Lien entre une variable et un ensemble de variables

Prof. Mustapha RACHDI

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau : C08 au Bât. Michel Dubois
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

Introduction

Les méthodes d'analyse des données prolongent et complètent les méthodes *classiques* de statistique descriptive.

Objectifs :

- rappeler un certain nombre de définitions et de concepts de statistique descriptive,
- montrer qu'une première description des données peut être obtenue grâce à des traitements élémentaires,
- proposer un rapide panorama des indicateurs calculés en statistique descriptive, en particulier ceux qui permettent de mesurer la liaison entre une variable et un ensemble de variables.

Différents types de variables : `str`

Une variable statistique décrit une caractéristique pour les différents individus pour lesquels elle est définie.

L'ensemble de ces individus constitue une population : l'ensemble des français, les branches de l'industrie allemande, ou encore les différentes régions de l'Europe.

On distingue deux types de variables (*caractères*) : les variables quantitatives (le poids d'un individu, le montant de son patrimoine, le volume du réservoir de son automobile) et les variables qualitatives (la couleur des yeux, le diplôme possédé ou encore la catégorie socioprofessionnelle).

Les variables quantitatives (variables *numériques*) :

- discrètes : elles prennent un nombre fini de valeurs, comme le nombre d'enfants d'une famille ou le nombre de fonctionnaires d'un ministère, ...
- continues : elles peuvent prendre toutes les valeurs intermédiaires, comme le poids ou la taille, ...

Description d'une variable quantitative : `summary`

Une variable quantitative est donc décrite par les valeurs qu'elle prend pour les n individus pour lesquels elle est définie.

Exemple

La variable taille*, pour une population de* $n = 4$ *individus, est décrite par le tableau suivant :*

Individu	*1*	*2*	*3*	*4*
Taille	*1.70*	*1.65*	*1.70*	*1.80*

Description d'une variable quantitative

Plus généralement, considérons une variable x, on note x_i la valeur prise par l'individu i pour $i = 1, \ldots, n$.
Pour synthétiser l'information donnée par cette variable quantitative, les deux indicateurs les plus fréquemment calculés sont :

- la moyenne de la variable, notée $\bar{x}$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  La moyenne est un indicateur de tendance centrale, reflétant l'ordre de grandeur de la variable.

- la variance de la variable, notée $var(x)$:

$$var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ ou la variance corrigée } \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

  La variance est un indicateur de dispersion, reflétant l'importance des fluctuations des valeurs de la série (*les* x_i) autour de leur valeur moyenne. Souvent, plutôt que de calculer $var(x)$, on préfère calculer l'écart-type de la variable x, noté σ_x. La valeur σ_x est la racine carrée positive de $var(x)$.

Description d'une variable quantitative : `mean, var, scale, median, ...`

Exemple

Dans l'exemple ci-dessus, la moyenne est 1.7125 *et la variance vaut* 0.0029*.*

Les propriétés essentielles de la moyenne et de la variance : sont les suivantes :

Propriétés

Pour tous réels a, b, on a

- $\overline{ax+b} = a\bar{x} + b$.
- $var(ax+b) = a^2\, var(x)$.

Remarque

Ces propriétés se déduisent immédiatement des définitions de la moyenne et de la variance.

Proposition

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \text{ : est une variable centrée-réduite.}$$

Description d'une variable qualitative

Tableau disjonctif complet : `tab.disjonctif` ou `tab.disjonctif.prop` `FactoMineR`

Comme une variable quantitative, une variable qualitative est décrite par les valeurs prises par les n individus pour lesquels elle est définie.

Exemple

Considérons la variable couleur des yeux *et* 6 *individus.*

Individu	*1*	*2*	*3*	*4*	*5*	*6*
Couleur des yeux	*Bleu*	*Vert*	*Marron*	*Vert*	*Bleu*	*Bleu*

La variable couleur des yeux *comporte* 3 *modalités : bleau, vert et marron. Pour chaque modalité, on peut calculer sa fréquence absolue N, c'est-à-dire le nombre d'individus qui possèdent cette modalité :*

$$N(bleu) = 3, \quad N(vert) = 2, \quad N(marron) = 1.$$

On appelle fréquence relative (notée f), le rapport de la féquence absolue au nombre total d'individus : $f(bleu) = \frac{3}{6}, \quad f(vert) = \frac{2}{6}, \quad f(marron) = \frac{1}{6}.$

