PSY38X2 : Traitement de données en Psychologie - TD N°5
Analyse Factorielle des Correspondances avec Minitab et Modalisa

Dossier REGIONS

Dans les deux premiers paragraphes, nous allons travailler sur les données "Régions".
Le tableau rapporte des données du recensement de 1968: il indique pour chacune des 22 régions françaises (en lignes) le nombre d'habitants (en milliers) par âge (en colonnes) HF00 signifie Hommes et Femmes de 0 à 4 ans, HF05 signifie Hommes et Femmes de 5 à 9 ans,... HF75 signifie Hommes et Femmes de plus de 75 ans.

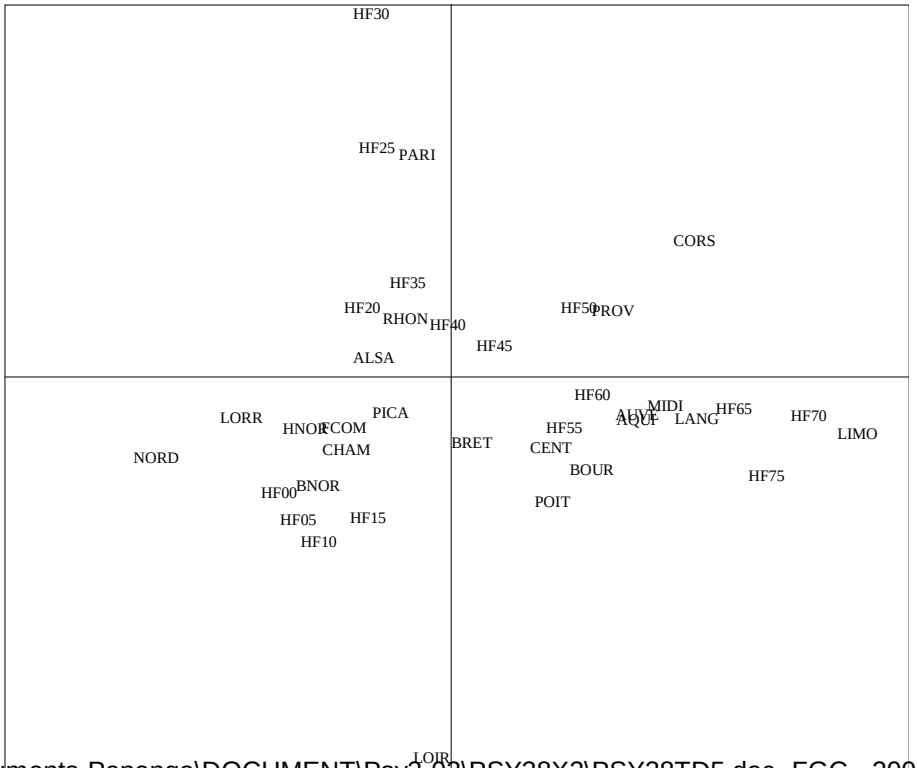
NB. Le fichier de données comporte au moins quatre erreurs de saisie, pour les combinaisons de modalités Picardie - HF15, Pays de la Loire - HF30, Nord Pas-de-Calais - HF65 et HF70. Le tableau de données erroné a été laissé tel quel ici, car certaines questions sont en lien avec les résultats effectivement trouvés sur ce tableau. Les résultats correspondant aux données rectifiées sont également sous forme de page html.

1) Analyse Factorielle des Correspondances avec Modalisa

Logiciel de traitement d'enquête, Modalisa est essentiellement conçu pour réaliser l'analyse factorielle des correspondances à partir du protocole des questionnaires saisis. Mais nous n'envisageons pas de saisir 50000 questionnaires indiquant la région de résidence et la classe d'âge de chaque groupe de 1000 personnes interrogées...
Nous nous servirons donc d'une autre fonctionnalité prévue par le logiciel : effectuer une AFC à partir d'un tableau de contingence enregistré au format texte. Mais les possibilités de traitement et de sauvegarde sont alors très réduites.

- A partir de la fenêtre du Pilote, exécutez le menu Analyse - AFC : Correspondances.
- Dans la fenêtre de dialogue suivante, sélectionnez l'item : "Nouvelle analyse à partir d'un tableau de nombres à importer".
- Conservez les paramètres par défaut des fenêtres de l'importation.

