| | |
|---|---|
| EISTI Ecole Internationale des Sciences du Traitement de l'Information | **GSI-MI-SIE : RESUME D'ANALYSE MULTIVARIEE**<br>**ANALYSE DE LA VARIANCE - RLM - ACP - AFC** |
| Rédigé par : les élèves et Hervé de Milleville | Ref : *SIE-SE-TD-TB-CL-IN* |
| A l'intention de : Etudiants de GSI – MI - SIE | Créé le : 18/12/2012 modifié le 05/01/2013 |

# Sommaire

# 1. Analyse de la variance

## *1.1 Le corpus de données*

Un ensemble de n individus observés sur une variable quantitative Y et une ou plusieurs variables $X_1$, …, $X_p$ qualitatives appelés facteurs

## *1.2 Quelle question se pose-t-on ?*

Les variations de la variable Y dépendent des facteurs (isolément et/ou simultanément )

## *1.3 Comment ?*

On étudie les différentes variances de Y d'où le nom de la méthode :

- Variance totale : celle calculée indépendamment du ou des facteurs
- Variance expliquée par le ou les facteurs
- Variance résiduelle : celle non expliquée par les facteurs

On teste le rapport dont le numérateur est directement proportionnel à la variance de Y expliquée par le ou les facteurs et le dénominateur est directement proportionnel à la variance non expliquée de Y dite variance résiduelle de Y.

### 1.3.1 Analyse de la variance à un facteur

#### 1.3.1.1 Le tableau des variations

| Source de variation de Y | Somme des carrés | Degré de liberté | Carré moyen |
|---|---|---|---|
| Variation expliquée par X | $SS_X^2 = n.S_X^2$ | k-1 | $\frac{SS_X^2}{k-1}$ |
| Variation résiduelle | $SS_R^2 = n.S_R^2$ | n-k | $\frac{SS_R^2}{n-k}$ |
| Variation totale | $SS_X^2 + SS_R^2$ | n-1 | |
| La variable de décision | | | $F_{obs} = \frac{\frac{SS_X^2}{k-1}}{\frac{SS_R^2}{n-k}}$ |

#### 1.3.1.2 La règle de décision

Dans ce qui suit, pour tout j ∈ {1, …, k}, on définit les moyennes $m_j$ de Y sur les sous populations { X = $x_j$ }.

Si le facteur X n'a pas d'influence sur les variations de Y alors toutes ces moyennes sont égales entre elles. On fait donc un test avec H0 : $m_1 = \ldots = m_k$ contre l'hypothèse $H_1$ : $\exists(j_1, j_2) \ / m_{j_1} \neq m_{j_2}$

Si $H_0$ est vraie alors la loi de la variable $F_{obs}$ est une loi de Fisher-Snedecor $F$(k-1,n-k).

Pour un seuil $\alpha$ donné, on cherche dans la table de Fisher-Snedecor la valeur f telle que P( F ≥ f) = $\alpha$ et on adopte la règle suivante :

- Si $F_{obs} \leq$ f alors on accepte l'hypothèse $H_0$ : le facteur X n'explique rien (ou si peu) des variations de Y.
- Si $F_{obs} \geq$ f alors on refuse l'hypothèse $H_0$ : les variations de Y sont significativement différentes sur au moins deux modalités de X

### 1.3.2 Analyse de la variance à deux facteurs

Pour plus de simplicité, on supposera le traitement est équilibré. Cela signifie que le nombre d'observations de la variable Y est le même pour chaque couple de modalité $(x_{1,j_1}, x_{1,j_2})$. On notera r ce nombre commun.

De plus, pour i ∈{1,2} on notera $k_i$ le nombre de modalités du facteur $X_i$.

#### 1.3.2.1 Le tableau des variations

| Source de variation de Y | Somme des carrés | Degré de liberté | Carré moyen | Variable de décision *F* |
|---|---|---|---|---|
| Variation expliquée par $X_1$ | $SS^2_{X_1} = k_2.r.S^2_{X_1}$ | $k_1$-1 | $\frac{SS^2_{X_1}}{k_1 - 1}$ | $F_1 = \frac{\frac{SS^2_{X_1}}{k_1 - 1}}{\frac{SS^2_R}{k_1.k_2.(r-1)}}$ |
| Variation expliquée par $X_2$ | $SS^2_{X_2} = k_1.r.S^2_{X_2}$ | $k_2$-1 | $\frac{SS^2_{X_2}}{k_2 - 1}$ | $F_2 = \frac{\frac{SS^2_{X2}}{k_2 - 1}}{\frac{SS^2_R}{k_1.k_2.(r-1)}}$ |
| Variation expliquée par l'interaction de $X_1$ et $X_2$ | $SS^2_{X_1X_2} = r.S^2_{X_1X_2}$ | ($k_1$-1)( $k_2$-1) | $\frac{SS^2_{X_1X_2}}{(k_1 - 1).(k_2 - 1)}$ | $F_{1,2} = \frac{\frac{SS^2_{X_1X_2}}{(k_1 - 1).(k_2 - 1)}}{\frac{SS^2_R}{k_1.k_2.(r-1)}}$ |
| Variation résiduelle | $SS^2_R$ | $k_1.k_2.(r-1)$ | $\frac{SS^2_R}{k_1.k_2.(r-1)}$ | |