Description d'une variable qualitative
Tableau disjonctif complet

Remarque

La présentation d'une variable qualitative sous sa forme disjonctive complète est celle qui se prête le mieux à des calculs statistiques. La forme disjonctive complète est obtenue en définissant une variable indicatrice pour chacune des modalités. Par exemple, si M est la variable indicatrice de la modalité marron*, pour l'individu i :*

$$M = \begin{cases} 1 & \textit{si l'individu} \text{ i } \textit{a les yeux marrons} \\ 0 & \textit{sinon} \end{cases}$$

On obtient donc pour la variable couleur des yeux *le tableau disjonctif complet X suivant, la ligne i correspondant à l'individu i :*

$$X = \begin{pmatrix} 100 \\ 010 \\ 001 \\ 010 \\ 100 \\ 100 \end{pmatrix}$$

Les 3 *colonnes de X sont constituées respectivement par des modalités* bleu*,* vert *et* marron*.*

Proposition

La somme des colonnes d'un tableau disjonctif complet X est égale au vecteur $\mathbb{I}_n$ de dimension n, dont tous les éléments sont égaux à 1.

Proposition

${}^tX\,X$ est une matrice diagonale dont les éléments sont les fréquences absolues des modalités.

Exemple

En reprenant l'exemple sur la couleur des yeux, on obtient

$$ {}^tX\,X = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} $$

Description d'une variable qualitative
Codage d'une variable qualitative

Quantifier ou coder une variable qualitative, c'est associer à chacune de ses modalités un nombre réel et ainsi transformer la variable qualitative en une variable quantitative.
Ainsi, si pour la variable *couleur des yeux*, on code a_1 la couleur *bleue*, a_2 la couleur *verte* et a_3 la couleur *marron*, on obtient la variable quantitative suivante définie pour les 6 individus.

Exemple

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_2 \\ a_1 \\ a_1 \end{pmatrix} = \begin{pmatrix} 100 \\ 010 \\ 001 \\ 010 \\ 100 \\ 100 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = X \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

L'ensemble des coefficients possibles est obtenu en faisant varier a_1, a_2 et a_3. Donc, l'ensemble des codifications possibles pour les individus, est l'ensemble des combinaisons linéaires des colonnes du tableau disjonctif complet X.

