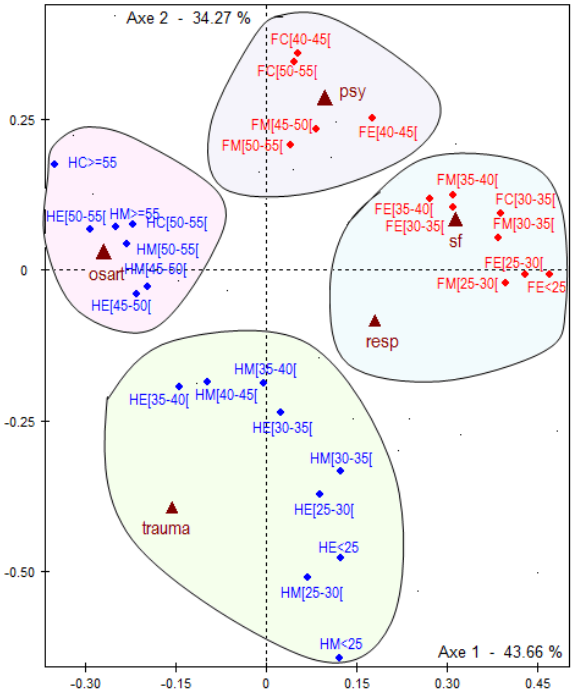


FIGURE 5.25 – Absentéisme MCD : principales liaisons identifiées par l'analyse des correspondances.

Remarque : Les mêmes classes sont obtenues en effectuant une classification ascendante hiérarchique des points projetés sur le premier plan principal.

Etude des groupes d'agents : différences selon le sexe, le collège et la classe d'âge. Sur le plan gauche de la figure 5.26, nous avons relié chaque groupe d'agents « hommes » au point moyen des groupes d'agents « hommes » et chaque groupe d'agents « femmes » au point moyen des groupes d'agents « femmes » (représentation « en étoile »). Nous avons fait de même pour chaque classe d'âge selon le sexe (au centre) et pour chaque collège (à droite). En complément, la table 5.6 contient la double décomposition de la variance pour les groupes déterminés par les variables *sexe*, *collège* et *classe d'âge* pour les deux premiers axes.

La variance inter des groupes déterminés par le sexe (hommes et femmes) est la plus importante sur l'axe 1, celle des groupes déterminés par la classe d'âge est importante sur l'axe 1, mais aussi sur l'axe 2. Pour le collège, les variances inter relatives à chaque axe sont négligeables.

Nous pouvons donc dire que le premier axe de l'analyse oppose les hommes (à gauche) aux femmes (à droite) (figure 5.26, gauche) et qu'il existe un alignement des classes d'âge le long de l'axe 1 et de l'axe 2 (classes d'âge jeunes en bas à droite et plus âgées en haut à gauche) (figure 5.26, centre). Nous ne pouvons rien dire concernant la position des trois collèges (figure 5.26, droite).

Ces graphiques mettent en évidence des différences entre les hommes et les femmes, mais aussi entre les classes d'âge au niveau de l'absentéisme de courte durée. Le collège, lui, est moins discriminant.

Remarque : Les résultats étaient différents pour l'absentéisme LM où seuls le sexe et le collège étaient discriminants (*cf.* p.150).

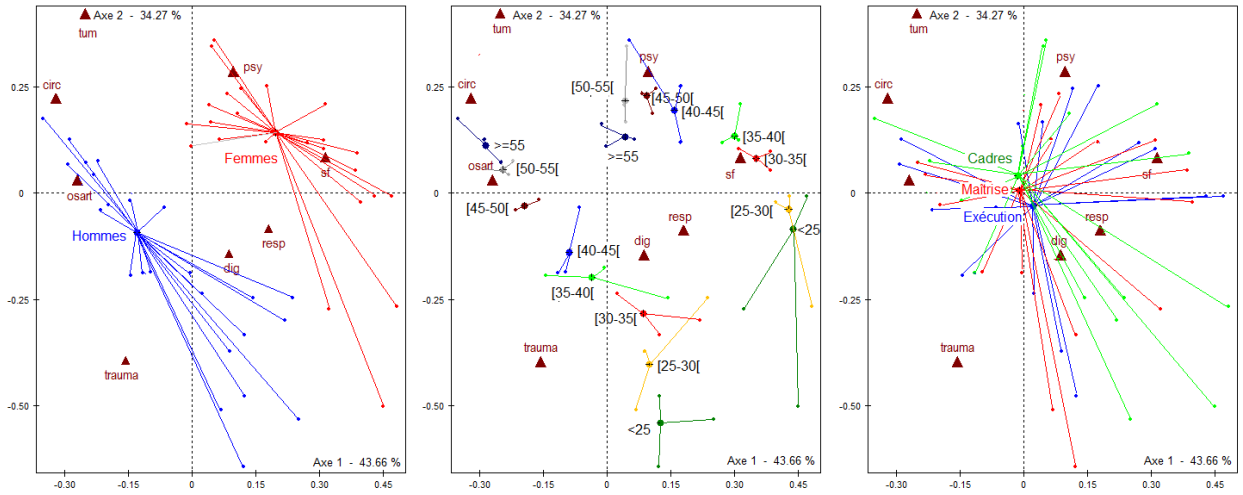


FIGURE 5.26 – Représentation « en étoile » des groupes hommes et femmes (à gauche), de toutes les classes d'âge pour chaque sexe (au centre) et des exécution, maîtrise, cadres (à droite).

Sexe	Axe 1	Axe 2
Variance inter	0.0254	0.0132
Variance intra	0.0209	0.0232
η^2	55%	36%
Classe d'âge	Axe 1	Axe 2
Variance inter	0.0243	0.0152
Variance intra	0.0220	0.0211
η^2	52%	42%
Collège	Axe 1	Axe 2
Variance inter	0.0002	0.0006
Variance intra	0.0461	0.0358
η^2	0.5%	1%

TABLE 5.6 – Tableau de double décomposition de la variance selon les deux premiers axes, pour les groupes déterminés par le sexe, la classe d'âge et le collège.

Commentaire :

En résumé, pour l'absentéisme de courte durée et pour l'année 2010, le premier plan principal de l'analyse des correspondances suggère les principales liaisons groupes d'agents/familles de pathologies suivantes :

- Les absences de courte durée motivées par des pathologies respiratoires et des pathologies du chapitre « signes fonctionnels » (correspondent à des signes de dysfonctionnement et à des plaintes du patient sans qu'il ne soit encore posé de réel diagnostic) sont sur–représentées chez les femmes de moins de 40 ans de chaque collège.
- Les absences de courte durée dues à des pathologies psychiatriques sont sur–représentées chez les femmes de plus de 40 ans de chaque collège.

- Les absences de courte durée motivées par de la traumatologie sont sur–représentées chez les hommes de moins de 40 ans, agents d'exécution et de maîtrise.
- Les absences de courte durée dues à des pathologies ostéo–articulaires sont sur–représentées chez les hommes de plus de 40 ans, agents d'exécution et de maîtrise, mais aussi chez les hommes cadres de plus de 55 ans.

Nous avons ici répondu à la question *Qui s'absente pour quoi ?* en maladie de courte durée. L'absentéisme MCD étant prépondérant dans les IEG (*cf.* figure 5.10, p.134), l'identification des liaisons entre groupes d'agents et familles de pathologies en cause dans l'absentéisme MCD est importante pour l'entreprise.

4.2.3 Evolutions des liaisons

L'étude des trajectoires des groupes d'agents sur le premier plan principal n'a pas permis d'identifier de réelle évolution de l'absentéisme MCD de 1995 à 2011. A titre d'exemple, nous avons tracé sur la figure suivante les trajectoires moyennes des groupes d'agents hommes et femmes. Nous constatons qu'elles sont assez condensées et que les années ne sont pas alignées. Depuis 1995, l'absentéisme MCD des femmes est donc principalement motivé par des pathologies psychiatriques et des signes fonctionnels tandis que celui des hommes est essentiellement motivé par des pathologies ostéo–articulaires et traumatologiques.

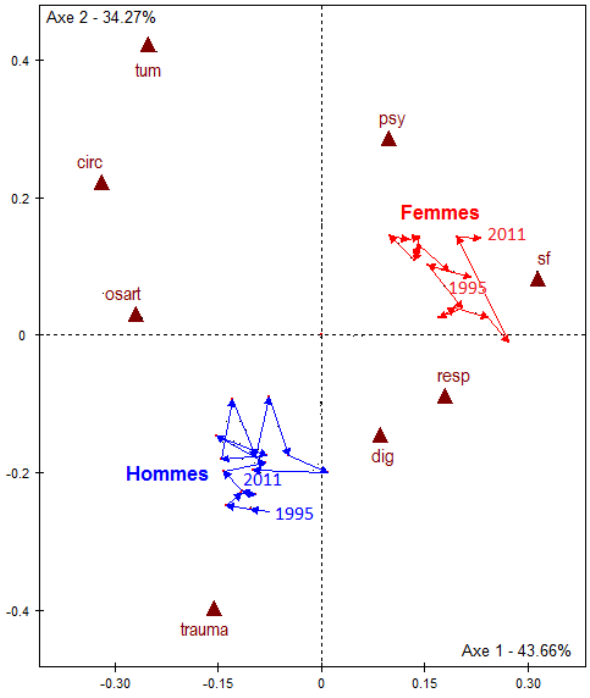


FIGURE 5.27 – Trajectoires des groupes hommes et femmes sur le premier plan principal.

Conclusion

Aboutissement d'une construction longue et fastidieuse, la cohorte EPIEG est un formidable outil pour le SGMC : elle contient (depuis 1995), les données socio–démographiques annuelles des milliers de salariés statutaires des IEG. La connaissance précise des embauches et des sorties de l'emploi permet d'être au plus près de l'effectif exact. Le croisement de la cohorte EPIEG avec les fichiers de données de santé du SGMC (épisodes d'arrêts notamment) permet d'étudier de façon longitudinale l'état de santé de la population salariale. Des études relatives à des sous–populations peuvent également être fournies. Nous avons par exemple étudié l'évolution de l'absentéisme dans une des directions d'EDF–SA de 2007 à 2011.

Dans ce chapitre, nous avons choisi d'étudier l'évolution de la santé de la totalité de la population salariale des IEG. De 1995 à 2011, le taux d'absentéisme global passe de 3.44% à 4.29%. Nous avons d'abord montré que cette augmentation est surtout la conséquence de l'augmentation du taux d'absentéisme dû aux longues maladies. En effet, ce taux a presque triplé de 1995 à 2011, tandis que les taux d'absentéisme dus aux maladies de courtes durée, aux accidents du travail/maladies professionnelles sont stables. L'absence étant fortement dépendante de la structure de la population (notamment de l'âge, du sexe et du collège), nous avons vérifié que les tendances observées ne sont pas dues à l'évolution de cette structure (population vieillissante, comprenant de plus en plus de femmes et de plus en plus de cadres).

Puis, plusieurs analyses ont permis d'identifier des familles de pathologies prépondérantes, des groupes d'agents sensibles et les liaisons qui existent entre ces groupes d'agents et ces familles de pathologies, d'une part pour l'absentéisme de longue maladie, et d'autre part pour l'absentéisme de courte durée. Concernant la longue maladie (identifiée comme étant la cause principale de l'augmentation du taux d'absentéisme global), la psychiatrie, l'ostéo–articulaire et les tumeurs sont les trois familles de pathologies prépondérantes, les taux d'absentéisme dus à chacune de ces pathologies sont en nette augmentation depuis 1995. Les groupes sensibles, c'est–à–dire présentant un absentéisme important et/ou croissant (en longue maladie) sont les agents d'exécution de plus de 40 ans, les agents de maîtrise de plus de 50 ans et les cadres de plus de 50 ans. Une analyse des correspondances, effectuée sur les données de longue maladie de l'année 2010, a permis d'identifier les liaisons groupes d'agents sensibles/familles de pathologies suivantes :

1. – femmes, agents d'exécution de plus de 40 ans : pathologies psychiatriques et affectant le système nerveux ;
   – hommes, agents d'exécution de plus de 40 ans : pathologies psychiatriques pour les plus jeunes d'entre eux, pathologies ostéo–articulaires pour les plus âgés ;
2. – femmes, agents de maîtrise de plus de 50 ans : pathologies psychiatriques et affectant le système nerveux ;
   – hommes, agents de maîtrise de plus de 50 ans : pathologies circulatoires ;
3. – femmes, cadres de plus de 50 ans : tumeurs et pathologies endocriniennes ;
   – hommes, cadres de plus de 50 ans : pathologies circulatoires et affectant le système nerveux.

D'autres analyses (trajectoires projetées sur le premier plan principal, *cf.* p.153) montrent que l'absentéisme de longue maladie des hommes a évolué depuis 1995 : il reste principalement lié à des pathologies circulatoires et ostéo–articulaires mais est de plus en plus motivé par des pathologies psychiatriques et tumorales. De même, l'absentéisme de longue maladie des femmes a évolué depuis 1995 : il reste principalement lié à des pathologies psychiatriques mais est de plus en plus motivé par des pathologies tumorales. L'étude des trajectoires des

groupes sensibles donne des résultats semblables (*cf.* p.155).

Ces méthodes pourraient être utilisées par le SGMC pour l'étude de sous-populations spécifiques. Les managers, services de ressources humaines et de santé pourraient alors appréhender au mieux la construction de leurs dispositifs d'actions de prévention.

Conclusion générale

Dans la première partie de cette thèse, nous avons présenté trois tests : le test géométrique de typicalité (comparaison d'un point moyen à un point de référence, chapitre 2), le test ensembliste de typicalité (comparaison d'un groupe d'observations à une population de référence, chapitre 3) et le test d'homogénéité (comparaison de plusieurs groupes indépendants, chapitre 4). Pour ces trois tests, nous construisons d'abord un *espace de référence* par rapport auquel nous situons les données. La construction de cet espace relève de techniques combinatoires, la plupart du temps permutationnelles. Ainsi pour le test géométrique de typicalité, l'*espace de typicalité* est engendré par un processus de symétrie (ou, dans le cas particulier de deux groupes appariés, par échange des individus entre les groupes) ; pour le test ensembliste de typicalité, l'*espace des échantillons* est construit par un processus d'échantillonnage dans la population de référence ; et pour le test d'homogénéité, l'*espace des emboîtements* est obtenu par échange des individus entre les groupes. Il s'agit ensuite de déterminer la proportion des éléments de l'espace de référence plus extrêmes que les données, selon une certaine statistique de test.

