INSA 4 GMM MMS – Exploration Statistique

Corrigé du Contrôle du 20 janvier 2012

1. *Le tableau ci-dessous fournit le « tri à plat » de quelques variables. Quels problèmes émergent dès la prise en main des données ?*
   Un volume très important de données manquantes (4228), les mêmes communes à plusieurs variables et des modalités à faibles effectifs qui poseront des difficultés.
2. *Le graphique ci-dessous illustre l'histogramme de la variable « Age ». La partie plus foncée correspond aux observations pour lesquelles la variable « Couple » est « NA ». Caractériser ces « NA ».*

Les NA sont massivement les plus jeunes enquêtés, sans autonomie.

3. *Même graphique avec les « NA » de la variable « Work ». Comment caractériser ces « NA » dans ce cas ? Compte tenu de ses objectifs, le statisticien de service choisit de supprimer les individus de moins de 20ans. Justifier ce choix*

On retrouve les mêmes « plus jeunes » mais aussi sans doute des inactifs de tous âges qui n'ont pas répondus. Ces derniers sont difficiles à caractériser. Les moins de 20 ans majoritairement pas « autonomes » ne sont pas considérés par ces aspects patrimoniaux.

4. *Il reste encore des « NA » dans la variable « Work ». Les graphiques ci-dessous (variables « Age » vs. Asvi ») sont tracés pour les individus de plus de 20 ans mais* avec *les « NA » restants de « Work » (à gauche) et* sans *ces « NA » (à droite). Que représentent ces graphiques, que suggèrent-ils ?*

Ce sont des diagrammes boîtes parallèles. Ces graphiques suggèrent que les NA restant dans la variable Work n'interviennent pas dans la liaison entre les variables Asvi et Age. Les supprimer n'introduira pas de biais à ce niveau.

5. *On se restreint toujours aux individus de plus de 20 ans. Que représente le tableau ci-dessous ? Quels problèmes sont toujours présents ?*

Il s'agit d'une table de contingence croisant les deux variables. On retrouve de nouveau les NA communs à ces deux variables et des modalités comme Dbact, Dbacg, Waut à trop faibles effectifs.

6. *Les commandes SAS :* `proc corresp ; table Diplome, Work ; run ;` *fournissent les résultats ci-après. D'où viennent les «* valeurs singulières *» ? Comment sont calculées les «* Inerties principales *? Quel rapport entre la valeur masquée xxxx.xx et la table précédente ? Que vaut-elle ?*

Analyse des correspondances simples : les inerties principales sont les valeurs propres de la matrice produit des matrices (AB ou BA) contenant les vecteurs profils lignes et colonnes. Ce sont les carrés des valeurs singulières de la matrice 1/n T. La troisième colonne décompose la statistique du Khi-2 (n fois les valeurs propres). `xxxx.xx=0.44152*17059=7531.9` est la statistique du khi-2 de la table de contingence précédente.

7. *Que valent les valeurs masquées par des $$? Comment interpréter le graphique ?*

Ce sont les parts de khi62 expliquées puis cumulées pour chaque valeur propre.
`$$= 5823.99/7531.9=0.77 ; $$=1103.18/7531.9=14.65 ; $$=0.77+0.15=0.92.`
On se limite à deux dimensions pour représenter l'AFC de cette table.

8. *Un autre graphe est obtenu ci-dessous en supprimant les « NA » des deux variables. Que conclure ? Pour les étudiants « francophones » : interpréter les proximités des modalités de ce graphique.*

La suppression des NA n'a pas modifié le graphique. Cette suppression est « validée ». Le graphique montre simplement la liaison entre la catégorie croissante de diplôme et le niveau professionnel.

9. *Le statisticien de service décide finalement d'exécuter le programme ci-après. Justifier ces choix.*

Les observations NA peu intéressantes et surtout gênantes pour l'étude sont supprimées car cette suppression ne semble pas introduire de biais spécifique aux objectifs de l'étude. Les modalités à faible effectifs (U6, Znor, Zest) sont regroupées ou supprimées (Meri, Peri). Les variables nationalité et xcapi sont supprimées car quasi constantes.

10. *Une analyse factorielle multiple des correspondances fournit les résultats ci-dessous. Quelle interprétation tirer des pourcentages cumulés ?*

Comme dans toute afc multiple, les pourcentages cumulés sont très faibles mais sans signification ; Ne pas en tenir compte de façon « absolue ». Les deux ou trois premières composantes semblent pertinentes.