Le modèle linéaire : selon le type de variables explicatives

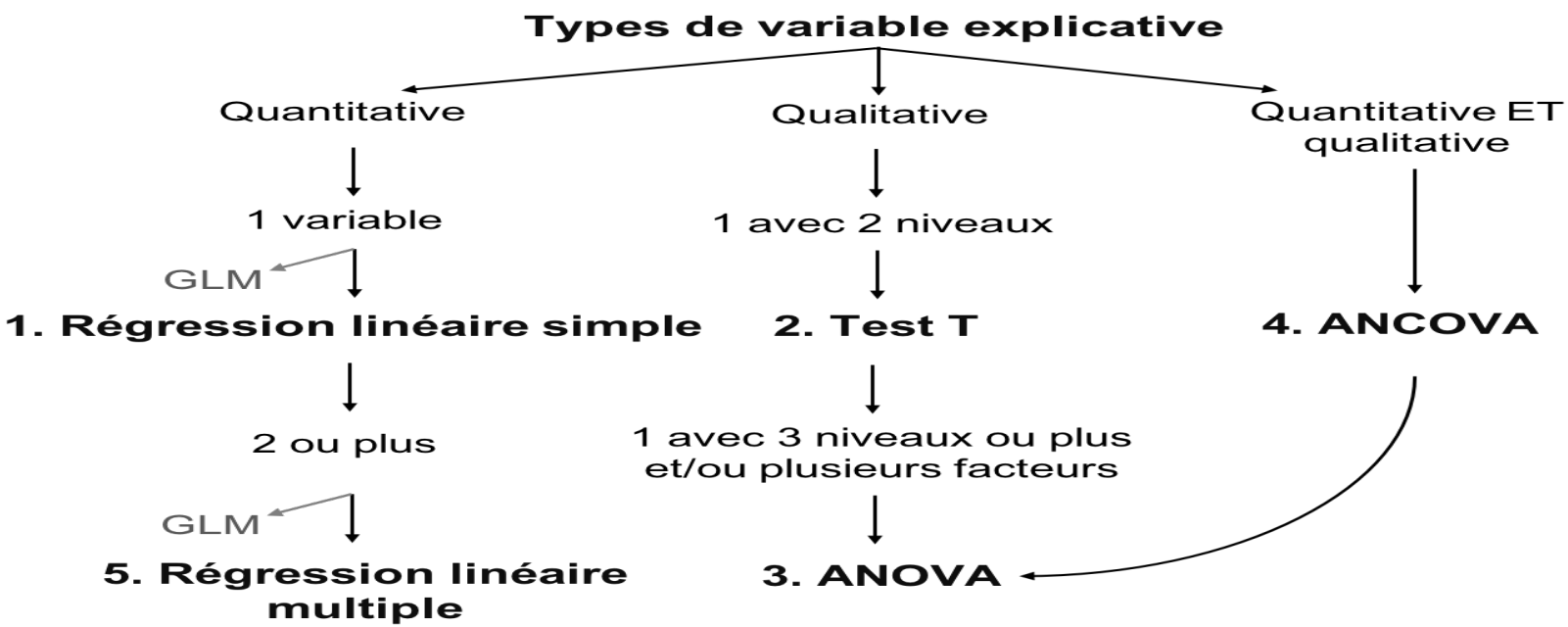


Figure – Modèle linéaire selon le type de la variable explicative

Relation entre deux variables quantitatives : la régression simple : `lm`

Le problème

Le poids et la taille sont deux variables qui varient généralement ensemble et dans le même sens : les individus les plus lourds sont souvent les plus grands, et les plus petits les plus légers.
Le lien entre les deux variables n'est cependant pas un lien absolu : certains individus petits sont plus lourds que d'autres plus grands.

Comment mesurer l'intensité de la relation entre deux variables ?

On note x et y les deux variables. Soit x_i (resp. y_i) la valeur prise par l'individu i pour la variable x (resp. y). Il s'agit de déterminer, s'elle existe, une relation linéaire vérifiée même approximativement par les 2 variables, c'est-à-dire s'il existe deux réels a et b tels que

$$y_i = ax_i + b + e_i \text{ pour } i = 1, \ldots, n$$

où e_i est un terme résiduel.

Conditions d'utilisation du modèle linéaire

- **adéquation** : nuage de points $(\widehat{y}, e)$ dans lequel les résidus ne doivent présenter aucune propriété intéressante
- **homoscédasticité** : nuage de points $(\widehat{y}, e)$ (transformation possible des données pour stabiliser la variance)
- **indépendance** des erreurs résiduelles :
  - hypothèse fondamentale du modèle linéaire
  - graphe (i, e_i) l'ordre des résidus ne doit pas avoir de sens
- **normalité** des erreurs résiduelles :
  - hypothèse la moins importante
  - normalité à vérifier pour petits échantillons (quelques dizaines)
  - tests de normalité (Kolmogorov-smirnov, shapiro-wilks,...) déconseillés car sensibles à la non indépendance
  - histogramme des résidus, QQ plot (quantiles empiriques des résidus e_i (ou normalisés) en fonction des quantiles de la gaussienne)
- **graphiques partiels** : tracé des p nuage de points (x_k, e) pour chaque variable explicative pour traquer les dépendances entre résidus et variables explicatives et détecter les points atypiques et/ou influents.

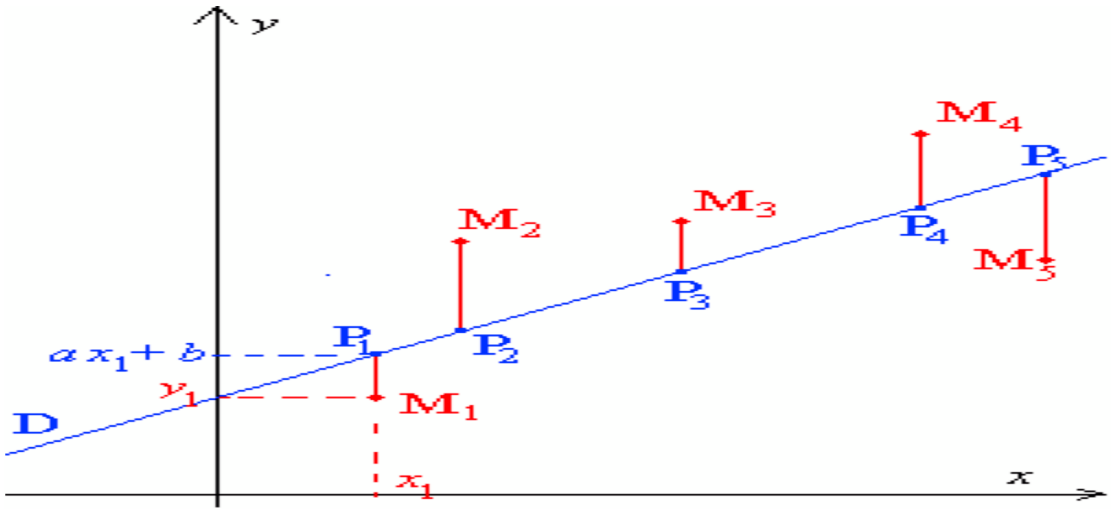
La régression : Résolution algébrique

La relation entre y et x sera d'autant plus proche d'une relation linéaire exacte que les valeurs de la série e, c'est-à-dire les valeurs de e_i seront petites.