Ces tests s'appliquent à des nuages euclidiens à valeurs dans des espaces multidimensionnels. Les statistiques de test, indices de typicalité ou d'homogénéité, sont toutes fonction de distances dérivées de la distance de Mahalanobis prenant en compte la structure de covariance du nuage soumis au test. Le choix de la statistique de test et des distances employées est crucial, il conduit à des prolongements géométriques originaux : interprétation géométrique du seuil observé et construction d'une zone géométrique de compatibilité des données avec l'hypothèse de typicalité ou d'homogénéité.

D'un point de vue pratique, nous effectuons les calculs exacts pour des petites tailles de données. Pour des tailles intermédiaires, dès que le nombre d'éléments de l'espace de référence devient trop grand, nous calculons une estimation du seuil observé par la méthode de Monte Carlo. Pour des gros volumes de données, des approximations basées sur des distributions gaussiennes ou du χ^2 peuvent être utilisées. Des programmes R ont été écrits dans ce sens pour les trois tests.

Les tests combinatoires mis en oeuvre dans cette thèse et le cadre d'interprétation sous–jacent sont libres de toute hypothèse concernant le processus d'obtention des données (échantillonnage au hasard, randomisation, observations indépendamment et identiquement distribuées, distributions gaussiennes, *etc.*). Les conclusions sont données en termes d'atypicalité du point moyen par rapport au point de référence (test géométrique de typicalité), d'atypicalité du groupe d'observations par rapport à la population de référence (test ensembliste de typicalité) ou d'hétérogénéité des groupes (test d'homogénéité). Elles portent donc exclusivement sur les données et ne s'étendent pas directement à des populations dont les données pourraient être issues. En ce sens, l'inférence combinatoire peut être considérée comme le premier degré de l'inférence. Elle est toujours applicable et présente un intérêt certain dans beaucoup d'applications lorsque les hypothèses requises pour le cadre fréquentiste conven-

tionnel ne sont pas valides et/ou satisfaites.

Les tests combinatoires peuvent également s'interpréter comme des *tests du hasard* (*cf.* Rouanet, Bernard & Le Roux, 1990, p.174–176 [72]). En effet, pour le test d'homogénéité l'hypothèse du hasard à tester est : « l'affectation des individus dans les groupes à comparer est assimilable à une affectation au hasard ». Sous l'hypothèse du hasard, le seuil observé (indice d'homogénéité) s'interprète comme la *probabilité* de construire (à partir des données) une affectation au moins aussi extrême que l'affectation observée selon la statistique de test choisie. Si cette probabilité est petite, nous pouvons rejeter l'hypothèse du hasard : nous concluons que l'affectation observée n'est pas assimilable à une affectation au hasard ; nous sommes alors fondés, pour l'interprétation, à prendre en compte l'hétérogénéité des groupes. De même, pour le test géométrique de typicalité, l'hypothèse du hasard est : « l'écart observé entre le point de référence O et le point moyen G est dû au hasard » et pour le test ensembliste de typicalité : « le groupe d'observations est assimilable à un échantillon au hasard de taille n de la population de référence ». Le test du hasard permet de convertir des proportions (permettant une interprétation combinatoire) en probabilités (permettant une interprétation probabiliste). Il peut donc être considéré comme étant la première marche menant à l'inférence probabiliste.

Des progrès extraordinaires ont été faits en termes de puissance de calcul au cours de ces dernières décennies ; il serait maintenant absurde de se borner aux procédures inférentielles traditionnelles et de négliger celles basées sur des algorithmes calculatoires. En effet, l'analyse mathématique supposant des hypothèses fortes (pourtant simplificatrices) est aujourd'hui remplaçable par des méthodes plus simples et moins contraignantes mais impliquant en contrepartie d'intensifs calculs. Tukey écrit en 1986 ([88]) : « In a world in which the price of calculation continues to decrease rapidly, but the price of theorem proving continues to hold steady or increase, elementary economics indicates that we ought to spend a larger and larger fraction of our time on calculation ». En ce sens, les procédures combinatoires ont toute leur place dans une perspective inférentielle. Cependant, elles ont le même défaut que les procédures inférentielles classiques : elles dépendent de la taille des données et peuvent mener à des conclusions contradictoires. Ainsi, si la taille des données est importante, un effet descriptif négligeable peut mener à un test combinatoire significatif, ce qui ne veut pas dire que l'effet est important. D'où l'importance de la phase descriptive qui doit être réalisée en amont de la phase inférentielle : en bonne méthodologie, un test statistique ne devrait être effectué que si l'écart descriptif observé est notable ou important. On vérifie ensuite grâce à la phase inférentielle que le nombre de données est suffisant pour conclure à un effet significatif. La phase inférentielle constitue donc un véritable garde–fou et permet de s'assurer que l'interprétation des résultats n'est pas basée sur des données trop peu nombreuses.

Dans la deuxième partie de cette thèse, nous avons présenté une étude concernant l'absentéisme de la population salariale des Industries Electriques et Gazières entre 1995 et 2011. La cohorte EPIEG, construite en 2009–2010 et mise à jour chaque année, a notamment permis la visualisation précise du taux d'absentéisme ventilé par types d'absence (maladie de courte durée, longue maladie, accidents du travail/maladies professionnelles) de 1995 à 2011. Nous avons ainsi montré que la tendance croissante du taux d'absentéisme global est en premier lieu la conséquence de l'augmentation du taux d'absentéisme de longue maladie. Nous avons également identifié des pathologies émergentes et des groupes d'agents sensibles pour la longue maladie mais aussi pour la maladie de courte durée. Ces éléments pourraient

permettre aux entreprises des IEG de mieux comprendre l'absentéisme de leurs salariés afin de mettre en place des campagnes de prévention spécifiques et ciblées. Une conclusion propre à cette partie figure page 163.

Plusieurs prolongements et compléments peuvent être apportés à ce travail de thèse :

- Concernant l'étude menée au Service Général de Médecine de Contrôle, l'idée première était de traiter les données individuelles de la cohorte EPIEG à l'aide des méthodes d'analyses géométriques telles que l'ACM ou l'ACP. Cependant, les données décrivant les individus par le nombre ou la durée de leurs épisodes d'absence conduisent à des tableaux « creux » : la très grande majorité des individus s'absente peu. Par conséquent, les analyses géométriques sur données individuelles n'apportent rien de concluant. Pour mettre en évidence des phénomènes de groupes, nous avons pensé à agréger les données afin de les traiter grâce à l'analyse des correspondances. Cette analyse s'est avérée être beaucoup plus adaptée pour aborder ce type de données. Mais, les nuages issus d'une analyse des correspondances sont pondérés, ils n'ont donc pas permis de mettre en oeuvre les tests combinatoires développés dans la première partie de cette thèse puisqu'ils s'appliquent pour l'instant à des nuages équipondérés. Un premier prolongement pourrait donc consister à construire des tests combinatoires pouvant être effectués à partir de nuages euclidiens pondérés.
- Des tests combinatoires pourraient également être construits pour certains cas classiques rencontrés en analyse de la variance. En particulier dans le cas du croisement de deux facteurs, un second prolongement du travail présenté ici pourrait consister à mettre en place un test combinatoire permettant de tester les effets principaux et les effets d'interaction.
- Concernant le test d'homogénéité, des zones de compatibilité ont été définies pour la comparaison (globale ou partielle) de deux groupes. Un troisième prolongement serait de construire des zones de compatibilité géométriques pour la comparaison de plus de deux groupes.
- L'étude de la puissance des trois tests combinatoires présentés ici est un problème ouvert auquel nous n'avons pour l'instant pas répondu. C'est un travail théorique qui reste à effectuer.

Annexes

Annexe A

Les bases de données du SGMC

Grandes régions	Régions administratives
Ile de France	Ile de France
Nord–Ouest	Basse–Normandie Bretagne Centre Haute–Normandie Pays–de–la–Loire
Nord–Est	Alsace Bourgogne Champagne–Ardennes Franche–Comté Lorraine
Sud–Est	Auvergne Corse Languedoc–Roussillon Paca Rhône–Alpes
Sud–Ouest	Aquitaine Limousin Midi–Pyrénées Poitou–Charentes
Outre–mer	France d'outre mer

TABLE A.1 – Répartition des régions administratives françaises dans les six « grandes régions ».

Variables populationnelles non évolutives			
Nom	*Libellé*	*Rôle*	*Stockage*
id	Identifiant	Identifiant	Chaîne
sexe	Sexe	Nominale	Chaîne
date_nais	Date de naissance	Nominale	Date
date_emb	Date d'embauche dans les IEG	Nominale	Date
age_emb	Age à l'embauche	Continue	Réel
diplome_emb	Niveau de diplôme à l'embauche	Nominale	Chaîne
gf_emb	Groupe fonctionnel à l'embauche	Nominale	Chaîne
college_emb	Collège à l'embauche	Nominale	Chaîne
ue_emb	Unité d'emploi à l'embauche	Nominale	Chaîne
structure_emb	Structure organisationnelle d'emploi à l'embauche	Nominale	Chaîne
motif_depart	Motif de sortie de la cohorte	Nominale	Chaîne
date_depart	Date de sortie de la cohorte	Nominale	Date
age_depart	Age lors de la sortie de la cohorte	Continue	Réel

Variables populationnelles évolutives			
Nom	*Libellé*	*Rôle*	*Stockage*
annee	Année considérée	Continue	Entier
age	Age	Continue	Réel
classage	Classe d'âge	Nominale	Chaîne
sit_fam	Situation familiale	Nominale	Chaîne
nb_enfant	Nombre d'enfants	Nominale	Chaîne
dpt_dom	Département d'habitation	Nominale	Entier
dpt_trav	Département de lieu de travail	Nominale	Entier
reg_dom	Région INSEE d'habitation	Nominale	Chaîne
reg_trav	Région INSEE de lieu de travail	Nominale	Chaîne
reg_trav_tel	Région de lieu de travail (selon la répartition des indices téléphoniques)	Nominale	Chaîne
gf	Groupe fonctionnel	Nominale	Chaîne
college	Collège	Nominale	Chaîne
ue	Unité d'emploi	Nominale	Chaîne
structure	Structure organisationnelle d'emploi	Nominale	Chaîne
coef	Durée de suivi (en année) sur l'année considérée	Continue	Réel

TABLE A.2 – Noms, libellés, rôles et stockages des variables de la cohorte EPIEG.

Chapitres de la nomenclature SGMC	
Chapitre 1	Certaines maladies infectieuses et parasitaires
Chapitre 2	Tumeurs
Chapitre 3	Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire
Chapitre 4	Maladies endocriniennes, nutritionnelles et métaboliques
Chapitre 5	Troubles mentaux et du comportement
Chapitre 6	Maladies du système nerveux
Chapitre 7	Maladies de l'oeil et de ses annexes
Chapitre 8	Maladies de l'oreille et de l'apophyse mastoïde
Chapitre 9	Maladies de l'appareil circulatoire
Chapitre 10	Maladies de l'appareil respiratoire
Chapitre 11	Maladies de l'appareil digestif
Chapitre 12	Maladies de la peau et du tissu cellulaire sous-cutané
Chapitre 13	Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif
Chapitre 14	Maladies de l'appareil génito-urinaire
Chapitre 15	Grossesse, accouchement et période puerpérale
Chapitre 17	Malformations congénitales et anomalies chromosomiques
Chapitre 18	Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoires non classés ailleurs
Chapitre 19	Morts mal définies, empoisonnements, suicides tentatives de suicides
Chapitre 20	Traumatologie et Causes de morbidité et de mortalité : travail et vie courante
Chapitre 21	Facteurs influant l'état de santé et motif de recours aux services de santé
Chapitre 22	Maladies professionnelles

TABLE A.3 – Chapitres constitutifs de la nomenclature utilisée par le SGMC pour le codage des pathologies (basée sur la CIM–10).

Chapitres de la CIM–10	
Chapitre 1	Certaines maladies infectieuses et parasitaires
Chapitre 2	Tumeurs
Chapitre 3	Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire
Chapitre 4	Maladies endocriniennes, nutritionnelles et métaboliques
Chapitre 5	Troubles mentaux et du comportement
Chapitre 6	Maladies du système nerveux
Chapitre 7	Maladies de l'oeil et de ses annexes
Chapitre 8	Maladies de l'oreille et de l'apophyse mastoïde
Chapitre 9	Maladies de l'appareil circulatoire
Chapitre 10	Maladies de l'appareil respiratoire
Chapitre 11	Maladies de l'appareil digestif
Chapitre 12	Maladies de la peau et du tissu cellulaire sous-cutané
Chapitre 13	Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif
Chapitre 14	Maladies de l'appareil génito-urinaire
Chapitre 15	Grossesse, accouchement et période puerpérale
Chapitre 16	Certaines affections dont l'origine se situe dans la période périnatale
Chapitre 17	Malformations congénitales et anomalies chromosomiques
Chapitre 18	Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoires non classés ailleurs
Chapitre 19	Lésions traumatiques, empoisonnements et certaines autres conséquence de causes externes
Chapitre 20	Causes externes de morbidité et de mortalité
Chapitre 21	Facteurs influant l'état de santé et motif de recours aux services de santé

TABLE A.4 – Chapitres constitutifs de la CIM–10.