| | | | | |
|---|---|---|---|---|
| Variation totale | $SS^2 = \begin{matrix} SS^2_{X_1} + SS^2_{X_2} \\ + \\ SS^2_{X_1,X_2} + SS^2_R \end{matrix}$ | $k_1.k_2.r - 1$ | | |

### 1.3.2.2 La règle de décision

Dans ce qui suit, pour tout $j_1 \in \{1, \ldots, k_1\}$ et pour tout $j_2 \in \{1, \ldots, k_2\}$, on définit les moyennes $m_{j_1,j_2}$ de Y sur les sous populations $\{X_1 = x_{1,j_1} et\ X_2 = x_{2,j_2}\}$.

Si les facteurs $X_1$ et $X_2$ n'ont pas d'influence sur les variations de Y alors toutes ces moyennes sont égales entre elles.

Sous l'hypothèse $H_0$ : $\forall(j_1, j_2)\, m_{j_1,j_2} = m$ alors les lois des variables $F_1$, $F_2$ et $F_{1,2}$ sont les lois de Fisher-Snedecor respectivement $F(k_1-1, k_1.k_2.(r-1))$, $F(k_2-1, k_1.k_2.(r-1))$ et $F((k_1-1).(k_2-1), k_1.k_2.(r-1))$ .

Pour un seuil $\alpha$ donné, on fait donc trois tests :

- On teste l'influence de l'interaction des facteurs $X_1$ et $X_2$ sur les variations de Y.
  On cherche dans la table de Fisher-Snedecor avec $((k_1-1).(k_2-1), k_1.k_2.(r-1))$ comme couple de degrés de liberté, la valeur f telle que $P(F \geq f) = \alpha$ et on adopte la règle suivante :
  - Si $F_{1,2} \leq f$ alors on accepte l'hypothèse : l'interaction des facteur $X_1$ et $X_2$ n'explique rien (ou si peu) des variations de Y.
  - Si $F_{1,2} \geq f$ alors on refuse l'hypothèse: les variations de Y sont significativement différentes sur au moins deux couples de modalités des variables $X_1$ et $X_2$.
- On teste l'influence du facteur $X_1$ sur les variations de Y
  On cherche dans la table de Fisher-Snedecor avec $((k_1-1), k_1.k_2.(r-1))$ comme couple de degrés de liberté, la valeur f telle que $P(F \geq f) = \alpha$ et on adopte la règle suivante :
  - Si $F_1 \leq f$ alors on accepte l'hypothèse : le facteur $X_1$ n'explique rien (ou si peu) des variations de Y.
  - Si $F_1 \geq f$ alors on refuse l'hypothèse: les variations de Y sont significativement différentes sur au moins deux modalités de la variable $X_1$.
- On teste l'influence du facteur $X_2$ sur les variations de Y
  On cherche dans la table de Fisher-Snedecor avec $((k_2-1), k_1.k_2.(r-1))$ comme couple de degrés de liberté, la valeur f telle que $P(F \geq f) = \alpha$ et on adopte la règle suivante :
  - Si $F_2 \leq f$ alors on accepte l'hypothèse : le facteur $X_2$ n'explique rien (ou si peu) des variations de Y.
  - Si $F_2 \geq f$ alors on refuse l'hypothèse: les variations de Y sont significativement différentes sur au moins deux modalités de la variable $X_2$.

# 2. Régression linéaire multiple

## 2.1 Le corpus de données

Un ensemble de n individus observés sur une variable quantitative Y et une ou plusieurs variables $X_1$, …, $X_p$ quantitatives.

## 2.2 Quelle question se pose-t-on ?

Existe-t-il un modèle mathématique qui permet d'estimer Y en fonction de $X_1$, …, $X_p$ ?

## 2.3 Comment ?

### 2.3.1 Hypothèses sur le modèle

On pose $Y = a_0 + \sum_{i=1}^{i=p} a_j.X_j + \varepsilon$ .