Algèbriquement, on détermine a et b selon le critère des moindres carrés., c'est-à-dire de telle manière que

$$\sum_{i=1}^{n} e_i^2 \text{ ait une valeur minimale.}$$

Graphiquement, chaque individu est représenté par un point de coordonnées x_i et y_i dans un repère d'axes x et y, et on recherche la droite qui passe *au plus près* du nuage de points.
L'écart e_i est l'écart entre la droite et la valeur observée de y, soit y_i



Proposition

La droite de régression des moindres carrés $y = ax + b + e$, est telle que :

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n\, var(x)} \ , \ b = \bar{y} - a\bar{x} \ et \ \sum_{i=1}^{n} e_i = 0.$$

Proposition

Par définition, la covariance entre 2 variables x et y est :

$$cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \ \ et\ donc \ \ a = \frac{cov(x, y)}{var(x)}.$$

Equation d'analyse de la variance

$$y_i = ax_i + b + e_i$$

permet de comprendre pourquoi la variable y_i prend des valeurs différentes d'un individu à l'autre : la valeur prise pour la variable y par l'individu i dépend de la valeur que cet individu prend pour la variable x. Les différents individus prennent des valeurs différentes pour cette variable x, et donc des valeurs différentes pour y.

Exemple

Reprenons l'exemple (poids, tailles), il y a des individus lourds et des individus plus légers parce que le poids est lié à la taille et certains individus ont une grande taille et d'autres une taille moins importante.

L'importance des fluctuations de la variable y est mesurée par var(y). Les fluctuations de y occasionnées par la variable x sont mesurées par var$(ax + b)$, et les fluctuations de y ne dépendant pas de x sont mesurées par var(e).

Equation d'analyse de la variance

Proposition

$$var(y) = var(ax + b) + var(e)$$

où $var(y)$ est la variance totale de y, $var(ax + b)$ est la variance expliquée par les variations de x et $var(e)$ est la variance des résidus.

Définition

$$R^2(x, y) = \frac{cov^2(x, y)}{var(x)\, var(y)} \text{ et } R(x, y) = \frac{cov(x, y)}{\sigma_x\, \sigma_y}.$$

De l'équation de l'analyse de la variance, il découle directement que le coefficient de détermination est compris entre 0 et 1, et donc que le coefficient de corrélation prend ses valeurs comprises entre -1 et 1.

Interprétation géométrique du coeff. de corrélation

Une variable x prenant n valeurs peut être représentée par un vecteur dans $\mathbb{R}^n$.
L'ensemble $\mathbb{R}^n$ est appelé *espace des variables*.
Dans $\mathbb{R}^n$, le produit scalaire usuel entre deux vecteurs x et y de coordonnées respectives $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$ est :

$$< x, y >= \sum_{i=1}^{n} x_i \, y_i.$$

En statistique, le produit scalaire utilisé dans l'espace des variables $\mathbb{R}^n$ est le produit scalaire suivant :

$$< x, y >= \frac{1}{n} \sum_{i=1}^{n} x_i \, y_i.$$

En effet, ce produit scalaire particulier permet de donner au coefficient de corrélation une interprétation géométrique simple.

Proposition

Dans l'espace $\mathbb{R}^n$, le cosinus de l'angle entre 2 variables centrées est égal au coefficient de corrélation entre ces variables.

Interprétation géométrique du coefficient de corrélation

Remarque

Si le coefficient de corrélation est égal à 1, les 2 vecteurs sont colinéaires, c'est-à-dire que les valeurs prises par x_i et y_i sont proportionnelles, donc qu'il existe une relation linéaire exacte entre les 2 variables. L'absence de corrélation se traduit par une valeur nulle pour R, et donc un angle droit entre x et y.

Proposition

Dans l'espace $\mathbb{R}^n$, la norme d'une variable centrée est égale à son écart-type.

Dans $\mathbb{R}^n$ donc, une variable centrée est normée si son écart-type (et, par conséquent, sa variance) est égal à 1. De la même manière, on montre sans difficulté que :

Proposition

Dans l'espace $\mathbb{R}^n$, la norme d'une modalité d'une variable qualitative est égale à la racine carrée de la fréquence relative de cette modalité.

Interprétation géométrique du coefficient de corrélation

Le coefficient de corrélation linéaire

- La plus ou moins grande dépendance entre les deux variables x et y peut être appréhendée par la valeur de l'angle que forment les droites de régression D et D'.
- Plus cet angle est ouvert, moins la liaison est forte.