Année	Sexe	Collège	Classe d'âge	Taux d'absentéisme LM (%)	Taux d'absentéisme MCD (%)
2010	Hommes	Exécution	[18-30[	0,1304	2,20894
2010	Hommes	Exécution	[30-40[	0,7177	2,91146
2010	Hommes	Exécution	[40-50[	2,3175	3,93527
2010	Hommes	Exécution	[50-70[	6,7202	5,13975
2010	Hommes	Maîtrise	[18-30[	0,0342	1,14775
2010	Hommes	Maîtrise	[30-40[	0,2673	1,44908
2010	Hommes	Maîtrise	[40-50[	0,6495	2,00983
2010	Hommes	Maîtrise	[50-70[	2,6216	2,78144
2010	Hommes	Cadres	[18-30[	0,0213	0,32387
2010	Hommes	Cadres	[30-40[	0,0466	0,62334
2010	Hommes	Cadres	[40-50[	0,2103	0,94156
2010	Hommes	Cadres	[50-70[	0,9259	1,31284
2010	Femmes	Exécution	[18-30[	0,9665	7,32776
2010	Femmes	Exécution	[30-40[	2,6624	8,12457
2010	Femmes	Exécution	[40-50[	5,9897	7,91707
2010	Femmes	Exécution	[50-70[	11,9767	6,6052
2010	Femmes	Maîtrise	[18-30[	0,2653	3,23351
2010	Femmes	Maîtrise	[30-40[	0,6455	3,82296
2010	Femmes	Maîtrise	[40-50[	2,0409	3,84849
2010	Femmes	Maîtrise	[50-70[	5,0283	4,67104
2010	Femmes	Cadres	[18-30[	0,126	0,87701
2010	Femmes	Cadres	[30-40[	0,3267	2,0389
2010	Femmes	Cadres	[40-50[	0,7044	1,86014
2010	Femmes	Cadres	[50-70[	1,9181	2,65301
2011	Hommes	Exécution	[18-30[	0,1374	2,39861
2011	Hommes	Exécution	[30-40[	0,5445	2,93074
2011	Hommes	Exécution	[40-50[	2,2564	4,05255
2011	Hommes	Exécution	[50-70[	7,1878	4,75578
2011	Hommes	Maîtrise	[18-30[	0,0294	1,06038
2011	Hommes	Maîtrise	[30-40[	0,2514	1,63753
2011	Hommes	Maîtrise	[40-50[	0,592	2,07006
2011	Hommes	Maîtrise	[50-70[	2,464	2,9154
2011	Hommes	Cadres	[18-30[	0	0,34364
2011	Hommes	Cadres	[30-40[	0,1029	0,74192
2011	Hommes	Cadres	[40-50[	0,2398	0,94081
2011	Hommes	Cadres	[50-70[	0,9663	1,38944
2011	Femmes	Exécution	[18-30[	0,7596	7,09811
2011	Femmes	Exécution	[30-40[	3,0544	8,15915
2011	Femmes	Exécution	[40-50[	5,9542	7,68069
2011	Femmes	Exécution	[50-70[	12,3834	6,80933
2011	Femmes	Maîtrise	[18-30[	0,1687	3,22497
2011	Femmes	Maîtrise	[30-40[	0,6064	4,33974
2011	Femmes	Maîtrise	[40-50[	2,2456	4,10722
2011	Femmes	Maîtrise	[50-70[	5,1646	4,61445
2011	Femmes	Cadres	[18-30[	0,019	1,04714
2011	Femmes	Cadres	[30-40[	0,3571	2,2836
2011	Femmes	Cadres	[40-50[	0,5611	2,17101
2011	Femmes	Cadres	[50-70[	2,0423	2,59912

TABLE A.5 – Taux d'absentéisme relatifs à la longue maladie et à la maladie de courte durée, par groupes d'agents, en 2010 et 2011.

Annexe B

Analyse Géométrique des Données, compléments

1 Analyse géométrique des données

1.1 Principales méthodes

Les trois paradigmes de l'Analyse Géométrique des Données que sont l'Analyse en Composantes Principales (ACP), l'Analyse des Correspondances (AC) et l'Analyse des Correspondances Multiples(ACM) ont toutes la même finalité et reposent toutes sur le même principe : elles visent à synthétiser l'information contenue dans un vaste tableau de données en réduisant la dimension de ce tableau. Conjointement à cette réduction de dimension, des visualisations graphiques, essentielles pour ces méthodes, permettent de représenter les associations entre les lignes et/ou les colonnes du tableau.

Le tableau de données est donc ici un tableau à deux entrées. Le type de tableau soumis à l'analyse détermine la méthode d'analyse géométrique à utiliser. Les trois principales méthodes sont décrites ci–après :

- **AC.** Lorsque le tableau de données est un tableau de contingence (croisement de deux variables nominales), l'*Analyse des Correspondances* permet d'étudier les relations existant entre les catégories des deux variables nominales. Les principes théoriques de cette méthode remontent à Fisher (1940 [32]), à Guttman (1941 [42]) et à Hayashi (1956 [44]). Cependant, c'est Benzécri qui a donné le nom d'*Analyse Factorielle des Correspondances* en 1963 [4], la présentation géométrique et les aides à l'interprétation qu'il a développées en font une des méthodes phares de l'Analyse de Données.
- **ACP**. Lorsque les lignes du tableau de données représentent des individus et les colonnes sont des variables quantitatives, l'*analyse en composantes principales* peut être appliquée. Introduite par Pearson en 1901 [63], elle est aujourd'hui une des méthodes d'Analyse de Données les plus répandues et les plus utilisées. Elle permet de décrire et d'interpréter les proximités qui existent entre les individus et de mettre en évidence les liaisons linéaires qui existent entre les variables.
- **ACM.** L'*Analyse des Correspondances Multiple* est applicable dans le cas où les lignes du tableau de données représentent des individus et les colonnes sont des variables nominales. Elle permet de décrire et d'interpréter les proximités qui existent entre les individus d'une part et entre les modalités des variables nominales considérées d'autre part. L'origine de cette méthode remonte à Guttman (1941 [42]).

1.2 Recherche des axes principaux : représentation géométrique

Comme nous l'avons vu précédemment, le type de tableau de données soumis à une analyse géométrique détermine le type de méthode à utiliser : AC, ACP ou ACM. Dans la suite, chaque ligne de ce tableau est qualifiée d'*unité statistique* ou d'*individu statistique*.

Pour chacune de ces méthodes, une distance (euclidienne) entre les unités statistiques est définie : distance du χ^2 pour l'AC, distance usuelle pour l'ACP, distance basée sur l'inverse des fréquences des catégories pour l'ACM [1]. La distance choisie permet de construire un nuage de N points dans un espace affine euclidien de dimension K. Ces méthodes consistent à déterminer les axes principaux de ce nuage et sont basées sur la propriété mathématique de la décomposition en valeurs singulières d'une application linéaire (ou la diagonalisation de l'endomorphisme symétrique associé). Nous donnons les deux formulations, géométrique et matricielle, de ce procédé (la décomposition en valeurs singulières a été présentée, pour des matrices rectangulaires, par Eckart et Young en 1936 [25] [2]).

Nuage étudié. Soit $I = \{1, \ldots, i, \ldots, N\}$ l'ensemble indexant les unités statistiques (ou individus). Chaque individu i peut être représenté par un point M^i à valeur dans un espace affine euclidien $\mathcal{U}$ de dimension K, soit $\mathcal{V}$ l'espace vectoriel directeur de $\mathcal{U}$ muni du produit scalaire noté $\langle .|. \rangle$ et de la norme euclidienne associée. Notons $\mathcal{M}$ le support affine du nuage et $\mathcal{L}$ le sous–espace vectoriel associé, de dimension L ($L \leq K$). L'importance accordée à chaque point du nuage peut être variable, par conséquent, à chaque point M^i est associée une masse n_i, la somme des masses des points du nuage est notée n [3].

Directions principales. Le but est ici d'approcher le nuage $\mathrm{M}^I = (\mathrm{M}^I)_{i \in I}$, de dimension L, par un nuage de dimension L' (avec $L' < L$). Pour ce faire, on recherche les L' premières directions d'« allongement maximum » du nuage : les L' meilleures directions principales (au sens des moindres carrés). Le nuage projeté sur le sous–espace engendré par ces L' directions, appelé sous–espace principal de dimension L', reconstitue « au mieux » les distances entre les points, il fournit donc le meilleur ajustement du nuage M^I par un nuage de dimension L'. Dans la suite, nous resteignons le nuage M^I à son support $\mathcal{M}$.

Les L directions principales du nuage M^I sont associées aux vecteurs propres de l'endomorphisme de covariance symétrique et positif *Som* associé au nuage M^I (*cf.* chapitre 1 p.16 pour une définition de *Som*).

Théorème 1.1 (Directions principales)**.** *Les vecteurs directeurs des droites principales du nuage* M^I*, notés* $(\overrightarrow{\alpha_\ell})_{\ell=1,\ldots,L}$*, sont les vecteurs propres de l'endomorphisme Som, associés aux valeurs propres* λ_ℓ*.*

$$\forall \ell = 1, \ldots L \ : \ Som(\overrightarrow{\alpha_\ell}) = \lambda_\ell \overrightarrow{\alpha_\ell}$$

Pour une démonstration de ce théorème, se reporter à Benzécri (1984 [6]) ou à Rouanet & Le Roux (1993 [74]).

Soit $L = \{1, \ldots, \ell, \ldots, L\}$ [4]. L'axe passant par G et de vecteur directeur $\overrightarrow{\alpha_\ell}$ est noté (G, $\overrightarrow{\alpha_\ell}$), c'est le ℓ–ème axe principal. Le nuage A^I_ℓ est le nuage projeté du nuage M^I sur

1. Se reporter à Le Roux & Rouanet (2004 [50]) ou à Lebart, Morineau & Piron (1995 [54]) pour le détail des distances utilisées.

2. Généralisation des travaux sur les matrices carrées de Sylvester en 1889 [83].

3. Nous utilisons ici les notations du chapitre 1.

4. L'ensemble et son cardinal sont identifiés par la même lettre, l'ensemble en italique et son cardinal en lettre droite.

cet axe, il est unidimensionnel. La ℓ–ème variable principale calibrée est définie par les coordonnées des points du nuage A_ℓ^I sur l'axe $(\mathrm{G}, \frac{\overrightarrow{\alpha_\ell}}{||\overrightarrow{\alpha_\ell}||})$, sa variance est égale à λ_ℓ.

Propriété 1.1. *Si $\lambda_\ell \neq \lambda_{\ell'}$, les ℓ–ème et ℓ'–ème variables principales sont non–corrélées.*

Décomposition principale. Le nuage M^I peut être décomposé en L nuages principaux unidimensionnels orthogonaux grâce à la formule suivante :

$$\forall i \in I \ : \ \overrightarrow{\mathrm{GM}}^i = \sum_{\ell=1}^{L} \overrightarrow{\mathrm{GA}}_\ell^i$$

D'après la propriété 1.1, on a :

$$\overrightarrow{\mathrm{GA}}_\ell^i \perp \overrightarrow{\mathrm{GA}}_{\ell'}^i, \text{ pour } \ell \neq \ell'$$

On en déduit que la *variance du nuage* est égale à la somme des variances des L nuages principaux $(\mathrm{A}_\ell^I)_{\ell=1,\ldots L}$, donc à la somme des L valeurs propres.

Remarque : La démarche reste la même dans le cas de valeurs propres multiples ; il n'y a plus alors unicité des axes principaux correspondants mais un choix d'axes principaux orthogonaux deux à deux peut toujours être fait.

Si les valeurs propres de *Som* : $\lambda_1 \ldots, \lambda_l \ldots, \lambda_L$, sont ordonnées de façon décroissante, le sous–espace principal de dimension L' cherché est défini par la propriété d'hérédité suivante :

Propriété 1.2 (Propriété d'hérédité)**.** *Le sous–espace principal de dimension L' est engendré par les L' premiers axes principaux :* $(\mathrm{G}, \overrightarrow{\alpha_1})$, …, $(\mathrm{G}, \overrightarrow{\alpha_\ell})$, …, $(\mathrm{G}, \overrightarrow{\alpha_{L'}})$.

Ajustement du nuage M^I *:*
Le nuage projeté du nuage M^I dans le sous–espace principal de dimension L' est noté $\widetilde{\mathrm{M}}^I$, il fournit le meilleur ajustement par un nuage de dimension L' du nuage M^I (au sens des moindres carrés) :

$$\forall i \in I \ : \ \overrightarrow{\widetilde{\mathrm{GM}}^i} = \sum_{\ell=1}^{L'} \overrightarrow{\mathrm{GA}}_\ell^i$$

On dit aussi que l'on effectue la *reconstitution d'ordre L'* du nuage M^I.

Formulations matricielles simples. Comme nous l'avons vu, la procédure centrale des méthodes d'analyse géométrique est la diagonalisation de l'endomorphisme symétrique *Som*. Or la matrice d'un endomorphisme symétrique, dans une base orthonormée, est symétrique. Par conséquent, si l'on munit l'espace vectoriel $\mathcal{L}$ d'un repère orthonormé $(\mathrm{G}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$, on peut donner des formulations matricielles de la recherche des axes principaux en dimension finie.

Notons $\mathbf{X} = [x^{i\ell}]_{i \in I, \ell \in L}$ la matrice à N lignes et L colonnes contenant les coordonnées des points du nuage M^I dans le repère orthonormé $(\mathrm{G}, (\overrightarrow{\varepsilon_\ell})_{\ell \in L})$. De même, notons $\mathbf{F}$ la matrice diagonale d'ordre N dont les éléments diagonaux sont les fréquences $(n_i/n)_{i \in I}$. On dit que les colonnes de $\mathbf{X}$ représentent les *variables coordonnées*.

Remarque : G étant l'origine du repère, les variables coordonnées ont pour moyenne 0.

Soit $\mathbf{V}$ la matrice de covariance entre les L variables coordonnées : $\mathbf{V} = \mathbf{X}'\mathbf{F}\mathbf{X}$; $\mathbf{V}$ est également la matrice associée à l'endomorphisme *Som* dans le repère orthonormé $(\mathrm{G}, (\overrightarrow{\varepsilon_\ell})_{\ell\in L})$, elle est symétrique et définie positive. Chercher les vecteurs propres de l'endomorphisme *Som* revient donc à chercher les vecteurs propres $(\mathbf{a}_\ell)_{\ell\in L}$ de la matrice $\mathbf{V}$, c'est–à–dire définis par :

$$\mathbf{V}\mathbf{a}_\ell = \lambda_\ell \mathbf{a}_\ell$$

avec $(\lambda_\ell)_{\ell\in L}$, les valeurs propres associées, ordonnées de façon décroissante. Pour faciliter l'exposé, nous prenons ici des vecteurs normés :

$$\forall \ell \in L,\ \mathbf{a}'_\ell \mathbf{a}_\ell = 1$$

Remarque : Dans le repère orthonormé $(\mathrm{G}, (\overrightarrow{\varepsilon_\ell})_{\ell\in L})$, les vecteurs–colonnes $(\mathbf{a}_\ell)_{\ell\in L}$ représentent les coordonnées des vecteurs $(\overrightarrow{\alpha}_\ell)_{\ell\in L}$ normés définis précédemment.

Le vecteur colonne $\mathbf{y}_\ell$ représentant la ℓ–ème variable principale calibrée est tel que :

$$\mathbf{y}_\ell = \mathbf{X}\mathbf{a}_\ell$$

Approximation de la matrice $\mathbf{X}$ *:*
Toute matrice $\mathbf{X}$ de rang L peut être écrite comme une somme de produits de matrices de rang 1. Cependant, la décomposition en valeurs singulières suivante est unique :

$$\mathbf{X} = \sum_{\ell\in L} \mathbf{X}\mathbf{a}_\ell\ \mathbf{a}'_\ell = \sum_{\ell\in L} \mathbf{y}_\ell\ \mathbf{a}'_\ell$$

Eckart et Young (1936 [25]) montrent que la meilleure approximation d'une matrice de rang L par une matrice de rang $L'(< L)$ est obtenue en se restreignant aux L' premiers termes de la somme :

$$\widetilde{\mathbf{X}} = \sum_{\ell\in L'} \mathbf{X}\mathbf{a}_\ell\ \mathbf{a}'_\ell = \sum_{\ell\in L'} \mathbf{y}_\ell\ \mathbf{a}'_\ell \quad (= \sum_{\ell\in L'} \sqrt{\lambda_\ell}\frac{\mathbf{y}_\ell}{\sqrt{\lambda_\ell}}\ \mathbf{a}'_\ell)$$

On dit aussi que l'on effectue la *reconstitution d'ordre* L' de la matrice $\mathbf{X}$.