Pour tout ce qui suit on fait les hypothèses suivantes :

1. Pour chaque individu, la variable $\varepsilon$ suit une loi normale centrée.
2. L'écart-type de la variable $\varepsilon$ est le même pour chaque individu. On parle l'homoscédasticité. On notera $\sigma$ cette valeur commune.
3. D'un individu à un autre individu, les variables $\varepsilon$ sont indépendantes.

Tout ce qui suit est basé sur la véracité de ces hypothèses. C'est pour cela que vous trouverez un dernier paragraphe à la fin du chapitre sur la régression qui donne une méthode pour vérifier à postériori la véracité des hypothèses et comment y remédier si ce n'est pas le cas. Ce paragraphe s'appelle "*analyse des résidus*".

### 2.3.2 Recherche du modèle

On cherche la combinaison affine ( $a_0 + \sum_{i=1}^{i=p} a_j.X_j$ ) des variables $X_j$ qui minimise l''écart entre les valeurs de Y et les valeurs de cette combinaison. Cette combinaison sera le modèle cherché. Dans un deuxième temps, on testera si ce modèle est valable ou pas : il explique une part significative des variations de Y.
Les valeurs trouvées pour les coefficients $a_j$ ne sont réellement que des estimations car les observations portent sur un échantillon. On notera donc ces estimations $\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p$ et $\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_p$ les estimateurs associés .

### 2.3.3 Tester la globalité du modèle

On note $\hat{Y} = \hat{a}_0 + \sum_{i=1}^{i=p} \hat{a}_j.X_j$ le modèle trouvé. On fait une analyse de la variance de la régression. Cela consiste à comparer les deux variances $Var(Y)$ et $Var(\hat{Y})$. La première est la variance des valeurs observées pour Y et la deuxième est la variance des valeurs estimées de Y par le modèle.
On a deux résultats à connaître :

1. $Var(\hat{Y}) \leq Var(Y)$ (d'où le terme régression). Le modèle sous estime les variations des valeurs de Y.
2. Plus $Var(\hat{Y})$ est proche de $Var(Y)$ alors meilleur est le modèle.

Le rapport $\frac{Var(\widehat{Y})}{Var(Y)}$ est appelé coefficient de détermination et il est égal à $R^2$ où $R = \rho(Y, \widehat{Y})$ (coefficient de corrélation linéaire entre $Y$ et $\widehat{Y}$.

#### 2.3.3.1 Ce qu'il faut tester

Pour tester que les variables Xj n'explique rien des variations de Y alors on fait le test suivant :

$H_0 : a_1 = \ldots = a_p = 0$

Si $H_0$ est vraie alors la loi de la variable $F_{obs} = \frac{R^2}{1-R^2}\frac{n-p-1}{p}$ est une loi de Fisher-Snedecor F(p, n-p-1).

#### 2.3.3.2 La règle de décision

Pour un seuil $\alpha$ donné, on cherche dans la table de Fisher-Snedecor la valeur f telle que $P(F \geq f) = \alpha$ et on adopte la règle suivante :

- Si $F_{obs} \leq f$ alors on accepte l'hypothèse $H_0$ : les variables $X_j$ n'explique rien des variations de Y.
- Si $F_{obs} \geq f$ alors on refuse l'hypothèse $H_0$ : au moins une des variables $X_j$ explique une partie significative des variations de Y.

### 2.3.4 Tester chaque coefficient du modèle

Si dans le test de la globalité du modèle on a refusé l'hypothèse $H_0$ alors il faut faire une analyse plus fine en testant la nullité des coefficients $a_j$ ($j \geq 1$) un par un.

#### 2.3.4.1 Ce qu'il faut tester

Pour toute valeur de j, on peut récupérer une estimation $\widehat{\sigma}_j$ de l'écart-type de l'estimateur $\hat{A}_j$. On fait le test $H_0 : a_j = 0$. Si l'hypothèse est vraie alors la loi de la variable $\frac{\hat{A}_j}{\hat{\sigma}_j}$ est une loi de Student à n-p-1 degré de liberté.

#### 2.3.4.2 La règle de décision

Pour un seuil $\alpha$ donné, on cherche dans la table de Student à n-p-1 degré de liberté, la valeur t telle que :

$P(-t \leq T \leq t) = 1 - \alpha$

On note $t_{obs}$ la valeur $\frac{\hat{a}_j}{\hat{\sigma}_j}$.