11. *Le graphique obtenu incite le statisticien à ré-exécuter le même programme avec une option du type :* `supplementary Tage Zeat Couple Mere Pere;` *Pourquoi ?*

Certaines modalités sont redondantes : les personnes veuves ou dont les parents sont décédés sont plus âgés, les jeunes ne sont pas en couple et ont toujours des grands parents. Elles sont considérées et représentées comme « supplémentaires » mais ne participent pas au calcul des axes.

12. *L'exécution avec cette option conduit au graphe ci-dessous. Expliquer brièvement les caractéristiques de la dispersion des modalités selon la direction du premier axe.*

Il s'agit d'un axe « richesse » : A droite les modalités associées à la possession de patrimoines et d'épargnes, les diplômes élevés et les catégories A. A gauche les catégories C et sans diplômes.

13. *Quelle est l'interprétation de la dispersion selon le 2ème axe ?*

Cet axe reste structuré par l'âge et ses conséquences même si cette variable n'est pas active.

14. *Le code suivant est exécuté. De quelle matrice, sera calculée la décomposition en valeur singulière pour réaliser cette analyse multiple des correspondances. Dans quel objectif ?*

Il s'agit de l'afcm par afc du tableau disjonctif complet dont la SVD est calculée. Cela permet de calculer les composantes principales de l'acp des profiles lignes qui sont les coordonnées des individus. Cela fournit une représentation quantitative des individus.

15. *Comment est obtenu ce graphique par rapport au précédent ? Que dire de la position relative des rouges et des noirs ? Que pouvons-nous penser de l'objectif de prévision de la possession d'un produit « assurance vie » ?*

Le calcul des composantes principales des individus permet de les représenter avec les modalités sur le même graphe. Les points rouges (possesseurs d'assurance vie) sont évidemment très largement majoritaires sur la droite, cela correspond aux personnes les plus aisées. Néanmoins, la séparation des deux classes de possesseurs et non possesseurs n'est pas nette : elle est même très confuse.

16. *De façon classique, tout du moins en France, le statisticien récupère de l'exécution précédente une matrice quantitative de représentation vectorielle des individus. Quelle est cette matrice ? Il l'utilise pour construire une classification non-supervisée. Pourquoi procéder ainsi ?*

Il s'agit de la matrice des composantes principales de l'ACP des profiles lignes représentant les individus. La démarche permet de récupérer des représentations quantitatives car les algorithmes de classification ne fonctionnent qu'avec de telles variables.

17. *Quel est l'intérêt de la CAH par rapport à kmeans ? Quel problème pourrait poser la CAH avec ces données ? Comment le contourner ? Quelle option faut-il fixer en CAH et quel choix est-il conseillé ?*

Le choix du nombre de classes se fait *a posteriori* avec la CAH. La taille élevée de l'échantillon peut poser des problèmes de mémoire pour stocker la matrice des distances des individus 2 à 2 en CAH. Néanmoins, ça marche avec SAS mais peut être pas avec R. Le contourner en enchaînant kmeans avec beaucoup de classes puis la CAH sur les barycentres des classes. Il faut fixer la distance entre groupes, le saut de Ward est généralement privilégié.

18. *La CAH conduit au graphe suivant. Que représente-t-il et quelle conclusion en tirer ? Quel graphe n'est pas représenté car trop complexe ?*

Le R-carré semi-partiel est la décroissance de la variance intra-classe. Avec le saut de Ward cela représente aussi les hauteurs des premières branches du dendrogramme ou arbre de classification hiérarchique pas représenté car trop complexes : beaucoup trop de feuilles.

19. *Comment est obtenue la représentation ci-dessous ?*

Après un choix de 5 classes pour faire simple, le même graphe des individus est représenté en coloriant les points en fonction de leur classe.

20. *Comment le graphe ci-dessous est-il obtenu ? Que représente les modalités K1 à K5. Caractériser brièvement celles les plus fréquentes K2, K4, K5 ?*

Les 5 classes obtenues sont les modalités d'une nouvelle variable. Celle-ci est ajoutée dans l'AFCM précédente pour aider l'interprétation. La classe K5 représente les personnes les plus aisées avec le plus d'épargne et donc de patrimoine. Ils sont généralement possesseurs de contrat assurance vie. Les deux autres classes sont les personnes les moins aisées, elles se distinguent principalement entre elles par l'âge : les plus jeunes dans K4, les plus âgées dans K5.