Modalisa affiche alors le graphique de l'AFC selon les deux premiers axes :

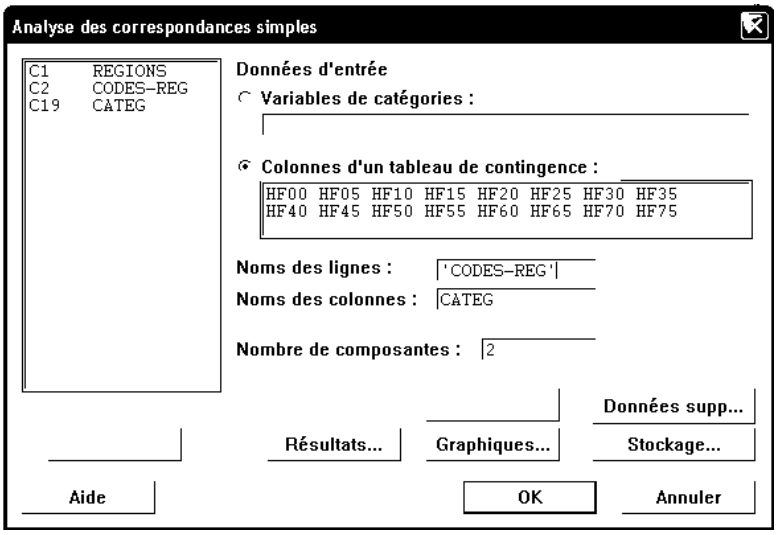


On peut alors :

- Explorer d'autres axes factoriels (menus Carte - Facteur X N° et Facteur Y N°)
- Afficher les contributions des individus-lignes et des individus-colonnes à l'inertie des facteurs (menu Carte - Editer les contributions) et enregistrer ces contributions dans un fichier texte.
- N'afficher qu'une sélection de points à l'aide du menu Carte - Seuils liens et contributions.
- Copier le graphique pour le coller dans un autre document (comme ici dans un document Word, par exemple).

2) Analyse Factorielle des Correspondances avec Minitab

Chargez Minitab et ouvrez le projet Minitab W:\PSY3\TD-Minitab\Regions.mpj
Utilisez le menu Stat - Tableaux - Analyse des Correspondances simples.
On pourra compléter la fenêtre de dialogue comme suit :



Sélectionnez également le sous-dialogue "Graphiques..." et demandez les trois premiers graphiques.

Outre les graphiques, Minitab affiche dans la fenêtre Session un certain nombre de résultats numériques :

- La liste des valeurs propres, avec la proportion de variance expliquée par le cumul des premières VP.
- Les contributions des lignes (Qualité, Masse et Inertie des individus lignes, ainsi que leurs scores, qualités et contributions sur chacune des composantes principales demandées.
- Les mêmes éléments concernant les colonnes.

On peut obtenir d'autres paramètres numériques en utilisant le sous-dialogue Résultats... de la fenêtre de dialogue Analyse de Correspondances Simples.
On peut ainsi obtenir également les profils-lignes, les profils-colonnes, le tableau des inerties relatives, etc.

Il peut également être intéressant d'examiner d'autres axes que les axes 1 et 2. Refaites l'analyse factorielle des correspondances en spécifiant 3 ou 4 axes et en demandant les graphiques entre les axes 1 et 3, 2 et 3.

Après avoir enregistré le fichier Regions.mpj, envoyez-le par mail à votre enseignant.

3) Interprétation des résultats obtenus

3.1 - Quelques questions relatives à l'interprétation des résultats obtenus à l'aide de Minitab ou de Modalisa.

1) Pourquoi y a-t-il 15 valeurs propres dans ce cas?

2) La somme des valeurs propres est égale à 0.01377. A quoi correspond cette valeur (deux réponses attendues) ?