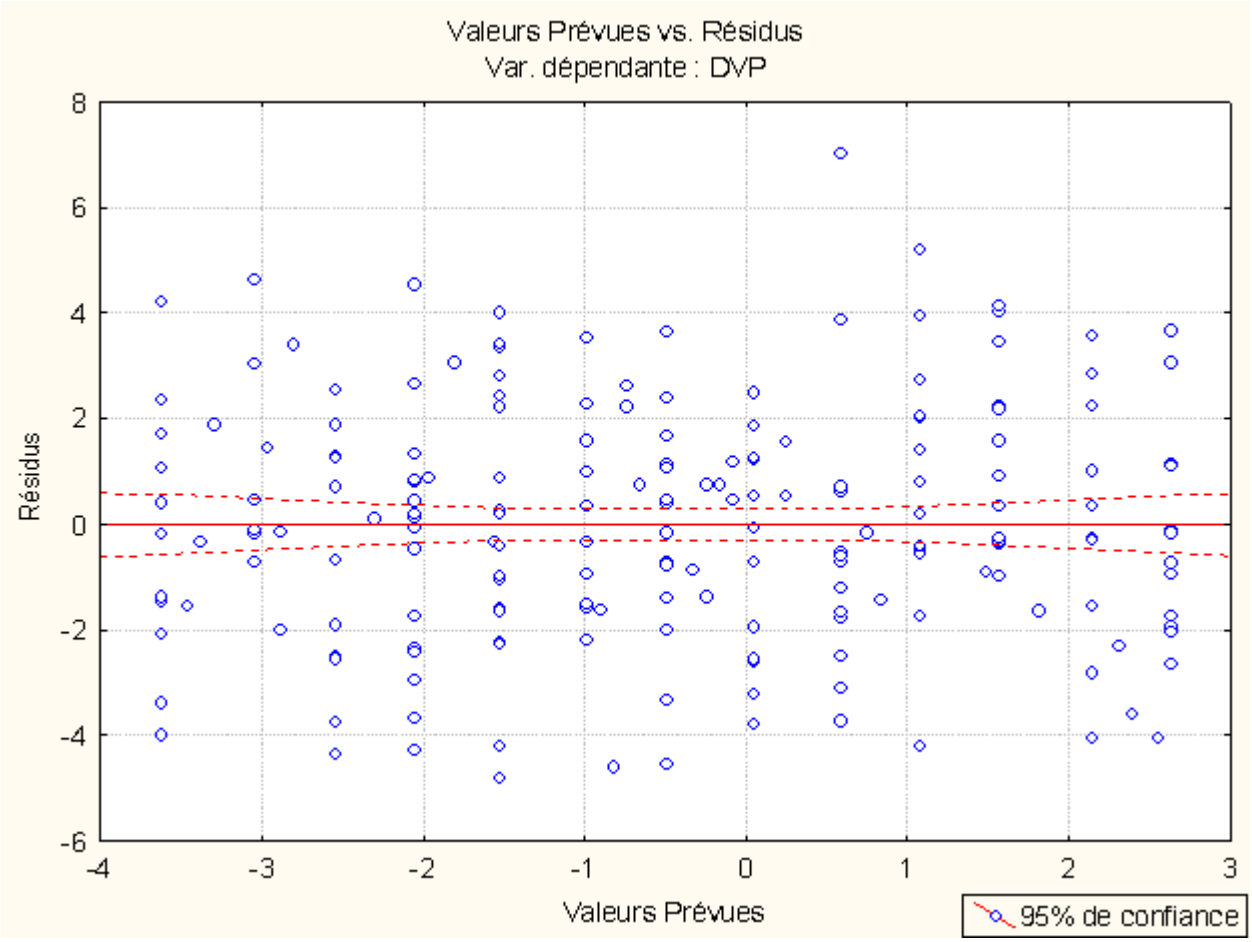
- Si $-t \leq t_{obs} \leq t$ alors on accepte l'hypothèse $H_0$ : la variables $X_j$ n'explique rien des variations de Y.
- Si $t_{obs} \leq -t$ ou $t \leq t_{obs}$ alors on refuse l'hypothèse $H_0$ : la variables $X_j$ explique une partie significative des variations de Y.

### 2.3.5 Analyse des résidus

Il s'agit ici de voir la méthode qui permet de vérifier les hypothèses de base de la régression linéaire.

On note $E = Y - \hat{Y}$. La variable *E* est l'estimateur du résidu du modèle.

Pour valider ces hypothèses de base, il faut vérifier qu'il y a une totale indépendance entre les valeurs estimées $\hat{Y}$ et les résidus estimés $E$. On trace donc le graphe des valeurs estimés (ou prévues) croisées avec les résidus estimés. Les résidus doivent être distribués presque rectangulairement avec une concentration de scores le long du centre. On vous donne ci-dessous un cas qui marche bien.