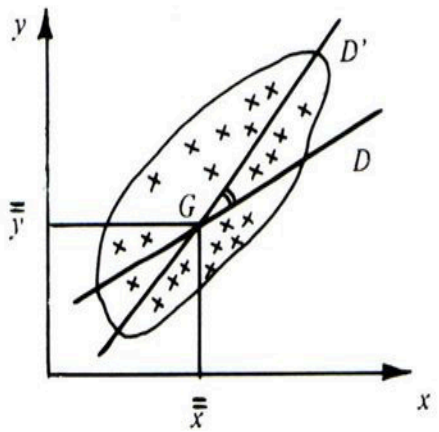


Figure – Le coefficient de corrélation dans $\mathbb{R}^n$

Relation entre une variable quantitative expliquée et un ensemble de variables quantitatives explicatives : la régression multiple

Position du problème

On généralise à plusieurs variables explicatives le modèle de régression simple : la relation linéaire entre Y et les variables explicatives X_j est vérifiée à un terme résiduel près, et pour l'individu i, on peut écrire :

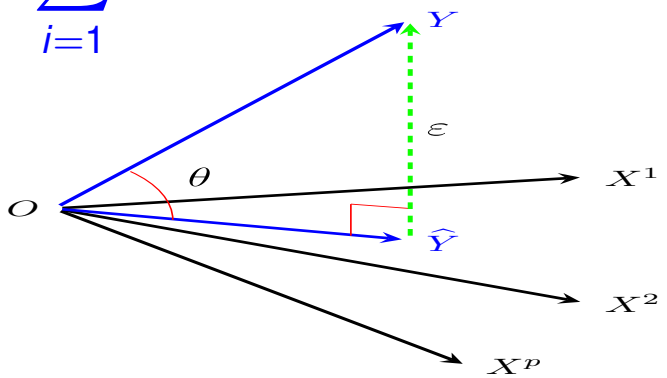
$$y_i = a_0 + a_1 x_{i1} + \cdots + a_j x_{ij} + \cdots + a_p x_{ip} + e_i$$

où y_i est l'observation i de la variable Y et x_{ij} est l'observation i de la variable X_j.

Relation entre une variable quantitative expliquée et un ensemble de variables quantitatives explicatives : la régression multiple

Les coefficients a_j (pour $j = 1, \ldots, p$) sont déterminés selon le critère des moindres carrées, c'est-à-dire de telle manière que

$$\sum_{i=1}^{n} e_i^2 \text{ ait une valeur minimale.}$$



Géométriquement, la régression est la projection $\widehat{Y}$ de Y sur l'espace vectoriel $vect\{1, X_1, \ldots, X_p\}$. De plus $R^2 = \cos^2(\theta)$.

Comme dans le cas de la régression simple, on montre facilement

$$\sum_{i=1}^{n} e_i = 0.$$

Relation entre une variable quantitative expliquée et un ensemble de variables quantitatives explicatives : la régression multiple

Position du problème

Par conséquent, le problème s'écrit matriciellement sous la forme

$$Y = X\,a + e$$

avec

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, \; a = \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ \vdots \\ a_p \end{pmatrix} \text{ et } \; e = \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}$$

Y est le vecteur-colonne des observations de la variable à expliquer.
X est la matrice de dimensions $n \times p$ des n observations de chacune des variables $X_j,\ j = 1, \dots, p$. X_j désigne la colonne j de la matrice X.

Interprétation dans l'espace des individus

L'espace des individus est l'espace dont les points sont des individus, c'est-à-dire des observations. Les coordonnées d'un point dans cet espace sont données par une ligne du tableau X et par la valeur de Y correspondante : on dispose par conséquent de n points dans un espace de dimension $(p+1)$.

Minimiser la somme des carrés des résidus revient donc à chercher un hyperplan d'équation :

$$Y = a_1 X_1 + \cdots + a_j X_j + \cdots + a_p X_p$$

de telle manière que cet hyperplan passe *au plus près* du nuage constitué par les n points.

Interprétation dans l'espace des variables

Dans l'espace des variables $\mathbb{R}^n$, on peut donner une autre interprétation géométrique du problème de la régression. L'espace des variables est un espace de dimension n : on représente dans cet espace la variable Y et chacune des variables X_j, $j = 1, \ldots, p$.

Les coordonnées de la variable à expliquer sont données par le vecteur Y, et les coordonnées des variables explicatives X_j sont données par les p colonnes de la matrice X. On a également e est un vecteur de $\mathbb{R}^n$ et on montre que :

Proposition

Les paramètres $a_1, \ldots, a_p$ tels que $\sum_{i=1}^{n} e_i^2$ soit minimal, sont donnés par la projection orthogonale de Y sur l'espace engendré par les vecteurs X_j, $j = 1, \ldots, p$.