3) On retiendra ici, pour la suite de l'analyse, 2 axes factoriels. Ceci peut se justifier par la décroissance forte des valeurs propres entre la deuxième et la troisième variable factorielle ("critère du coude"). Cependant, deux arguments conduiraient à retenir plutôt 3 axes factoriels. Indiquer ces deux arguments:

4) D'après l'examen de ce plan factoriel (1x2), indiquer les deux régions qui semblent avoir la répartition des âges la plus éloignée de la répartition moyenne (toutes régions confondues)

5) Toujours d'après l'examen de ce graphique factoriel, indiquer une région qui semble avoir un profil d'âge proche du profil moyen de la population française.

6) Pourquoi faut-il retourner aux données pour vérifier l'exactitude de ce qui est ainsi suggéré par ce graphique factoriel ?

7) Pour la région Picardie (PICA) Minitab indique Qual = 0,079.
a) Que signifie cette valeur?
b) Cette région (Picardie) apparaissait proche du centre de gravité sur le nuage des régions . Quel commentaire peut-on faire maintenant, après avoir pris connaissance de la valeur de Qual?

8) Pour Paris (PARI), la colonne Mass indique 0,187.
Pourquoi cette valeur est-elle si élevée ?

9) La colonne Inert indique, pour les Pays de la Loire, une valeur relativement élevée : 0,184.
a) A quoi correspond cet indice?
b) Pourquoi cette valeur est-elle si élevée (2 réponses attendues) ?

10) Pour l'Auvergne (AUVE), on constate une valeur élevée de Corr pour l'axe 1 (0,918) et une valeur faible de ce même indice pour l'axe 2 (0,032). Interpréter ces valeurs.

11) Pour qualifier d'importante la contribution d'un point à un axe factoriel, on calcule une valeur-repère. Pour les données Régions,on pourra prendre comme valeur repère 1/22=0,046
Utiliser cette valeur repère pour déterminer les régions contribuant aux axes 1 et 2.
b) Indiquer en quoi consiste, en résumé, l'axe 1.

12) Sur le graphe factoriel relier les points-âges entre eux, selon l'ordre naturel.
En déduire une première interprétation de l'axe 1 (axe horizontal) du point de vue de l'âge, au vu de ce graphique
Que peut-on en déduire concernant la région Limousin (LIMO), située à l'extrême gauche du graphe des régions ?

13) L'examen des contributions à l'axe 2, montre que cet axe oppose les 5 et 10 ans (côté négatif de l'axe, en bas du graphe), aux 25 et 30 ans (côté positif, en haut du graphe). Que peut-on en déduire sur les caractéristiques de la région parisienne (PARI), située en haut de la figure ?

3.2 - Quelques indications de réponses

1) Plus petite dimension du tableau (16 âges) moins 1 = 15

2) Variance totale du nuage. Phi-2 du tableau de contingence initial.

3) Le troisième axe a une contribution supérieure à la contribution moyenne (1/15) = 7%)

La part de variance expliquée serait plus grande (90% au lieu de 81%) Certaines régions seraient mieux représentées avec 3 axes, par exemple la Picardie (QLT = 79/1000 avec 2 axes)

4) Le Limousin (LIMO) et les Pays de Loire (LOIR) car ce sont les plus éloignées du centre de gravité du nuage (intersection des 2 axes factoriels)

5) La Bretagne (ou Picardie ou Alsace ou Rhone)

6) Il est possible que les distances lues sur ce graphe factoriel aient été déformées par la projection du nuage de points, situé en fait dans un espace de dimension 15, sur cet espace de dimension 2 (le plan factoriel)

7) a) La Picardie est mal représentée par ce plan 1x2
b) La distance de ce point Picardie au centre de gravité du nuage est peut-être plus importante qu'il n'y paraît à la lecture du graphe factoriel. Il est préférable de ne pas interpréter ce point sur la seule base de ce plan factoriel.

8) La Région Parisienne est la plus peuplée des régions de France (plus de 9 millions d'habitants)

9) a) C'est la contribution relative (en millièmes), de la Région Pays de la Loire (LOIR), à la variance totale du nuage.
b) Son poids relatif est important (c'est une région fortement peuplée 2 468 000 h). La distance de son profil au profil moyen est grande (on l'a vu sur le graphe factoriel).

10) Cette région est bien représentée sur l'axe 1 et très mal sur l'axe 2.