2 Analyses sur tableaux à trois entrées

La décomposition en valeurs singulières, sous–jacente à l'ensemble des méthodes d'analyse géométrique des données, ne se généralise pas directement au cas des tableaux à plus de deux entrées : les *tableaux multiples*. Pour pallier ce problème, plusieurs solutions ont été proposées, elles s'inscrivent soit dans le cadre de la modélisation multi–tableaux (comme par exemple dans Kroonenberg, 1989 [46]), soit dans le cadre géométrique (*cf.* Cazes, 2004 [13]). Nous donnons ici un bref aperçu des solutions proposées.

L'exposé est basé, sans perte de généralité, sur les tableaux à trois dimensions, le cas multiple en est une simple extension.

On ne dispose plus d'un tableau (ou matrice) $\mathbf{X}$ à N lignes et L colonnes mais d'un ensemble de T tableaux (matrices) à N lignes et L colonnes : $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_T$. L'ensemble des T matrices forme un *tenseur d'ordre 3*, de taille $N \times L \times T$, noté $\mathcal{X}$ (figure B.1 ci–après). Notons $T = \{1, \ldots, t, \ldots, T\}$ l'ensemble indexant les T matrices.

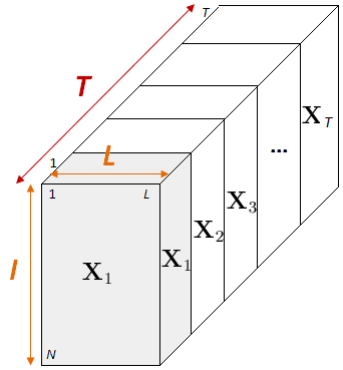


FIGURE B.1 – Tenseur d'ordre 3.

2.1 Approche « Modèles »

De nombreux modèles de décomposition tensorielle ont été développés. L'idée même est due à Cattell (1944 [11]), mais ce n'est qu'après les travaux de Tucker (1966 [87]), de Carroll & Chang (1970 [10]) et de Harshman (1970 [43]), initialement réalisés pour des applications en psychométrie, que grandit l'intérêt pour ce concept.

Dans les paragraphes suivants, nous présentons un des principaux modèles de décomposition : le modèle PARAFAC. Le modèle de Tucker est aussi introduit brièvement.

2.1.1 Décomposition PARAFAC

Nous nous intéressons ici à une décomposition des tenseurs connue sous le nom de Parallel Factor Analysis (PARAFAC). Elle a été introduite sous ce nom par Harshman en 1970 [43] et développée indépendamment, sous le nom de Canonical Decomposition (CANDECOMP), par Carroll et Chang la même année [10]. C'est pourquoi, on trouve souvent l'appellation CP décomposition (CANDECOMP/PARAFAC décomposition).

PARAFAC propose de décomposer un tenseur sous la forme d'une somme minimale de tenseurs de rang un. Le rang d'un tenseur peut être vu comme une généralisation du rang des matrices, sa définition est donnée dans la suite.

Définitions préliminaires :

Définition 2.1 (Produit–externe)**.** *Le produit–externe (noté $\circ$) de k vecteurs $\mathbf{u}^{(1)}$, $\mathbf{u}^{(2)}$, ..., $\mathbf{u}^{(k)}$, de tailles respectives I_1, I_2, ..., I_k est un tenseur d'ordre k de taille $I_1 \times I_2 \times \ldots \times I_k$ dont l'élément d'indice $(i_1, i_2, \ldots, i_k)$ est défini par :*

$$(\mathbf{u}^{(1)} \circ \mathbf{u}^{(2)} \circ \ldots \circ \mathbf{u}^{(k)})_{i_1,i_2,\ldots,i_k} = \mathbf{u}^{(1)}_{i_1}\mathbf{u}^{(2)}_{i_2}\ldots\mathbf{u}^{(k)}_{i_k}.$$

Définition 2.2 (Tenseur de rang un)**.** *Un tenseur qui se décompose sous la forme d'un produit externe de vecteurs est dit de rang un.*

Définition 2.3 (Tenseur de rang R)**.** *Un tenseur qui se décompose sous la forme d'une somme (minimale) de R tenseurs de rang un est un tenseur de rang R.*

La décomposition PARAFAC d'un tenseur de rang R est la décomposition de ce tenseur sous la forme de la somme de R tenseurs de rang un. Dans le cas d'un tenseur $\mathcal{X}$ de rang 3 et de taille $N \times L \times T$, cette décomposition s'écrit :

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

avec R un entier positif et $\mathbf{a}_r$, $\mathbf{b}_r$ et $\mathbf{c}_r$ des vecteurs de tailles respectives N, L et T. On peut également donner l'écriture suivante :

$$x_{i\ell t} = \sum_{r=1}^{R} a_{ir} b_{\ell r} c_{tr}, \text{ pour } i = 1, \ldots, N, \ \ell = 1, \ldots, L \text{ et } t = 1, \ldots, T.$$

avec $x_{i\ell t}$ l'élément d'indice (i, ℓ, t) du tenseur $\mathcal{X}$, a_{ir} l'élément d'indice i du vecteur $\mathbf{a}_r$, $b_{\ell r}$ l'élément d'indice ℓ du vecteur $\mathbf{b}_r$ et c_{tr} l'élément d'indice t du vecteur $\mathbf{c}_r$.

Le schéma de la décomposition PARAFAC d'un tenseur de rang 3 est donné sur la figure suivante :

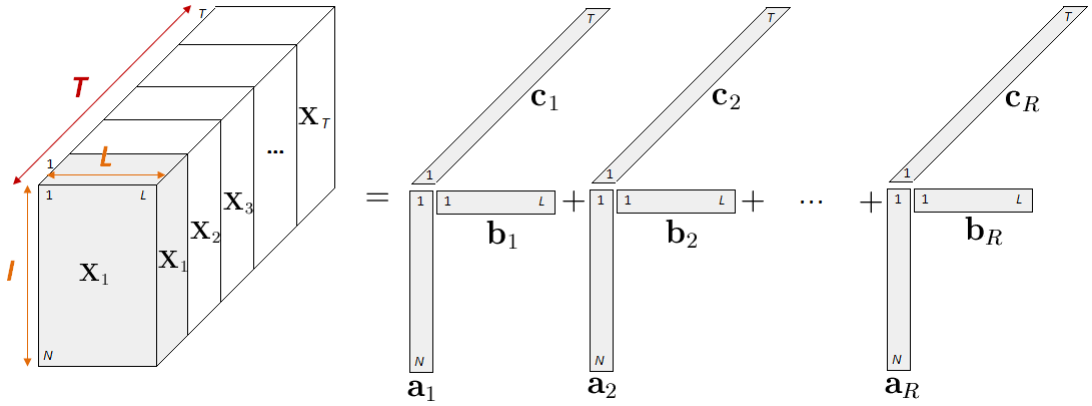


FIGURE B.2 – Décomposition PARAFAC d'un tenseur de rang 3.

Notons $\mathbf{A}$, la matrice formée des R vecteurs–colonnes $\mathbf{a}_1, \ldots, \mathbf{a}_R$: $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_R)$. De même, notons $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_R)$ et $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_R)$.

Unicité :

De façon générale, sous des conditions peu restrictives, ce modèle a une solution unique, ce qui a fait son succès. Le résultat d'unicité de la décomposition est dû à Harshman (1970 [43]). Cependant, le théorème d'unicité le plus général et le plus connu est celui de Kruskal (1977 [47]), il est énoncé ci–après, tel que l'a fait Kruskal, pour un tenseur d'ordre 3, mais a été démontré plus tard par Sidiropoulos et Bro pour un tenseur d'ordre quelconque (2000 [82]).

Définition 2.4 (k–rang de Kruskal)**.** *Le k–rang de Kruskal de la matrice $\mathbf{A}$, noté $k_{\mathbf{A}}$ est le nombre maximum $k_{\mathbf{A}}$ de colonnes de $\mathbf{A}$ tel que toute sous–matrice de $\mathbf{A}$ de $k_{\mathbf{A}}$ colonnes soit de rang plein.*

Théorème 2.1 (Théorème d'unicité de Kruskal)**.** *Si : $k_{\mathbf{A}} + k_{\mathbf{B}} + k_{\mathbf{C}} \leq 2(R+1)$, alors la décomposition PARAFAC est unique.*

Calcul de la décomposition PARAFAC :

Dans la pratique, il n'existe pas d'algorithme permettant de calculer le rang R d'un tenseur. On doit donc fixer R *a priori* (Bro & Kiers ont proposé en 2003 une procédure, appelée CONCORDIA, pour comparer différentes valeurs de R [9]). En supposant que R est fixé, il existe de nombreux algorithmes pour effectuer une décomposition PARAFAC mais le plus utilisé est l'algorithme ALS (Alternating Least Squares) proposé dans les articles fondateurs de Harshman (1970 [43]) et de Carroll et Chang (1970 [10]). On doit approcher le tenseur $\mathcal{X}$ par un tenseur $\widetilde{\mathcal{X}}$ de rang R (c'est le tenseur $\widetilde{\mathcal{X}}$ que l'on décompose). Le modèle s'écrit de la façon suivante :

$$\mathcal{X} = \widetilde{\mathcal{X}} + \mathcal{E} = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \mathcal{E}$$

avec $\mathcal{E}(= \mathcal{X} - \widetilde{\mathcal{X}})$ un tenseur « écart » d'ordre 3. Ou, de façon équivalente :

$$x_{i\ell t} = \widetilde{x_{i\ell t}} + e_{i\ell t} = \sum_{r=1}^{R} a_{ir} b_{\ell r} c_{tr} + e_{i\ell t}, \text{ pour } i = 1, \ldots, N, \ \ell = 1, \ldots, L \text{ et } t = 1, \ldots, T.$$

avec $e_{i\ell t}$ l'élément d'indice (i, ℓ, t) du tenseur $\mathcal{E}$.

Le principe de l'algorithme ALS est de minimiser $\mathcal{E}$ au sens des moindres carrés. Pour ce faire, il met à jour, de manière alternée, chacune des matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, en gardant les deux autres fixées (on initialise deux des trois matrices aléatoirement, la troisième pouvant en être déduite). Lorsque deux matrices parmi les trois sont fixées, le système à résoudre devient un simple problème de moindres carrés.

2.1.2 Décomposition de TUCKER

La décomposition présentée ici a été introduite par Tucker en 1963 [86], elle est souvent appelée *décomposition Tucker 3*, son principe est de décomposer un tenseur en un tenseur « coeur » multiplié par une matrice le long de chacun de ses modes. Dans le cas ternaire, on a :

$$\mathcal{X} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \ \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r + \mathcal{E}$$

où $\mathbf{a}_p$, $\mathbf{b}_q$ et $\mathbf{c}_r$ sont les composantes (colonnes) des matrices $\mathbf{A} \in \mathbb{R}^{N \times P}$, $\mathbf{B} \in \mathbb{R}^{L \times Q}$ et $\mathbf{C} \in \mathbb{R}^{T \times R}$, matrices étant souvent prises orthogonales et pouvant être considérées comme composantes principales sur chaque mode. Le tenseur $\mathcal{G}$ est appelé le tenseur « coeur ».

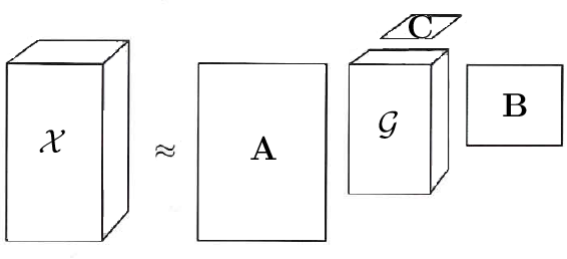


FIGURE B.3 – Décomposition Tucker 3 d'un tenseur de rang 3.

La plupart des algorithmes supposent que les matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$ sont orthogonales, mais ce n'est pas obligatoire. L'algorithme de base pour calculer une solution à Tucker 3 consiste :

- pour le calcul des matrices $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$, à juxtaposer les trois modes du tenseur initial ; on obtient ainsi des tableaux à 2 entrées sur lesquels on peut effectuer des ACP et obtenir les matrices orthogonales $\mathbf{A}$, $\mathbf{B}$ et $\mathbf{C}$;
- pour le calcul de $\mathcal{G}$, on utilise la formule :

$$g_{pqr} = \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{t=1}^{T} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r x_{ilt}$$

D'autres modèles de Tucker, notamment Tucker 2, peuvent aussi être utilisés. De nombreux travaux (comme par exemple Carroll & Chang, 1970 [10]) montrent les relations qui existent, sous diverses contraintes, entre les modèles PARAFAC et les modèles Tucker. Par exemple, PARAFAC peut être vu comme un cas particulier de Tucker 3 lorsque le tenseur coeur est superdiagonal et que l'on a $P = Q = R$.

2.2 Autres approches

Nous citons ici deux autres approches :

Juxtaposition/Superposition de tableaux à deux entrées. C'est la procédure la plus classique : il s'agit d'effectuer l'analyse d'un des T tableaux (tableau de référence) et de placer les autres en supplémentaires. Dans cette approche, les tableaux peuvent être superposés (verticalement) et/ou juxtaposés (horizontalement).

Des procédures plus complexes, traitant notamment des cas où l'on dispose d'une suite de tableaux définis seulement sur le même ensemble d'individus ou seulement sur le même ensemble de variables, sont détaillées dans Cazes (2004 [13]). Citons par exemple les approches STATIS (Hermier des Plantes, 1976 [55]) et AFM (Escofier & Pagès, 1998 [29]) dans lesquelles il s'agit d'associer à chaque tableau une pondération adéquate.

Approche fonctionnelle. Lorsqu'on dispose d'un tableau à deux dimensions où chaque variable est une donnée « complexe », on entre dans le cadre de l'analyse fonctionnelle. Pour plus de détails, se reporter aux travaux de Besse & Ramsay (1986 [8]) et de Ramsay & Silverman (2005 [68]) relatifs à l'ACP fonctionnelle.

Annexe C

Mise en oeuvre informatique

Pour les trois tests développés dans cette thèse, trois fonctions R ont été écrites : test_geo_typicalite.R (mise en oeuvre du test géométrique de typicalité), test_ens_typicalite.R (mise en oeuvre du test ensembliste de typicalité) et test_homogeneite.R (mise en oeuvre du test d'homogénéité). Ces programmes donnent pour chaque test les seuils et zones de compatibilité associées [1].

R étant un langage interprété et non un langage compilé, les temps de calculs peuvent être conséquents lorsque le volume de données est important. De plus, nous n'avons pas cherché réellement à optimiser les procédures de calculs à cette étape de la recherche.