Dans le cas où ne constate cette distribution aléatoire, il faut essayer de comprendre le phénomène. Sans être exhaustif, on peut citer deux cas de figure classiques :

1. Une ou plusieurs données sont aberrantes (on le voit clairement sur le graphique) et elles contredisent l'hypothèse d'homoscédacité. Pour régler le problème, on retire ces individus de l'échantillon et on refait la régression.
2. Une forme particulière apparaît entre les valeurs estimées et les résidus estimés.
   Dans ce cas il faut :
   a. Soit vérifier (visuellement) si l'une des variables Xj a une relation autre que linéaire avec Y. Si c'est le cas, il faut opérer une transformation adéquate de la variable $X_j$ pour faire apparaître une relation plus linéaire avec Y. Ensuite, on refait la RLM en remplaçant $X_j$ par la transformation.

b. Soit vérifier s'il ne manque pas une variable explicative Z qui aurait une relation non linéaire avec Y. Dans ce cas, il faut intégrer cette variable dans la régression en ajoutant une p+1ème variable qui est une transformation de Z. Cette transformation de Z doit avoir une relation linéaire avec Y.

# 3. Analyse en composantes principales

## *3.1 Le corpus de données*

Un ensemble de n individus observés sur plusieurs variables $X_1$, …, $X_p$ quantitatives.

## *3.2 Quelles questions se pose-t-on ? :*

- Quelles sont les différences les plus significatives entre les individus observés sur ces variables.
- Quelles sont les significations de ces différences ?
- Pour chaque individu, quelles sont ses différences par rapport à tous les autres individus ?
- Une question sous jacente : peut-on découvrir des groupes d'individus homogènes (qui se ressemblent) ?

## *3.3 Comment ?*

### 3.3.1 Centrage et réduction des données

- On calcule pour chaque variable $X_j$, la moyenne $m_j$ et l'écart-type $\sigma_j$ de cette variable. Chaque valeur $x_{i,j}$ du corpus de données est changée avec la transformation $x_{i,j} = \frac{x_{i,j} - m_j}{\sigma_j}$. Cette transformation permet de centrer le nuage de points formé par les observations dans l'espace $R^p$ (centrage des données) et de ramener toutes les variables dans la même échelle (réduction des données).
- Après cette transformation, pour mesurer les différences entre les individus, on munit $R^p$ de la distance euclidienne classique : $d^2(x_i,\vec{0}) = \sum_{j=1}^{p} x_{i,j}^2$

### 3.3.2 Recherche des axes factoriels et des composantes principales

Le principe général consiste à projeter le nuage de points dans des espaces de faible dimension en maximisant la dispersion du nuage projeté.

Le premier axe consiste à chercher une droite vectorielle $\Delta_{u_1}$ telle que :

$\sum_{i=1}^{p} d^2(pr_{\Delta_{u_1}}(x_i),0) = \underset{u/d^2(u)=1}{Min} \sum_{i=1}^{p} d^2(pr_{\Delta_u}(x_i),0)$ où $pr_{\Delta_{u_1}}(x_i)$ est la projection orthogonale de $x_i$ sur $\Delta_{u_1}$

Le deuxième axe consiste à chercher une droite vectorielle $\Delta_{u_2}$ qui ramène le plus de dispersion possible du nuage de points non ramenée par le 1er axe.

$$\sum_{i=1}^{p} d^2(pr_{\Delta_{u_2}}(x_i),0) = \underset{\substack{u/d^2(u)=1 \\ u \perp u_1 = 0}}{Min} \sum_{i=1}^{p} d^2(pr_{\Delta_u}(x_i),0)$$

Le j$^{ème}$ axe est construit sur le même principe. Parmi les droites vectorielles orthogonales aux j-1 premiers axes, on cherche celui qui ramène le plus dispersion possible du nuage projeté.

A chaque axe $\Delta_{u_j}$, on associe la j$^{ème}$ composante principale CP$^j$. CP$^j$ est vecteur de R$^n$ donc la i$^{ème}$ coordonnée est définie comme suit : $CP_i^j = pr_{\Delta_{u_j}}(x_i)$

### 3.3.3 Interprétation des résultats

#### 3.3.3.1 Récupération de la dispersion

La dispersion totale du nuage est donnée par : $\frac{1}{n}.\sum_{i=1}^{n} d^2\left(x_i,\vec{0}\right)$. Chaque axe en récupère une partie de cette dispersion et on a la formule suivante : $\frac{1}{n}.\sum_{i=1}^{n} d^2\left(x_i,\vec{0}\right) = \sum_{j=1}^{p}\left(\frac{1}{n}.\sum_{i=1}^{n} d^2\left(pr_{\Delta_{u_j}}(x_i),\vec{0}\right)\right)$. On peut faire la somme des dispersions car les axes sont orthogonaux et récupèrent donc chacun une partie de la dispersion que n'est récupérée par aucun autre axe.