11) b) Il oppose des régions situées au sud de la France à des régions situées au nord de celle-ci (+)

12) L'axe oppose les âges élevés (-) aux jeunes (+)
Cette analyse factorielle suggère que le Limousin est peuplée d'une proportion relativement forte de personnes âgées et d'une proportion relativement faible de jeunes (relativement à la répartition moyenne de la population française par âge).

13) Paris apparaît peuplée d'une forte proportion de jeunes adultes (25 à 30 ans) et d'une faible proportion d'enfants et adolescents (0 à 19 ans)

4) Exercice

Les données "Conjoint" étudiées en TD de statistiques, ont été saisies dans le projet Minitab W:\PSY3\TD-Minitab\Conjoint.mpj et dans le fichier texte W:\PSY3\TD-Modalisa\Conjoint.txt.

Utilisez Modalisa et Minitab pour retrouver les résultats indiqués sur la fiche de TD de statistiques.

5) Travail de monitorat à rendre

Un corpus de 200 publicités visant la consommation de produits alimentaires plus spécialement destinés aux enfants a été analysé.

Ces publicités concernaient onze catégories de produits: laitages nature (LNAT), laitages fruités (LFRU), fromages frais (FROF), fromages-pâtes (FROP), eaux minérales (EAUX), boissons sucrées (BOIS), produits pour petits déjeuners (PDEJ), jus de fruits (JUSF), confiseries (CONF), barres chocolatées (BARR), chocolats (CHOC).

Ces publicités s'appuient sur la mise en scène de différentes valeurs : Gourmandise/Plaisir (GOUR), Nature/Ecologie (NATU), Forme/Santé (SANT), Aventure/Evasion (AVEN), Tendresse/Affection (TEND), Prestige/Luxe (PRES), Séduction-Erotisme (SEDU), Convivialité/Partage (CONV), Tradition/Gastronomie (TRAD), Innovation/Modernisme (INOV), Folie/Délire (FOLI).

Une publicité pour un produit peut s'appuyer sur une ou plusieurs valeurs.

Les données rassemblées dans le tableau de contingence ci-dessous indiquent le nombre de fois où une valeur a été associée à un produit.

PROD	GOUR	NATU	SANT	AVEN	TEND	PRES	SEDU	CONV	TRAD	INOV	FOLI
LNAT	1	4	9	2	7	2	1	2	2	2	0
LFRU	7	5	6	0	5	1	0	2	1	0	0
FROF	12	10	1	0	7	5	8	4	5	2	0
FROP	12	8	0	4	3	6	4	6	9	6	0
EAUX	1	9	13	4	2	4	1	0	0	0	2
BOIS	3	3	8	10	2	0	6	2	0	1	1
PDEJ	11	5	10	7	0	0	0	3	0	1	2
JUSF	0	4	3	1	0	3	2	2	0	0	0
CONF	5	4	1	5	1	0	4	4	0	2	3
BARR	19	8	6	15	4	1	2	0	1	1	3
CHOC	11	1	4	4	5	12	5	7	1	0	4

Ces données se trouvent également dans les fichiers W:\PSY3\TD-Minitab\Publicites.MPJ et W:\PSY3\TD-Modalisa\Publicites.txt.

Traiter des données à l'aide d'une analyse factorielle des correspondances, en se limitant aux deux premiers axes factoriels. On s'intéressera en particulier aux questions suivantes.

1) Caractériser le profil de Chocolat (CHOC) par rapport au profil moyen.
2) On réalise une analyse factorielle des correspondances, en ne retenant que les deux premiers axes. Compte tenu des informations fournies par le tableau des valeurs propres, ce choix semble-t-il pertinent ?
3) On constate que EAUX et FROP sont éloignés l'un de l'autre sur le graphique. Qu'est-ce que cela suggère ?
4) Expliciter ce qui distingue principalement les groupes de produits qui s'opposent sur l'axe 1.
5) De même, expliciter ce qui distingue les groupes de produits qui s'opposent sur l'axe 2.

Composez un document Word comprenant un diagramme selon les axes principaux 1 et 2, et les réponses aux questions indiquées ci-dessus.
Envoyez ce document par mail à votre enseignant.