Dans ce chapitre annexe, nous donnons d'abord une notice d'utilisation pour chaque test, puis nous détaillons les étapes et astuces de développement pour le test géométrique (celles des deux autres tests s'inscrivant dans la même ligne).

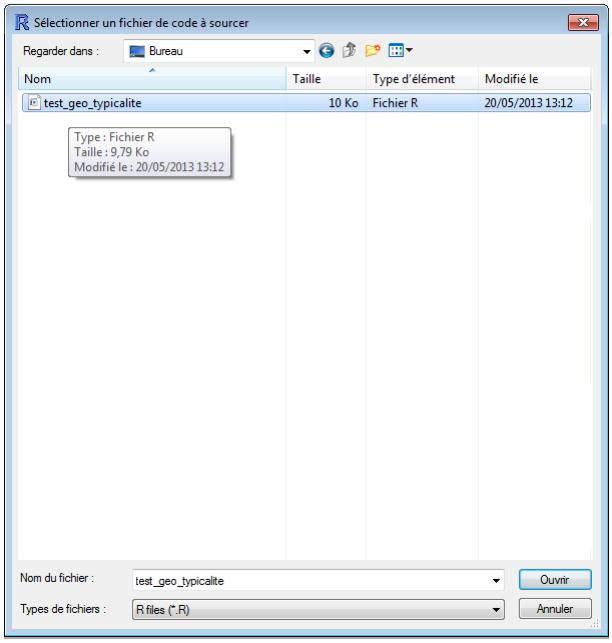
1 Notices d'utilisation

Le logiciel R doit être préalablement installé.

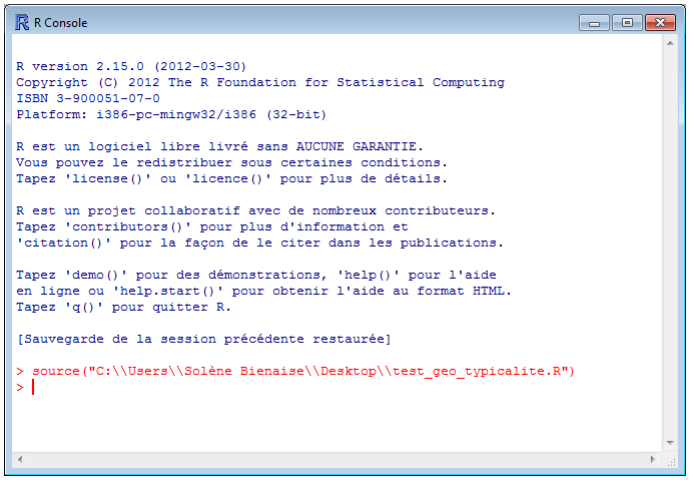
1.1 Test géométrique de typicalité

– Étape 0 ⇒ Charger la fonction test_geo_typicalite dans R (elle doit être rechargée à chaque démarrage) : cliquer sur « Fichier » puis « Sourcer du code R... ». La fenêtre suivante apparaît :

1. Ces programmes sont encore en phase de test ; toutefois ils sont disponibles sur demande.

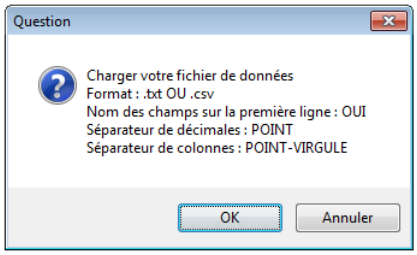


sélectionner le fichier test_geo_typicalite.R et cliquer sur « Ouvrir ». La console R affiche :



Exécution de la fonction test_geo_typicalite :

– Étape 1 ⇒ Dans la console R, taper :
  `test_geo_typicalite()`
– Étape 2 ⇒ La fenêtre suivante apparaît, elle indique les formats possibles du fichier de données (.txt ou .csv) :



Pour l'exemple cible (traité au chapitre 2, p.26), le fichier de données (.txt ou .csv) a la forme suivante :

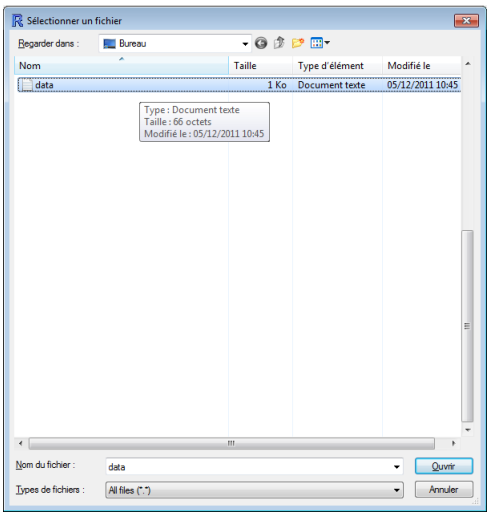
```
Var1;Var2
0;-12
```

```
6;-10
14;-6
6;-2
12;0
-8;2
2;4
6;4
10;10
12;10
```

où « Var1 » et « Var2 » sont les variables–coordonnées des points du nuage soumis au test.

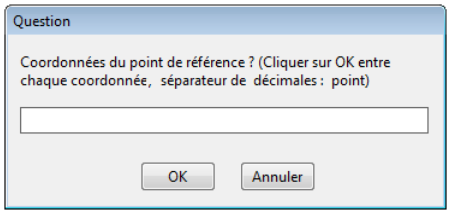
Remarques :
- Ce programme a été écrit dans l'optique de pouvoir être utilisé par des non utilisateurs de R, nous avons donc fait le choix de « cacher » l'import des données (read.table) dans la fonction test_geo_typicalite().
- L'exemple précédent comporte deux variables, mais il peut bien sûr y en avoir plus.
- Si l'utilisateur clique sur « Annuler », le programme stoppe, il peut alors transformer son fichier de données dans un format adapté puis revenir à l'Étape 1.

– Étape 3 ⇒ La fenêtre suivante apparaît :



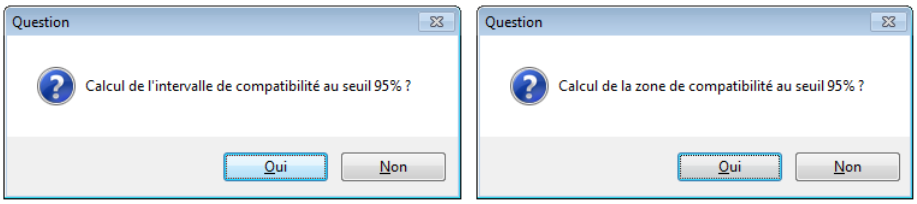
Indiquer la localisation du fichier de données, puis cliquer sur « Ouvrir ».

– Étape 4 ⇒ La fenêtre suivante apparaît :



Saisir les coordonnées du point de référence une par une : entrer la première, cliquer sur « OK », puis entrer la seconde, cliquer sur « OK », *etc.* Le séparateur de décimales doit être un point.

– Étape 5 ⇒ Une des fenêtres suivantes apparaît (celle de gauche si les données sont unidimensionnelles, celle de droite sinon) :



Cliquer « Oui » pour obtenir l'intervalle de compatibilité (cas unidimensionnel), ou le kappa de l'ellipse de compatibilité (cas multidimensionnel), cliquer « Non » pour obtenir uniquement le résultat du test.

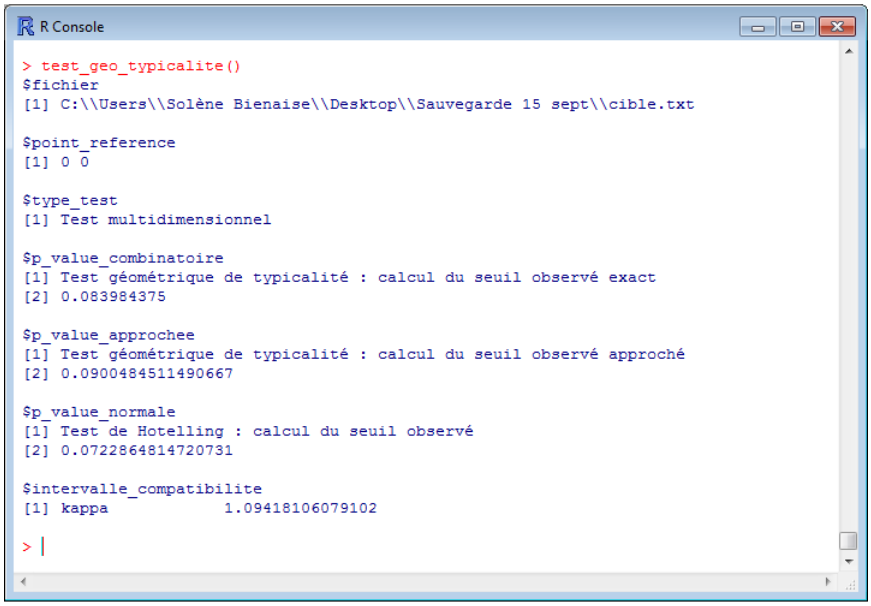
– Étape 6 ⇒ Le programme s'exécute et les résultats sont affichés :

Remarque : Le test exact est effectué pour un petit nombre d'observations ($N \leq 15$). Pour un plus grand nombre d'observations ($15 < N \leq 100$), nous donnons une estimation du seuil observé du test à l'aide de la méthode de Monte Carlo. Lorsque $N > 100$, nous donnons le seuil observé approché. Les seuils approché et de Student (ou de Hotelling dans le cas multidimensionnel) sont également fournis quelque soit la taille des données.
De la même façon, lorsqu'il est demandé par l'utilisateur, l'intervalle de compatibilité (ou le κ de l'ellipse délimitant la zone de compatibilité dans le cas multidimensionnel) est donné de manière exacte pour $N \leq 15$ et en utilisant la méthode de Monte Carlo pour $15 < N \leq 100$. Pour $N > 100$, seul l'intervalle (ou la zone) approché(e) est donné(e). Le tableau suivant résume ces éléments :

Nombre d'observations	Seuil exact	Seuil approché	Seuil de Student/Hotelling	Zone exacte	Zone approchée
$0 < N \leq 15$	X	X	X	X	
$15 < N \leq 100$	X (Monte Carlo)	X	X	X (Monte Carlo)	
$N > 100$		X	X		X

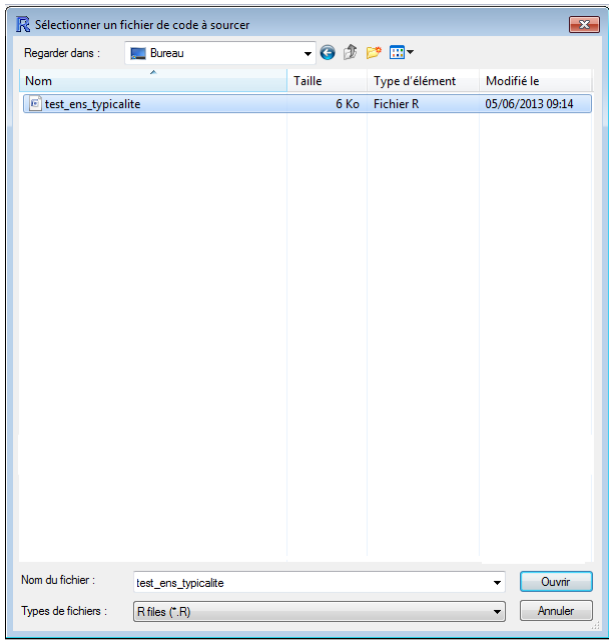
Exemple cible (multidimensionnel) :

– Fichier cible.txt (voir données ci–avant)
– Point de référence de coordonnées :
0 0
– Zone de compatibilité au seuil 95% : demandée

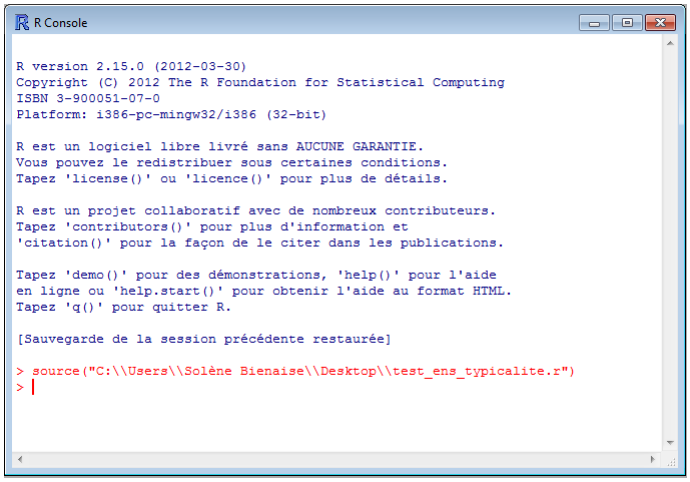


1.2 Test ensembliste de typicalité

– Étape 0 ⇒ Charger la fonction test_ens_typicalite dans R (elle doit être rechargée à chaque démarrage) : cliquer sur « Fichier » puis « Sourcer du code R... ». La fenêtre suivante apparaît :

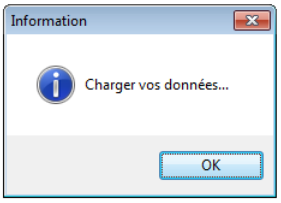


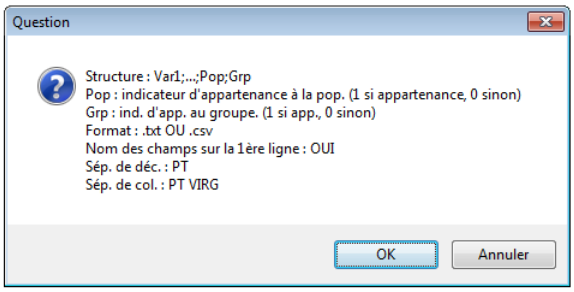
sélectionner le fichier test_ens_typicalite.R et cliquer sur « Ouvrir ». La console R affiche :



Exécution de la fonction test_ens_typicalite :

– Étape 1 ⇒ Dans la console R, taper :
```
test_ens_typicalite()
```
– Étape 2 ⇒ Les fenêtres suivantes apparaissent. Elles indiquent les formats possibles (.txt ou .csv) et la structure du fichier de données nécessaires au bon fonctionnement de la fonction :





Pour l'exemple des races canines (traité au chapitre 3, p.73), le fichier de données (.txt ou .csv) a la forme suivante :

```
Var1;Var2;Pop;Grp
0.3172000;-0.4177020;1;0
-0.2541100;1.1012300;1;0
0.4863950;-0.4644500;1;0
-0.4473650;-0.8817780;1;1
-1.0133500;0.5498800;1;1
0.7525750;0.5469120;1;0
-0.9123020;-0.0161873;1;1
-0.8407990;0.8438520;1;1
-0.7332950;0.0790735;1;1
0.1173250;-0.5261080;1;1
-0.6472400;-0.9901840;1;1
0.8732100;-0.3154810;1;0
1.0470200;0.5069570;1;0
-0.4780450;-1.0369300;1;0
0.1449100;-0.5157830;1;0
0.8765680;0.0252396;1;0
-0.8816220;0.1389670;1;1
0.5173380;-0.1134040;1;0
-0.6472400;-0.9901840;1;0
0.6766930;-0.0831668;1;0
0.7559320;0.8876330;1;0
-0.8407990;0.8438520;1;1
0.6733350;-0.4238880;1;0
0.5833790;0.5936600;1;0
0.5041400;-0.3771390;1;0
-1.0133500;0.5498800;1;1
0.3835050;0.4852540;1;0
```

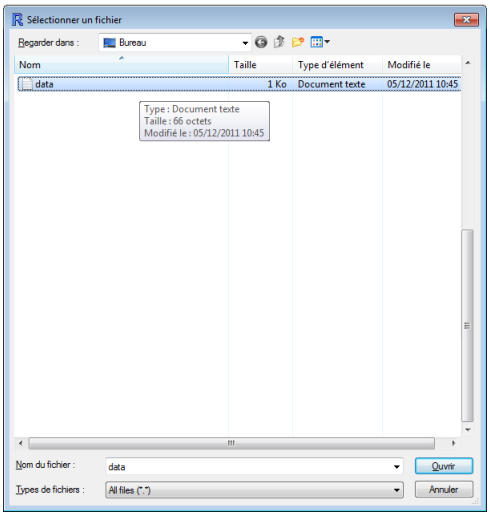
où « Var1 » et « Var2 » sont les variables–coordonnées des points du nuage soumis au test et où « Pop » et « Grp » sont respectivement des indicateurs d'appartenance des individus à la population de référence et au groupe d'observations (0 : appartenance, 1 : non appartenance) [2].