Pour chaque axe, on peut donc calculer le rapport $\frac{\sum_{i=1}^{n} d^2\left(pr_{\Delta_{u_j}}(x_i),\vec{0}\right)}{\sum_{i=1}^{n} d^2\left(x_i,\vec{0}\right)}$ qui représente le pourcentage de dispersion récupérée par le j$^{ème}$ axe.

Ces différents pourcentages sont les premières valeurs à étudier pour l'interprétation des résultats. On peut aussi pour les j premiers axes, étudier le rapport $\frac{\sum_{k=1}^{j}\sum_{i=1}^{n} d^2\left(pr_{\Delta_{u_k}}(x_i),\vec{0}\right)}{\sum_{i=1}^{n} d^2\left(x_i,\vec{0}\right)}$. Ce rapport représente le pourcentage de dispersion récupérée par les j premiers axes.

Pour j donné, la valeur $\frac{1}{n}\sum_{i=1}^{n} d^2\left(pr_{\Delta_{u_j}}(x_i),\vec{0}\right)$ correspond à une valeur propre d'une certaine matrice. C'est pour cela que les logiciels parlent de valeur propre dans l'affichage des résultats. En général, les logiciels présentent ces informations comme suit :

| Axe | Valeur propre | Dispersion en % ou information en % | Cumul Disp. Ou cumul info en % |
|---|---|---|---|
| N° 1 | ??? | ??? | ??? |
| … | ??? | ??? | ??? |
| N° p | ??? | ??? | ??? |

#### 3.3.3.2 Interprétation des axes

D'un point de vue mathématiques, un axe factoriel est une combinaison linéaire des variables $X_1$, …, $X_p$. Cela ne permet pas donner une signification à l'axe. En revanche, l'axe $\Delta_{u_j}$ est associé à la composante principale $CP_j$. Une composante principale est une nouvelle variable qui est une synthèse des variables $X_1$,

…, $X_p$. Pour interpréter les axes, on récupère les coefficients de corrélation linéaires entre les variables et les composantes principales.
Pour chaque composante principale $Cp_k$, on cherche les variables qui sont les plus fortement corrélées en valeur absolues. On sépare ces variables en deux groupes : celles qui sont corrélées négativement avec la composante d'un côté et celles qui sont corrélées positivement. Donner un sens à l'axe consiste à trouver la signification de l'opposition de ces deux groupes de variables.
Une première signification triviale consiste à dire que :

- Tout individu placé sur la partie négative de l'axe aura tendance à avoir de plus fortes valeurs sur les variables corrélées négativement avec l'axe par rapport à la moyenne des individus (représentée par l'origine) et aura tendance à avoir de plus faibles valeurs sur les variables corrélées positivement avec l'axe par rapport à la moyenne des individus (représentée par l'origine).
- Tout individu placé sur la partie positive de l'axe aura tendance à avoir de plus faibles valeurs sur les variables corrélées négativement avec l'axe par rapport à la moyenne des individus (représentée par l'origine) et aura tendance à avoir de plus fortes valeurs sur les variables corrélées positivement avec l'axe par rapport à la moyenne des individus (représentée par l'origine).

Dans certain cas, l'un des deux groupes est vide. Si, par exemple, il n'y a que des variables fortement corrélées positivement avec l'axe alors cela signifie :

- que les individus placés sur la partie positive de l'axe auront tendance à avoir de plus fortes valeurs sur ces variables par rapport à la moyenne des individus (représentée par l'origine).
- que les individus placés sur la partie négative de l'axe auront tendance à avoir de plus faibles valeurs sur ces variables par rapport à la moyenne des individus (représentée par l'origine).

#### 3.3.3.3 Caractérisation des individus

Pour caractériser un individu par rapport à la moyenne des tous les individus, il faut regarder sa position sur les axes sur lesquels il est bien représenté. Il faut donc récupérer pour chaque individu i, sa qualité de représentation (souvent appelé $\cos^2$) sur chaque axe $\Delta_{u_j}$.