Estimation du modèle

Solution

L'interprétation géométrique dans l'espace des variables permet de calculer le vecteur-colonne des paramètres a.
Le vecteur $(X\,a)$ est la projection orthogonale de Y sur l'espace engendré par les vecteurs X_j, $j = 1, \ldots, p$, c'est-à-dire que

$$X\,a = PY$$

où P est le projecteur orthogonal sur l'espace engendré par les colonnes de X, c'est-à-dire l'application qui à un vecteur fait correspondre sa projection orthogonale sur cet espace.
Par conséquent

$$X\,a = X\,({}^tX\,X)^{-1}\,{}^tX\,Y. \qquad \text{(cf. TD)}$$

Soit finalement, comme les colonnes de X sont linéairement indépendantes

Proposition

Le vecteur "a" minimisant la somme des carrés des résidus est

$$a = ({}^tX\,X)^{-1}\,{}^tX\,Y.$$

Equation de l'analyse de la variance

Le but de la régression est d'expliquer pourquoi la valeur de la variable à expliquer fluctue. L'élément explicatif de ces fluctuations est constitué par les variations de

$$X\,a = a_1 X_1 + \cdots + a_j X_j + \cdots + a_p X_p$$

lorsque $X_1, \ldots, X_j, \ldots, X_p$ varient.

Notons Y_{ic} la valeur de la variable à expliquer pour l'individu i calculée par l'équation de régression

$$Y_{ic} = a_1 X_{i1} + \cdots + a_j X_{ij} + \cdots + a_p X_{ip}$$

c'est-à-dire

$$Y_{ic} = Y_i - e_i$$

Si Y_c désigne la série des Y_{ic} :

$$\bar{Y}_c = \bar{Y} \text{ car } \bar{e} = 0.$$

Equation de l'analyse de la variance

D'après le théorème de Pythagore, puisque (Xa) est orthogonal à e

$$\|Y\|^2 = \|Xa\|^2 + \|e\|^2$$

ce qui s'écrit aussi, en utilisant les séries non centrées, selon des notations évidentes

Proposition

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{ic} - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_{ic} - y_i)^2$$

Variance totale = Variance expliquée + Variance résiduelle

Equation de l'analyse de la variance

Remarque

*Cette équation porte le nom d'*équation d'analyse de la variance *: elle généralise l'équation d'analyse de la variance de la régression simple.*

Et comme pour la régression simple, on calcule un coefficient de détermination R^2, rapport de la variance expliquée à la variance totale, pour mesurer la qualité de l'ajustement obtenu

$$R^2 = \frac{\textit{Variance expliquée}}{\textit{Variance totale}} = \frac{\|Xa\|^2}{\|Y\|^2}$$

La racine carrée positive du coefficient de détermination, R, est le coefficient de corrélation multiple.

Matrice de variances-covariances / covariances et matrice des corrélations : `cov` et `cor`

- La matrice des variances-covariances / covariances :

$$V = \begin{pmatrix} cov(X_1, X_1) = var(X_1, X_1) & \dots & cov(X_1, X_j) & \dots & cov(X_1, X_p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(X_i, X_1) & \dots & var(X_i, X_i) & \dots & cov(X_i, X_p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cov(X_p, X_1) & \dots & cov(X_p, X_j) & \dots & var(X_p, X_p) \end{pmatrix}$$

- La matrice des corrélations : $R = Diag(1/\sigma_X)\ V\ Diag(1/\sigma_X)$:

$$R = \begin{pmatrix} cor(X_1, X_1) = 1 & \dots & cor(X_1, X_j) & \dots & cor(X_1, X_p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cor(X_i, X_1) & \dots & cor(X_i, X_i) = 1 & \dots & cor(X_i, X_p) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ cor(X_p, X_1) & \dots & cor(X_p, X_j) & \dots & cor(X_p, X_p) = 1 \end{pmatrix}$$

Matrice de variances-covariances : variables centrées

La matrice à inverser pour calculer le projecteur P est la matrice $({}^tX\,X)$.
L'élément situé à l'intersection de la ligne k et de la colonne j de $({}^tX\,X)$ est égal à $n\,cov(X_k, X_j)$.

Proposition

$\frac{1}{n}({}^tX\,X)$ la matrice de variances-covariances entre les variables. Ses éléments diagonaux sont les variances des variables et l'élément situé à l'intersection de la ligne k et de la colonne j est égal à $cov(X_k, X_j)$.