Remarques :

– L'exemple précédent comporte deux variables, mais il peut bien sûr y en avoir plus.
– Si l'utilisateur clique sur « Annuler », le programme stoppe, il peut alors transformer son fichier de données dans un format adapté puis revenir à l'Étape 1.

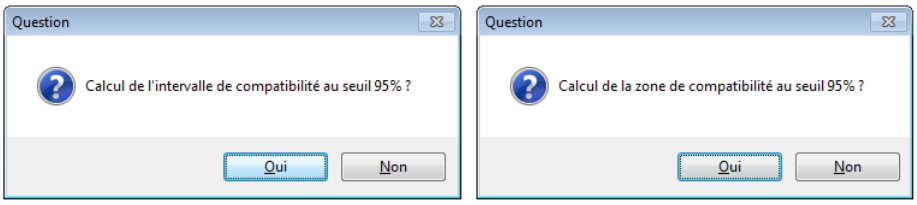
2. Le nom des colonnes n'a aucune importance, seules comptent leurs positions dans le fichier de données : « Pop » doit être l'avant dernière colonne et « Grp » la dernière.

– Étape 3 ⇒ La fenêtre suivante apparaît :



Indiquer la localisation du fichier de données, puis cliquer sur « Ouvrir ».

– Étape 4 ⇒ Une des fenêtres suivantes apparaît (celle de gauche si les données sont unidimensionnelles, celle de droite sinon) :



Cliquer « Oui » pour obtenir l'intervalle de compatibilité (cas unidimensionnel), ou le kappa de l'ellipse de compatibilité (cas multidimensionnel), cliquer « Non » pour obtenir uniquement le résultat du test.

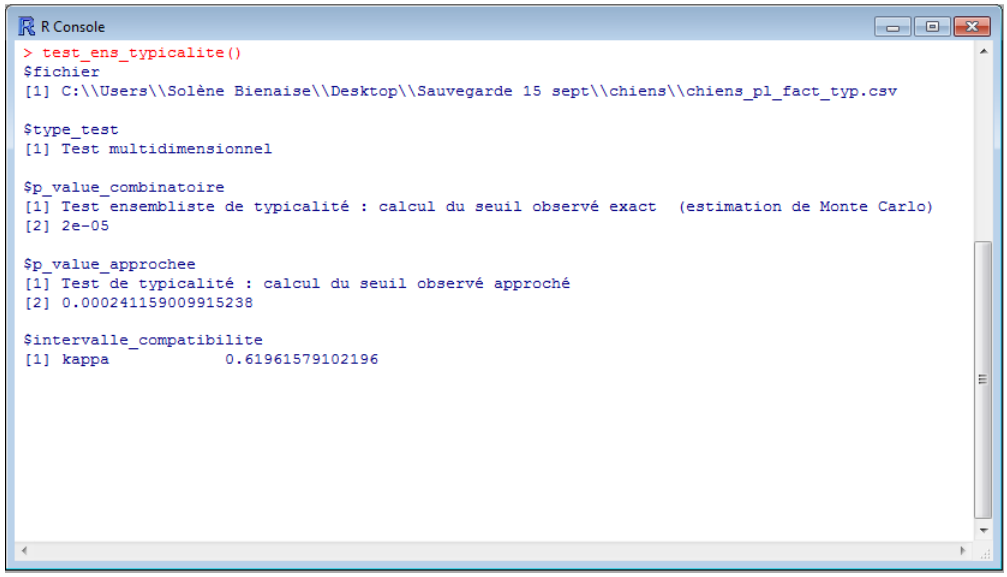
– Étape 5 ⇒ Le programme s'exécute et les résultats sont affichés :

Remarque : Le test exact est effectué lorsque le nombre d'échantillons (noté N_{ech}) à construire est raisonnable ($N_{ech} < 30000$)(*cf.* chapitre 3, paragraphe 1.1.1, p.54). Pour un plus grand nombre d'échantillons ($N_{ech} \geq 30000$) et lorsque le nombre d'individus n appartenant au groupe d'observations est inférieur ou égal à 100, nous donnons une estimation du seuil observé du test à l'aide de la méthode de Monte Carlo. Lorsque $N_{ech} \geq 30000$ et $n > 100$, nous ne donnons que le seuil observé approché (fourni quelque soit la taille des données).
De la même façon, lorsqu'il est demandé par l'utilisateur, l'intervalle de compatibilité (ou le κ de l'ellipse délimitant la zone de compatibilité dans le cas multidimensionnel) est donné de manière exacte pour $N_{ech} < 30000$ et en utilisant la méthode de Monte Carlo pour $N_{ech} \geq 30000$ et $n \leq 100$. Pour $N_{ech} \geq 30000$ et $n > 100$, seul l'intervalle (ou la zone) approché(e) est donné(e). Le tableau suivant résume ces éléments :

Nombre d'observations	Seuil exact	Seuil approché	Zone exacte	Zone approchée
$N_{ech} < 30000$	X	X	X	
$N_{ech} \geq 30000$ et $n \leq 100$	X (Monte Carlo)	X	X (Monte Carlo)	
$N_{ech} \geq 30000$ et $n > 100$		X		X

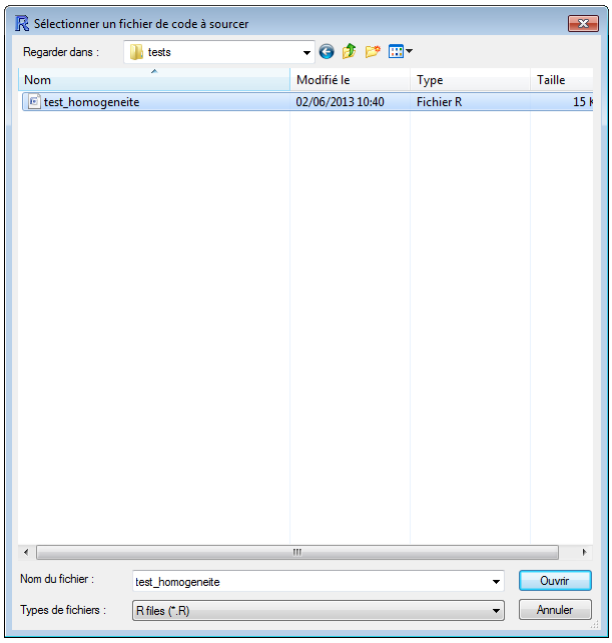
Exemple des races canines (typicalité du groupe 1 par rapport à la population de référence constituée de la réunion des trois groupes) :

- Fichier races_canines.txt (voir données ci–avant)
- Zone de compatibilité au seuil 95% : demandée

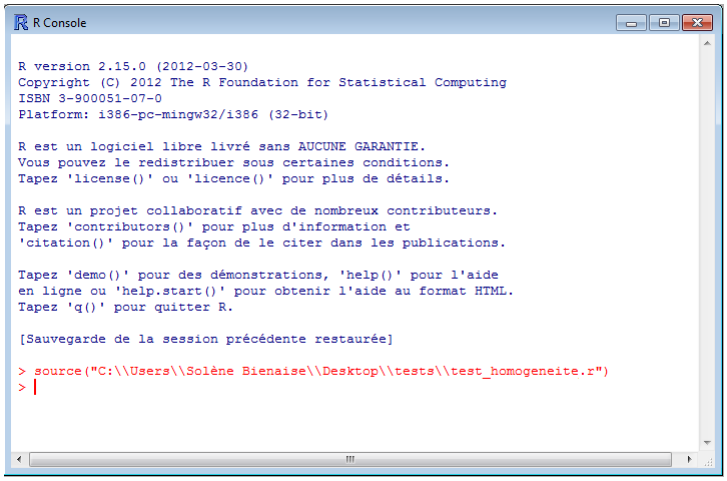


1.3 Test d'homogénéité

- Étape 0 ⇒ Charger la fonction test_homogeneite dans R (elle doit être rechargée à chaque démarrage) : cliquer sur « Fichier » puis « Sourcer du code R... ». La fenêtre suivante apparaît :



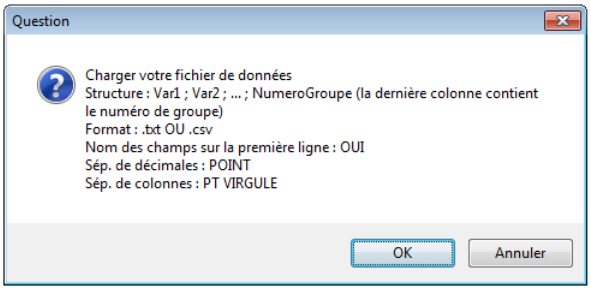
sélectionner le fichier test_homogeneite.R et cliquer sur « Ouvrir ». La console R affiche :



Exécution de la fonction test_homogeneite :

– Étape 1 ⇒ Dans la console R, taper :
`test_homogeneite()`

– Étape 2 ⇒ La fenêtre suivante apparaît, elle indique les formats possibles du fichier de données (.txt ou .csv) :



Pour l'exemple des races canines (traité au chapitre 4, p.114), le fichier de données (.txt ou .csv) a la forme suivante :

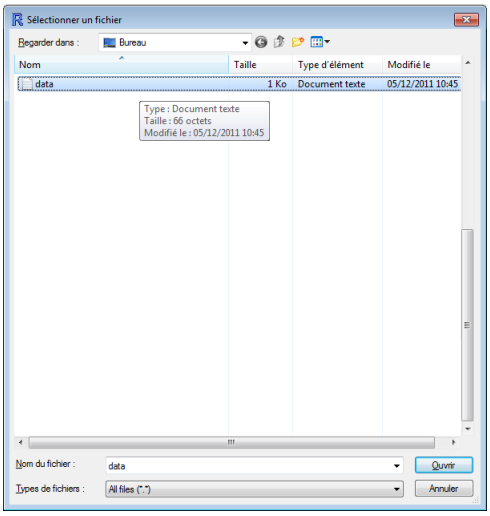
```
Var1;Var2;Groupe
0.3172000;-0.4177020;3
-0.2541100;1.1012300;2
0.4863950;-0.4644500;3
-0.4473650;-0.8817780;1
-1.0133500;0.5498800;1
0.7525750;0.5469120;3
-0.9123020;-0.0161873;1
-0.8407990;0.8438520;1
-0.7332950;0.0790735;1
0.1173250;-0.5261080;1
-0.6472400;-0.9901840;1
0.8732100;-0.3154810;3
1.0470200;0.5069570;3
-0.4780450;-1.0369300;2
0.1449100;-0.5157830;2
0.8765680;0.0252396;2
-0.8816220;0.1389670;1
0.5173380;-0.1134040;2
-0.6472400;-0.9901840;2
```

```
0.6766930;-0.0831668;2
0.7559320;0.8876330;3
-0.8407990;0.8438520;1
0.6733350;-0.4238880;2
0.5833790;0.5936600;3
0.5041400;-0.3771390;2
-1.0133500;0.5498800;1
0.3835050;0.4852540;3
```

où « Var1 » et « Var2 » sont les variables–coordonnées des points du nuage soumis au test et « Groupe » est le numéro des groupes [3].

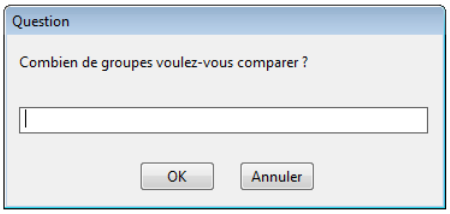
Remarques :
- L'exemple précédent comporte deux variables et trois groupes, mais il peut bien sûr y en avoir plus.
- Si l'utilisateur clique sur « Annuler », le programme stoppe, il peut alors transformer son fichier de données dans un format adapté puis revenir à l'Étape 1.

– Étape 3 ⇒ La fenêtre suivante apparaît :



Indiquer la localisation du fichier de données, puis cliquer sur « Ouvrir ».

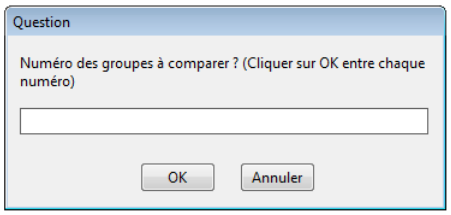
– Étape 4 ⇒ La fenêtre suivante apparaît :



Entrer le nombre de groupes à comparer puis cliquer sur « OK ».

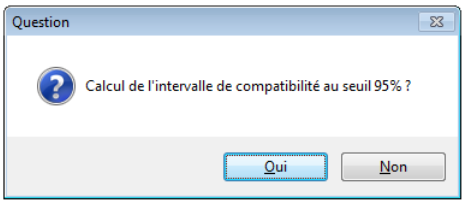
– Étape 5 ⇒ La fenêtre suivante apparaît :

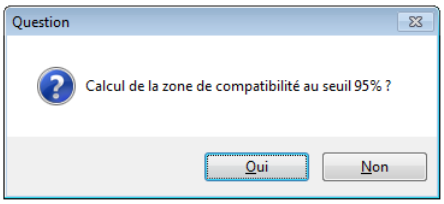
3. Le nom des colonnes n'a aucune importance, seules comptent leurs positions dans le fichier de données : « Groupe » doit être la dernière colonne.