Pour chaque individu, la caractérisation se fait en deux temps :

1. Chercher les axes sur lesquels l'individu a une forte qualité de représentation
2. Pour chacun de ces axes, regarder sa position sur l'axe et conclure en fonction de la signification de l'axe (voir paragraphe Interprétation des axes)

#### 3.3.3.4 Contributions des individus et individus aberrants

Un axe doit traduire une tendance générale et non pas la caractérisation d'un tout petit nombre d'individus. Pour vérifier cette éventuelle aberration, on vous fournit pour chaque individu et pour chaque axe, la contribution de l'individu à l'axe. Pour un même axe, les contributions doivent être à peu près égales entre tous les individus pour traduire une tendance générale. Si pour un axe, un petit ensemble d'individus ont des contributions nettement plus grande que la grande majorité des individus, il faut probablement penser à faire les deux choses suivantes :

1. Noter cette caractéristique pour ces quelques individus
2. Refaire l'analyse en supprimant ces individus ou en leur affectant un poids de 0 (individus supplémentaires)

### 3.3.3.5 Individus : coordonnées, cos² et contributions

En général sur les logiciels, les coordonnées, les cos² et les contributions sont présentés comme suit :

| | Axe n° 1 | … | Axe n° j | … | Axe n° p |
|---|---|---|---|---|---|
| Individu n° 1 | Coord<br>Cos²<br>contrib | | Coord<br>Cos²<br>contrib | … | Coord<br>Cos²<br>contrib |
| … | … | … | … | | |
| Individu n° i | Coord<br>cos²<br>contrib | | Coord<br>cos²<br>contrib | … | Coord<br>cos²<br>contrib |
| … | | … | | | |
| Individu n° n | Coord<br>cos²<br>contrib | | Coord<br>cos²<br>contrib | … | Coord<br>cos²<br>contrib |

# 4. Analyse factorielle des correspondances

## *4.1 Le corpus de données*

Un tableau d'effectifs qui est le croisement de 2 variables qualitatives X et Y ayant respectivement k et l modalités. Le contenu $n_{i,j}$ d'une case du tableau donne le nombre d'individus qui ont simultanément la i$^{ème}$ modalité de X et la j$^{ème}$ modalité de Y. Le tableau ci-dessous présente la forme du corpus de données.

| Y<br>X | Mod n° 1 | … | Mod n° j | … | Mod n° l | Effectifs de X |
|---|---|---|---|---|---|---|
| Mod n° 1 | $n_{1,1}$ | … | $n_{1,j}$ | … | $n_{1,l}$ | $n_{1,.} = \sum_{j=1}^{l} n_{1,j}$ |
| … | … | … | … | … | … | … |
| Mod n° i | $n_{i,1}$ | … | $n_{i,j}$ | … | $n_{i,l}$ | $n_{i,.} = \sum_{j=1}^{l} n_{i,j}$ |
| … | … | … | … | … | … | … |
| Mod n° k | $n_{k,1}$ | … | $n_{k,j}$ | … | $n_{k,l}$ | $n_{k,.} = \sum_{j=1}^{l} n_{k,j}$ |
| Effectifs de Y | $n_{.,1} = \sum_{i=1}^{k} n_{i,1}$ | … | $n_{.,j} = \sum_{i=1}^{k} n_{i,j}$ | … | $n_{.,l} = \sum_{i=1}^{k} n_{i,l}$ | $n = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{i,j}$ |

## *4.2 Quelles questions se pose-t-on ?*

- Y-a-t-il un lien entre les deux variables qualitatives ?
- Si oui, quelles sont les correspondances entre les modalités de la 1$^{ère}$ variable et les modalités de la 2$^{ème}$ variable ?

## *4.3 Comment ?*

Le tableau d'effectif n'est pas directement utilisable car deux modalités d'une même variable peuvent avoir des effectifs très différents et en conséquence une grande valeur d'un $n_{i,j}$ peut être dû au fait que la modalité i de la première variable a elle-même un grand effectif et non pas dû à une spécificité du croisement des 2 variables.

### 4.3.1 Les profils lignes (ou profils colonnes)

Pour pallier, on va étudier ce qu'on appelle des profils.
A la $i^{ème}$ modalité de la variable X, on associe un profil ligne définie par :

$$PL_i = \left( \frac{n_{i,1}}{n_{i,.}}, ..., \frac{n_{i,j}}{n_{i,.}}, ..., \frac{n_{i,l}}{n_{i,.}} \right).$$

Un profil ligne est donc un vecteur de $R^l$. Les $j^{ème}$ coordonnées de deux profils ligne sont maintenant directement comparables. En fait, on comparera chaque profil ligne au profil moyen définie par :

$$PLM = \left( \frac{n_{.,1}}{n}, ..., \frac{n_{.,j}}{n}, ..., \frac{n_{.,l}}{n} \right)$$

Pour un profil ligne $Pl_i$ donné et un indice j, un écart significatif entre la $j^{ème}$ coordonnée du profil $Pl_i$ et la $j^{ème}$ coordonnée du profil moyen PLM traduira une spécificité du profil par rapport à l'ensemble des autres profils représentés par le profil moyen PLM. On pourra donc établir par ce biais les correspondances énoncées dans les questions que l'on se pose.
On peut faire l'étude duale sur les profils colonnes. On choisit l'un des deux types de profils suivant que l'on désire expliquer X en fonction de Y ou expliquer Y en fonction de X. Quelle est la cause ? Quel est l'effet ?