Remarque

Si les variables de départ sont non seulement centrées, mais de plus réduites, alors les éléments de $\frac{1}{n}({}^tX\,X)$ sont les coefficients de corrélation entre les variables et la diagonale de $\frac{1}{n}({}^tX\,X)$ est constituée de 1.

Relation entre une variable quantitative et un ensemble de variables quelconques

Le modèle de régression multiple peut être étendu aux cas où l'ensemble des variables explicatives est constitué par des variables quantitatives et/ou des variables qualitatives.

Rapport de corrélation

On dispose d'une variable Y quantitative centrée *à expliquer* et d'une seule variable qualitative explicative dont les q modalités sont décrites par le tableau disjonctif complet X.

Comme dans le cas de régression multiple, Y s'écrit comme une combinaison linéaire des colonnes (c'est-à-dire comme un codage ou une quantification de la variable qualitative, comme cela a été indiqué au début de ce chapitre), et vérifie approximativement

$$Y = X\,a + e.$$

En reprenant le même raisonnement qu'en régression multiple, le vecteur (Xa) est alors la projection orthogonale de Y sur l'espace engendré par les colonnes de X.

Relation entre une variable quantitative et un ensemble de variables quelconques

Le vecteur $\mathbb{I}$ dont toutes les composantes sont égales à 1 appartient à cet espace. Donc, puisque le vecteur e est orthogonal à tout vecteur de l'espace engendré par les colonnes de (Xa), $\bar{e} = \frac{1}{n}\,\mathbb{I}\,e = 0$.

Par hypothèse, la variable Y est centré. Par conséquent la variable

$$X\,a = Y - e,$$

est centrée.

Par analogie avec le coefficient de détermination, le rapport de corrélation η^2 est défini comme étant le rapport de la variance expliquée à la variance totale et mesure la qualité de l'ajustement obtenu. Ce rapport mesure l'intensité de la liaison entre la variable quantitative Y et la variable qualitative X.

Proposition

Le rapport de corrélation η^2 est égal à

$$\frac{\|X\,a\|^2}{\|Y\|^2} = \frac{\textit{variance expliquée}}{\textit{variance totale}}$$

Relation entre une variable quantitative et un ensemble de variables quelconques

Remarque

$({}^tX\,X)^{-1}\,{}^tX\,Y$ est le vecteur dont les q composantes sont les q valeurs moyennes à l'intérieur des classes et $\|X\,a\|^2$ mesure la variance inter-classes. Cette variance est donnée par :

$$\|X\,a\|^2 = \sum_{j=1}^{q} \frac{n_j}{n} (\bar{Y}_j)^2.$$

Remarque

Les deux cas extrêmes pouvant se produire sont :

- *la variable qualitative détermine entièrement la variable quantitative : les valeurs de la variable quantitative sont identiques à l'intérieur de chaque classe et différentes d'une classe à l'autre. Dans ce cas, le rapport de corrélation est maximal et égal à 1 car $\|X\,a\|^2 = \|PY\|^2$.*
- *la variable qualitative n'a aucun pouvoir explicatif, les valeurs moyennes sont égales d'une classe à l'autre (et donc nulles) : le rapport de corrélation est nul.*