Entrer les numéros des groupes à comparer un par un : entrer le premier, cliquer sur « OK », puis entrer le second, cliquer sur « OK », *etc.*

– Étape 6 ⇒ Si le nombre de groupes à comparer vaut 2, une des fenêtres suivantes apparaît (celle de gauche si les données sont unidimensionnelles, celle de droite sinon) :





Cliquer « Oui » pour obtenir l'intervalle de compatibilité (cas unidimensionnel), ou le kappa de l'ellipse de compatibilité (cas multidimensionnel), cliquer « Non » pour obtenir uniquement le résultat du test.

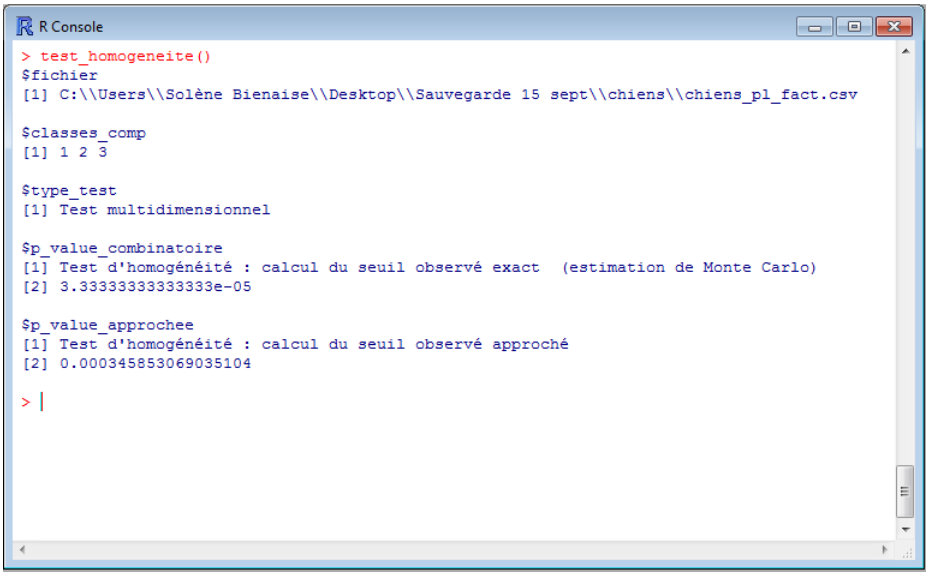
– Étape 7 ⇒ Le programme s'exécute et les résultats sont affichés :

Remarque : Le test exact est effectué lorsque le nombre d'emboîtements (noté N_{emb}) à construire est raisonnable ($N_{emb} < 30000$)(*cf.* chapitre 4, paragraphe 1.1.1, p.85). Pour un plus grand nombre d'emboîtements ($30000 \leq N_{emb} \leq 10^{15}$), nous donnons une estimation du seuil observé du test à l'aide de la méthode de Monte Carlo. Lorsque $N_{emb} > 10^{15}$, nous ne donnons que le seuil observé approché (fourni quelque soit la taille des données).
De la même façon, lorsqu'il est demandé par l'utilisateur, l'intervalle de compatibilité (ou le κ de l'ellipse délimitant la zone de compatibilité dans le cas multidimensionnel) est donné de manière exacte pour $N_{emb} < 30000$ et en utilisant la méthode de Monte Carlo pour $30000 \leq N_{emb} \leq 10^{15}$. Pour $N_{emb} > 10^{15}$, seul l'intervalle (ou la zone) approché(e) est donné(e). Le tableau suivant résume ces éléments :

Nombre d'observations	Seuil exact	Seuil approché	Zone exacte	Zone approchée
$N_{emb} < 30000$	X	X	X	
$30000 \leq N_{emb} \leq 10^{15}$	X (Monte Carlo)	X	X (Monte Carlo)	
$N_{emb} > 10^{15}$		X		X

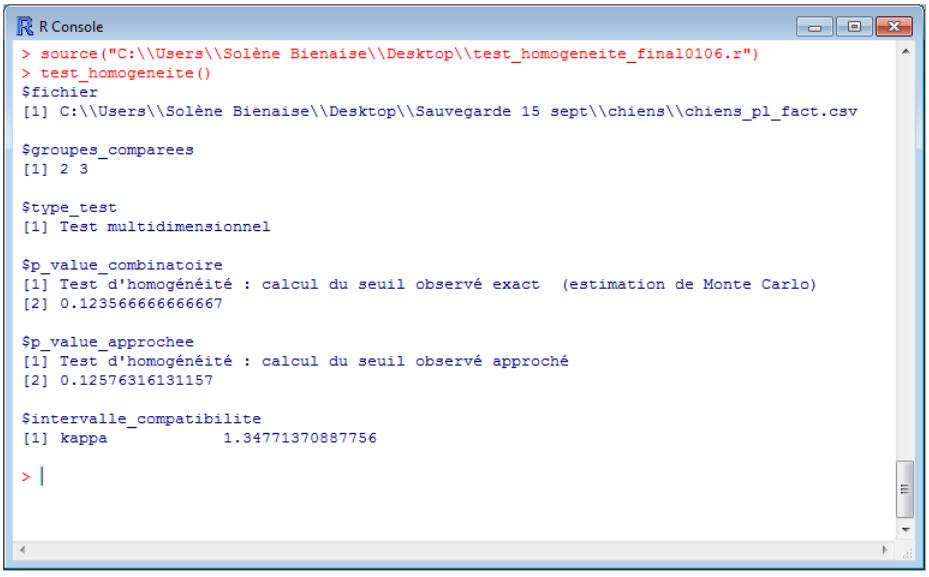
Exemple des races canines (comparaison des groupes 1, 2 et 3) :

– Fichier races_canines.txt (voir données ci–avant)
– Nombre de groupes à comparer :
3



Exemple des races canines (comparaison partielle des groupes 2 et 3) :

- Fichier races_canines.txt (voir données ci–avant)
- Nombre de groupes à comparer :
  2
- Numéro des groupes à comparer :
  2 3
- Zone de compatibilité au seuil 95% : demandée



2 Étapes de développement (test géométrique de typicalité)

Nous présentons ici les différentes étapes et astuces de développement pour le test géométrique de typicalité. Les programmes du test ensembliste de typicalité et du test d'homogénéité s'inscrivent dans la même ligne.

La structure globale du programme (fonction) qui effectue le test est la suivante :

```
#fonction globale
test_geo_typicalite()=function(){

    #fonction qui crée l'ensemble des possibles
    ensemble_typicalite(param1,param2,...)=function(){
    }

    #fonction qui calcule le seuil observé (p-value)
    seuil_typicalite(param1,param2,...)=function(){
    }

    #fonction qui calcule l'intervalle ou la zone de compatibilité
    zone_typicalite(param1,param2,...)=function(){
    }

#Dialogue avec l'utilisateur :
#Données ?
#Point de référence ?
#Zone de compatibilité ?

#Appels combinés des 3 fonctions précédentes
}
```

– *Construction de l'espace de typicalité (« complet » pour $N \leq 15$, « restreint » pour $15 < N \leq 100$).*

Lorsque $N \leq 15$, nous construisons l'espace de typicalité « complet », c'est–à–dire obtenu en effectuant les $\mathrm{J} = 2^n$ échanges possibles de $1, 2, \ldots n$ points du nuage M^I avec leurs symétriques (*cf.* chapitre 2, p.24). Lorsque $15 < N \leq 100$, nous construisons un espace de typicalité « restreint » en engendrant $\mathrm{J} = 50\,000$ échanges possibles [4] (méthode de Monte Carlo).
La fonction suivante construit la matrice $\mathbf{E} = [e^{ji}]_{\substack{j=1,\ldots,\mathrm{J} \\ i=1,\ldots,n}}$ à J lignes et n colonnes qui permet d'obtenir les J nuages de l'espace de typicalité (complet ou restreint). Pour le j–ème nuage de l'espace de typicalité, $e^{ji} = 1$ si le point M^i doit être remplacé par son symétrique et $e^{ji} = 0$ sinon.

```
#fonction ensemble_typicalite
#param 1 : nombre de points du nuage MI
#param 2 : nombre de tirages pour la méthode de Monte Carlo
   ensemble_typicalite=function(N,n_tirages_mc){

      if (N<16){
         #indique que l'espace de typicalité complet est
            #construit, c'est-à-dire que le test exact est effectué
         methode="Test géométrique de typicalité : calcul du
            seuil observé exact "

         #nombre d'échanges (par symétrie) possibles
```

4. Afin d'optimiser le temps de calcul, nous ne vérifions pas ici le fait que les échanges engendrés soient différents. En effet, la probabilité d'obtenir plusieurs fois le même échange est très faible.

```
13            Nperm=2^N
14
15            #initialisation de la matrice E, appelée Iperm ici
16            Iperm= matrix(1,nrow=Nperm,ncol=N)
17
18            for ( IP in c(2:Nperm)){
19               compt=0
20               for (j in c(1:N)) {
21                  Iperm[IP,j]= Iperm[IP-1,j]}
22               for (j in c(1:N)) {
23                  if (compt==0) {
24                     Iperm[IP,j]= Iperm[IP,j]+1
25                     if(Iperm[IP,j]==2)
26                     { Iperm[IP,j] =0
27                        compt=compt+1}}
28               } }}
29
30         if (N>15 & N<101) {
31         #indique que nous construisons un espace de typicalité
32           #restreint à l'aide de la méthode de Monte Carlo
33         methode="Test géométrique de typicalité : calcul du
34           seuil observé exact  (estimation de Monte Carlo)"
35         Nperm=n_tirages_mc
36         Iperm= matrix(1,nrow=Nperm,ncol=N)
37         for ( IP in c(1: n_tirages_mc)){
38               Iperm[IP,]= rbinom(N,1,0.5)
39               }}
40
41         #renvoie la matrice Iperm, sa méthode de construction
42           #et son nombre de lignes (nombre de nuages l'espace
43           #de typicalité)
44            list(Iperm=Iperm,methode=methode,Nperm=Nperm)
45         }
```

– *Calcul du seuil observé.*

Rappelons que le calcul de la statistique de test D consiste en un calcul de distance particulier puisqu'il s'agit de la M_O–distance des points du nuage de typicalité au point de référence O (*cf.* chapitre 2, p.25) :

$$\begin{array}{rccl} D: & \mathcal{M} & \to & \mathbb{R}^+ \\ & C^j & \mapsto & |\overrightarrow{OC^j}|_O = \sqrt{\mathbf{d}^{j\prime}\mathbf{B}_O^{-1}\mathbf{d}^j} \end{array}$$

où $\mathbf{d}^j$ est le vecteur–colonne associé à $\overrightarrow{OC^j}$ et où $\mathbf{B}_O$ est telle que définie au chapitre 2 (p.28).

Afin d'alléger les calculs et d'optimiser leurs temps d'exécution, la fonction seuil_typicalite effectue d'abord le changement de base consistant à se placer dans la base ortho–calibrée des vecteurs propres de la matrice $\mathbf{B}_O$, c'est–à–dire la base de vecteurs orthogonaux dont les normes sont égales aux racines carrées des valeurs propres.

Le nuage de typicalité est construit à partir des données dans la base ortho–calibrée (appel de la fonction ensemble_typicalite). Enfin, le calcul de la statistique de test est effectué pour chaque point du nuage de typicalité, il se ramène à un calcul de distance élémentaire, c'est–à–dire à une simple somme de carrés. En effet, la nouvelle matrice correspondant à la matrice $\mathbf{B}_{\mathrm{O}}$ après changement de base est la matrice identité.

– *Construction de la zone de compatibilité.*

Le théorème 1.1 (chapitre 2, p.31) permet de mettre en place une procédure de calcul permettant la construction de la zone de compatibilité.

Afin d'obtenir des points–limites d'incompatibilité (au seuil .05 par défaut), nous appelons la fonction seuil_typicalite un certain nombre de fois en modifiant à chaque itération les coordonnées du point de référence : nous choisissons de procéder par approximations successives en balayant uniquement les axes principaux du nuage M^I. Pour chaque axe principal, 2 points–limites sont obtenus (un du côté négatif et un du côté positif). Pour un point–limite d'incompatibilité P donné, nous calculons la valeur k telle que $k = |\overrightarrow{\mathrm{PG}}|$. En effectuant ce procédé pour les $2 \times L$ points–limites obtenus (L est la dimension du nuage), nous obtenons un ensemble de valeurs dont nous prenons la moyenne comme valeur estimée du κ de l'ellipse ajustée à cet ensemble de points–limites (théorème 1.1, p.31).

La valeur κ est alors affichée et permet la construction de la zone de compatibilité.

Dans le cas unidimensionnel, il s'agit de trouver les 2 points–limites d'incompatibilité qui constituent les bornes de l'intervalle de compatibilité.

Communications

Les travaux présentés dans cette thèse ont fait l'objet des communications suivantes :

- S. Bienaise and B. Le Roux, *Combinatorial inference in geometric data analysis : typicality test*. International Conference on Correspondence Analysis and Related Methods, Rennes, 2011.
- S. Bienaise and M. Gettler–Summa, *Validation of Trajectories on Factorial Plans after a Tandem Clustering Approach*. International Classification Conference, Saint–Andrews, 2011.
- S. Bienaise, M. Gettler–Summa, C. Godard and N. Hafner, *Multiple Small and Graphical Validation for Visual Decision Aid*. Symposium on Learning and Data Science, Florence, 2012.
- S. Bienaise and B. Le Roux, *Inférence combinatoire en Analyse Géométrique des Données : tests de typicalité*. 44èmes Journées de Statistique, Bruxelles, 2012.
- S. Bienaise and B. Le Roux, *Inférence combinatoire en Analyse Géométrique des Données : tests d'homogénéité*. 45èmes Journées de Statistique, Toulouse, 2013.

Références

[1] M.J. Anderson. A new method for non–parametric multivariate analysis of variance. *Austral Ecology*, 26(1) :32–46, 2001.

[2] M.J. Anderson and P. Legendre. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3) :271–303, 1999.

[3] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons : New–York, 1958.

[4] J.P. Benzécri. Cours Peccot au Collège de France, 5–ème leçon. 1963.

[5] J.P. Benzécri. *Pratique de l'analyse des données, Vol.1. Analyse des correspondances : Exposé élémentaire*. Dunod : Paris, 1980.

[6] J.P. Benzécri. *Pratique de l'analyse des données, Vol.1. Analyse des correspondances : Exposé élémentaire*. Dunod : Paris, 1984, 2ème éd.