Pour la suite de ce résumé, on choisira de façon arbitraire d'étudier les profils lignes. Mais les résultats sont parfaitement symétriques.

### 4.3.2 La distance du chi-2 pour mesurer les différences

Comme pour l'ACP, il faut définir une distance pour mesurer les différences entre chaque profil et le profil moyen. Elle est définit comme suit : $d^2(PL_i, PLM) = \sum_{j=1}^{l} \left( \left( PL_{i,j} - PLM_j \right)^2 \Big/ PLM_j \right)$. Elle est appelée distance du khi-2. En effet, si on fait la somme de ces distances de tous les profils, on obtient au coefficient n (nombre d'individus) près, la valeur utilisée dans le test du chi-2 d'indépendance de deux variables qualitatives. Cette distance a deux propriétés intéressantes :

- Elle donne plus d'importance aux écarts constatés sur des modalités de faibles effectifs.
- Si dans le tableau initial (tableau des effectifs), deux colonnes sont proportionnelles alors la fusion de ces deux colonnes ne change pas la valeur des distances entre les profils et le profil ligne moyen. C'est le principe d'équivalence distributionnelle.

### 4.3.3 Interprétation des résultats

#### 4.3.3.1 Récupération de la dispersion

Le principe est exactement le même qu'avec l'ACP. A ceci près qu'en ACP, on peut avoir jusqu'à p axes alors qu'en AFC, on ne peut avoir au maximum min(k-1,l-1). Sans rentrer dans les détails mais ce résultats est du à deux choses :

- tous les profils sont liés par une même relation : la somme de leurs coordonnées est égale à 1.
- l'étude des profils lignes est duale de l'étude des profils colonnes.

#### 4.3.3.2 Interprétation des axes

Contrairement à l'ACP, il n'y a pas de coefficient de corrélation linéaire entre les composantes principales et les variables car X et Y sont qualitatives et dans ce cas le concept de corrélation linéaire n'a aucun sens. En revanche, on peut récupérer ce qu'on appelle les contributions partielles des colonnes aux axes. Elles joueront le même rôle que les coefficients de corrélations linéaires. On devra pour un axe ne retenir que les plus fortes contributions. Pour savoir, parmi les colonnes retenues, si elle contribue négativement ou positivement sur l'axe, on se reportera à la position de la colonne sur l'axe.

#### 4.3.3.3 Caractérisation des individus

Pour caractériser un profil ligne par rapport au profil ligne moyen, il faut regarder sa position sur les axes sur lesquels il est bien représenté. Il faut donc récupérer pour chaque profil ligne i, sa qualité de représentation (souvent appelé $\cos^2$) sur chaque axe $\Delta_{u_j}$. C'est donc le même principe que pour l'ACP.

Pour chaque profil ligne, la caractérisation se fait en deux temps :

3. Chercher les axes sur lesquels le profil ligne a une forte qualité de représentation
4. Pour chacun de ces axes, regarder sa position sur l'axe et conclure en fonction de la signification de l'axe (voir paragraphe Interprétation des axes)

## 5. Etude simultanée : ACP et AFC

Normalement, l'ACP ne concerne que les variables quantitatives et l'AFC ne concerne que les variables qualitatives (et donc des effectifs).
Cependant, on peut déroger à cette règle. Si le corpus de données est un ensemble de n individus observés sur plusieurs variables $X_1$, …, $X_p$ quantitatives et positives alors on peut bien sûr faire une ACP mais aussi faire une AFC en assimilant chaque individu à un profil ligne.

ATTENTION à l'interprétation : la constitution d'un profil ligne consiste à mélanger les valeurs des différentes variables alors que pour une ACP les variables ne sont pas mélangées car considérées comme étant de nature différente (elles ne sont en général pas exprimées dans la même unité).
En particulier, si pour deux individus, les 2 lignes des valeurs des variables sont proportionnelles alors dans l'AFC, les deux individus seront confondus et dans l'ACP, ils seront d'autant plus éloignés que le rapport de proportion sera très différent de 1.
En résumé, il est important pour bien interpréter de comprendre comment dans les deux analyses , on a constitué les points individus (ACP) ou profils lignes (AFC)