[7] J.P. Benzécri and Al. *L'analyse des données. Vol.1 : Taxinomie. Vol.2 : Analyse des Correspondances*. Dunod : Paris, 1973.

[8] P. Besse and J.O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2) :285–311, 1986.

[9] R. Bro and H.A.L. Kiers. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, 17(5) :274–286, 2003.

[10] J.D. Carroll and J.J. Chang. Analysis of individual differences in multidimensional scaling via an N–way generalization of "Eckart–Young" decomposition. *Psychometrika*, 35(3) :283–319, 1970.

[11] R.B. Cattell. "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. *Psychometrika*, 9(4) :267–283, 1944.

[12] P. Cazes. Note sur les éléments supplémentaires en analyse des correspondances. II. Tableaux multiples. *Cahiers de l'analyse des données*, 7(2) :133–154, 1982.

[13] P. Cazes. Quelques méthodes d'analyse factorielle d'une série de tableaux de données. *Revue Modulad*, (31) :1, 2004.

[14] A. Chevalier, C. Blanc, D. Luce, and M. Goldberg. Facteurs de l'absentéisme médical dans une grande entreprise nationalisée. *Arch. mal. prof.*, 48(3) :223–232, 1987.

[15] A. Chevalier, J.F. Chastang, and C. Blanc. L'absentéisme pour raisons médicales à Electricité et Gaz de France : comparaisons selon le sexe. *Sciences sociales et santé*, 5(3–4) :19–40, 1987.

[16] A. Chevalier, M. Golberg, C. Godard, P. Guénel, B. Callet, B. Antonini, A. Médard, and F. Coing. Incidence des cancers dans la population masculine des salariés en activité à Electricité de France et Gaz de France. *Rev. Epidém. et Santé Publ.*, 44 :25–36, 1996.

[17] A. Chevalier and M. Goldberg. L'absence au travail : indicateur social ou indicateur de santé ? *Sciences sociales et santé*, 10(3) :47–65, 1992.

[18] A. Chevalier, M. Zins, C. Godard, J. Morin, V. Jourdain, F. François, J. Lambrozo, M. Golberg, and P. Ducimetière. Un registre des cardiopathies ischémiques chez les salariés en activité d'EDF et Gaz de France. Mise en place et premiers résultats. *Rev. Epidém. et Santé Publ.*, 49 :51–60, 2001.

[19] S. Coulondre, A. Chevalier, A. Verrier, C. Ravault, T. De Chazal, and F. Coing. Pathologies liées à l'amiante indemnisées à Electricité et Gaz de France. Bilan et évolution sur huit ans. *Revue Médicale de l'Assurance Maladie*, (3–4) :9–18, 1999.

[20] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press : Princeton, 1946.

[21] P.H. Crowley. Resampling methods for computation–intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*, 23 :405–447, 1992.

[22] J.J. Daudin, C. Duby, and P. Trecourt. Stability of principal component studied by the bootstrap method. *Statistics*, 19 :241–258, 1988.

[23] A.C. Davison and D.V. Hinkley. Saddlepoint approximations in resampling methods. *Biometrika*, 75(3) :417–431, 1988.

[24] Service Général de Médecine de Contrôle. *Classification statistique internationale des maladies et des problèmes de santé connexes, dixième version, pour Aramis*. 2009.

[25] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3) :211–218, 1936.

[26] E.S. Edgington. *Randomization Tests*. Marcel Dekker Inc : New–York, 1995, 3ème éd.

[27] B. Efron. Bootstrap methods : another look at the jackknife. *The annals of Statistics*, 2 :1–26, 1979.

[28] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall : New York, 1993.

[29] B. Escofier and J. Pagès. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation*. Dunod, 2008, 3ème éd.

[30] A.M. Ferrandez and O. Blin. A comparison between the effect of intentional modulations and the action of L–dopa on gait in Parkinson's disease. *Behavioural brain research*, 45(2) :177–183, 1991.

[31] R.A. Fisher. The design of experiments. *Oliver & Boyd : Edinburgh*, 1935.

[32] R.A. Fisher. The precision of discriminant functions. *Annals of Human Genetics*, 10(1) :422–429, 1940.

[33] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, pages 292–298, 1983.

[34] N. Gault. *Incidence du cancer du sein parmi les salariées en activité des IEG, période 1990–2006*. PhD thesis, Université Paris Diderot, 2009.

[35] M. Gettler Summa and Al. Multiple time series : new approches and new tools in data mining ; applications to cancer epidemiology. *Modulad*, (34) :37–46, 2006.

[36] M. Gettler Summa and K. Pak. Pyramidisation procedure for a hierarchy of times series based on the Kullback Leibler Divergence. Porto, 2008. Compstat.

[37] C. Godard. Indicateurs d'absence et prédiction de la survenue d'événements de santé graves : haut risque de décès dans une population salariée en apparente bonne santé. *Sciences sociales et santé*, 5(3–4) :85–104, 1987.

[38] C. Godard. Réflexion sur l'intérêt de la régression de Poisson pour la surveillance épidémiologique systématique en population salariée. Master's thesis, Université Paris Sud, 2009.

[39] C. Godard, A. Chevalier, B. Siret, J.F. Giorla, T. Hergueta, Y. Lecrubier, J.G. Bauer, C. Bolle, J. Coste, J.P. Sperte, T. Lault, and G. Lahon. Prévention des troubles anxieux et dépressifs par une action d'éducation pour la santé en consultation : résultats du programme APRAND. *Rev. Epidém. et de Santé Publ.*, 2007.

[40] M. Goldberg and Al. Une base de données concernant la santé de la population des travailleurs d'une entreprise à l'échelle nationale et son utilisation à des fins épidémiologiques. *Arch. mal. prof.*, 41(2) :85–91, 1980.

[41] M. Goldberg, A. Leclerc, J.F. Chastang, and Al. Une cohorte épidémiologique à EDF–GDF : l'opération 20 000 volontaires pour la recherche médicale. *Hommes et Santé*, 60(7–14), 1991.

[42] L. Guttman. The quantification of a class of attributes : A theory and method of scale construction. *The prediction of personal adjustment*, pages 251–264, 1941.

[43] R.A. Harshman. Foundations of the PARAFAC procedure : models and conditions for an "explanatory" multi–modal factor analysis. *University of California at Los Angeles, Working Papers in Phonetics*, 16 :1–84, 1970.

[44] C. Hayashi. Theory and examples of quantification.(II). 4(2) :19–30, 1956.

[45] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics, vol. 2.* 1973, 3ème éd.

[46] P.M. Kroonenberg and Al. Singular value decompositions of interactions in three-way contigency tables. *Elsevier*, pages 169–184, 1989.

[47] J.B. Kruskal. Three-way arrays : Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2) :95–138, 1977.

[48] B. Le Roux. Inférence combinatoire en analyse géométrique des données. *Mathématiques et sciences humaines*, 144, 1998.

[49] B. Le Roux and H. Rouanet. Interpreting axes in Multiple Correspondence Analysis : Method of the contributions of points and deviations. *Visualization of Categorical Data*, pages 197–220, 1998.

[50] B. Le Roux and H. Rouanet. *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis.* Kluwer, 2004.

[51] B. Le Roux and H. Rouanet. *Multiple Correspondence Analysis*, volume 163 of *QASS*. Thousand Oaks : Sage Publications, 2010.

[52] L. Lebart. Validité des résultats en Analyse des Données. *Rapport CREDOC–DGRST*, 1975.

[53] L. Lebart. The significance of eigenvalues issued from correspondence analysis. Vienne, 1976. Compstat.

[54] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle.* Dunod : Paris, 1995.

[55] H. L'Hermier des Plantes. *Structuration des tableaux à trois indices de la statistique : théorie et application d'une méthode d'analyse conjointe.* PhD thesis, Université des sciences et techniques du Languedoc, 1976.

[56] D. Luce, A. Leclerc, A. Chevalier, M. Goldberg, and C. Blanc. Facteurs de risques individuels des accidents du travail à Electricité de France–Gaz de France. *Arch. mal. prof.*, 48(6) :461–465, 1997.

[57] E. Malinvaud. *Statistical Methods of Econometrics*. Rand McNally : Chicago, 1980.

[58] B.F.J. Manly. *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman & Hall, 1997, 2ème éd.

[59] Pierre Merle. *La démocratisation de l'enseignement (nouvelle édition)*. La découverte, 2009.

[60] H. Morin. *Théorie de l'échantillonnage*. Presses de l'Université de Laval, 1993.

[61] D.F. Morrison. *Multivariate statistical methods*. McGraw–Hill, 1967.

[62] H. Oja. On permutation tests in multiple regression and analysis of covariance problems. *Australian Journal of Statistics*, 29(1) :91–100, 1987.

[63] K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag*, 2(6) :559–572, 1901.

[64] F. Pesarin and L. Salmaso. *Permutation tests for complex data : theory, applications and software*. Wiley, 2010.

[65] E.J.G. Pitman. Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society, suppl*, 4(1) :119–130, 1937.

[66] M. Poncet, C. Chevalier, F. Bumsel, and G. Lahon. La mortalité des salariés d'EDF–GDF : disparités socioprofessionnelles et évolution. *Rev. Epidém. et de Santé Publ.*, 51 :481–491, 2003.

[67] J.M Pouplet, A. Chevalier, and J. Lambrozo. Les séquelles graves d'accidents du travail à Electricité et Gaz de France. *Arch. mal. prof.*, 48(2) :101–107, 1987.

[68] J.O. Ramsay and B.W. Silverman. *Functional data analysis*. Springer : New York, 2005, 2ème éd.

[69] J. Rice. Functional and longitudinal data analysis : Perspectives on smoothing. *Statistica Sinica*, 14(3) :631–648, 2004.

[70] H. Rouanet. Bayesian methods for assessing importance of effects. *Psychological Bulletin*, 119(1) :149–158, 1996.

[71] H. Rouanet, J.M. Bernard, M.C. Bert, B. Lecoutre, M.P. Lecoutre, and B. Le Roux. *New Ways in Statistical Methodology*. Peter Lang, 1998.

[72] H. Rouanet, J.M. Bernard, and B. Le Roux. *Analyse Inductive des Données*. Dunod, 1990.

[73] H. Rouanet, J.M. Bernard, and B. Lecoutre. Nonprobabilistic statistical inference : A set–theoretic approach. *The American Statistician*, 40 :60–65, 1986.

[74] H. Rouanet and B. Le Roux. *Analyse des Données Multidimensionnelles*. Dunod, 1993.

[75] H. Rouanet and B. Lecoutre. Specific inference in ANOVA : From significance tests to Bayesian procedures. *British Journal of Statistical and Mathematical Psychology*, 36 :252–268, 1983.

[76] H. Rouanet, D. Lepine, and J. Pelnard-Considere. Bayes–fiducial procedures as practical substitutes for misplaced significance testing : An application to educational data. *Advances in psychological and educational measurement*, pages 33–48, 1976.

[77] C. Rumeau-Rouquette, B. Blondel, and M. Kaminski. *Epidémiologie : Méthodes et pratique*. Flammarion médecine-sciences, 1993.

[78] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006, 2ème éd.

[79] G. Saporta and G. Hatabian. Régions de confiance en analyse factorielle. *Data analysis and informatics*, pages 499–508, 1986.

[80] H. Savelli, A. Chevalier, C. Puppinck, and F. Coing. L'entrée en longue maladie dans les entreprises Electricité de France–Gaz de France. *Revue Médicale de l'Assurance Maladie*, (1) :19–31, 1994.

[81] A. Shadrokh. *Analyse comparative des tests de permutations en régression multiple et application à l'analyse de tableaux de distances*. PhD thesis, Université Joseph Fourier – Grenoble I, 2006.

[82] N.D. Sidiropoulos and R. Bro. On the uniqueness of multilinear decomposition of N–way arrays. *Journal of chemometrics*, 14(3) :229–239, 2000.

[83] J.J. Sylvester. On the reduction of a bilinear quantic of the nth order to the form of a sum of n products by a double orthogonal substitution. *Messenger of Mathematics*, 19 :42–46, 1889.

[84] M. Tenenhaus. *Statistique : Méthodes pour décrire, expliquer et prévoir*. Dunod, 2007.

[85] C.J.F. ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques*, pages 79–85. Springer, 1992.

[86] L.R. Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change*, pages 122–137, 1963.

[87] L.R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3) :279–311, 1966.

[88] J.W. Tukey. Sunset salvo. *The American Statistician*, 40(1) :72–76, 1986.

[89] S. Van Aelst and G. Willems. Fast and Robust Bootstrap for Multivariate Inference : The R Package FRB. *Journal of Statistical Software*, 53(3) :1–32, 2013.

[90] X. Zhao, J.S. Marron, and M.T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14(3) :789–808, 2004.

Résumé. La première partie de la thèse traite d'*inférence combinatoire* en *Analyse Géométrique des Données* (AGD). Nous proposons des tests multidimensionnels sans hypothèse sur le processus d'obtention des données ou les distributions. Nous nous intéressons ici aux problèmes de *typicalité* (comparaison d'un point moyen à un point de référence ou d'un groupe d'observations à une population de référence) et d'*homogénéité* (comparaison de plusieurs groupes). Nous utilisons des procédures combinatoires pour construire un ensemble de référence par rapport auquel nous situons les données. Les statistiques de test choisies mènent à des prolongements originaux : interprétation géométrique du seuil observé et construction d'une *zone de compatibilité*.

La seconde partie présente l'étude de l'absentéisme dans les Industries Electriques et Gazières de 1995 à 2011 (avec construction d'une cohorte épidémiologique). Des méthodes d'AGD sont utilisées afin d'identifier des *pathologies émergentes* et des *groupes d'agents sensibles*.

Mots clefs. Analyse Géométrique des Données, inférence combinatoire, tests de typicalité, tests d'homogénéité, zones de compatibilité, absentéisme, cohorte.

Abstract. The first part of this PhD thesis deals with *combinatorial inference methods* for *Geometric Data Analysis* (GDA). We propose multidimensional tests that make no assumption on the process of generating data or distributions. We focus particularly on problems of *typicality* (comparison of a mean point to a reference point or comparison of a group of observations to a reference population) and on problems of *homogeneity* (comparison of several groups). These methods consist in using combinatorial procedures to build a reference set with respect to which we situate the data. The chosen test statistics lead to original extensions : geometric interpretation of the observed level and construction of a *compatibility zone*.

The second part of this thesis presents the study of absenteeism in the French Electricity and Gas Industries from 1995 to 2011 (with construction of an epidemiological cohort). GDA methods are used to identify *emerging diseases* and *sensitive groups of agents*.

Keywords. Geometric Data Analysis, combinatorial inference, typicality tests, homogeneity tests, compatibility zone, absenteeism, cohort.