Decomposition de la variance totale : inter et intra

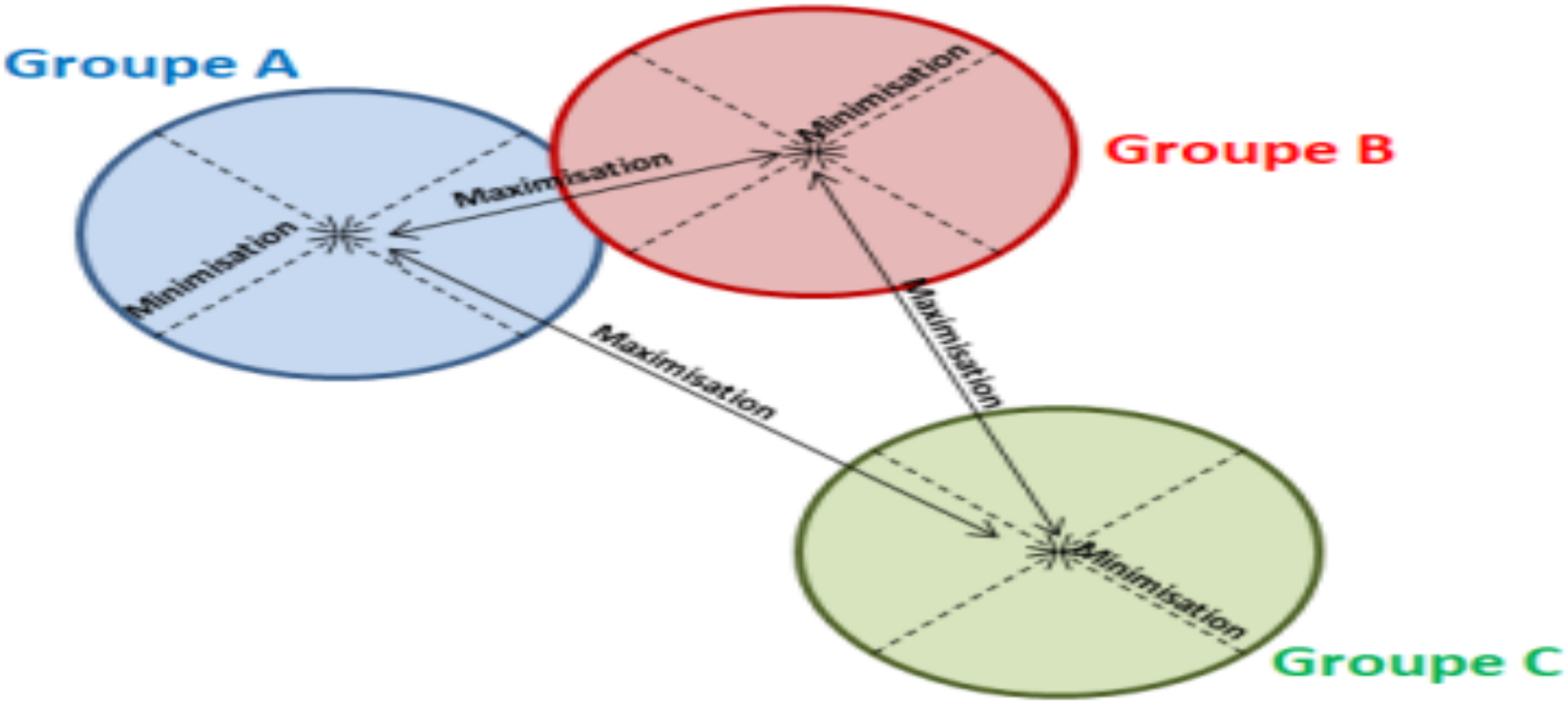


Figure – Illustration de la décomposition de la variance

Modèles d'analyse de la variance et de la covariance : ANOVA et ANCOVA

- La variable à expliquer étant quantitative, lorsque le tableau X est constitué par les modalités d'une ou plusieurs variables qualitatives, le modèle est un modèle d'analyse de la variance. Par exemple, le salaire (variable quantitative) peut être expliqué par le diplôme possédé (variable qualitative), le sexe (variable qualitative) ou le secteur de l'entreprise qui embauche (variable qualitative), ...

- Si la variable explicative X est constituée d'un mélange de variables quantitatives et de modalités de variables qualitatives, il s'agit d'un modèle d'analyse de la covariance. Par exemple, le prix d'un appartement peut être défini à partir de variables quantitatives (le nombre de m^2 de l'appartement, sa distance du centre-ville, ...) ou qualitatives (présence ou non d'ascenseur, d'un concierge, le type de quartier, ...).

Dans les deux cas, les calculs sont identiques aux calculs effectués dans le cadre de la régression multiple.

Relation entre deux variables qualitatives : Construction d'un tableau de contingence

Exemple

Considérons deux variables qualitatives X (la couleur des yeux, qui possède 3 *modalités : vert, bleu et marron) et Y (la couleur des cheveux, qui possède* 2 *modalités : châtin, aubrun). Chacune des* 2 *variables qualitatives est observée pour* 10 *individus et on obtient les tableaux disjonctifs complet suivants :*

$$X = \begin{pmatrix} 100 \\ 001 \\ 100 \\ 001 \\ 100 \\ 010 \\ 010 \\ 001 \\ 010 \\ 100 \end{pmatrix} \qquad Y = \begin{pmatrix} 01 \\ 01 \\ 01 \\ 10 \\ 10 \\ 01 \\ 10 \\ 10 \\ 10 \\ 01 \end{pmatrix}$$

Relation entre deux variables qualitatives : Construction d'un tableau de contingence

Exemple

On construit le tableau croisant les deux variables. Ce tableau indique le nombre d'individus possèdant à la fois une certaine modalité de la première variable qualitative et une certaine modalité de la seconde variable qualitative.

	Châtin	*Aubrun*
Vert	*1*	*3*
Bleu	*2*	*1*
Marron	*2*	*1*

Matriciellement, on peut obtenir ce tableau C selon la formule

$$C = {}^tX\,Y$$

Relation entre deux variables qualitatives : Construction d'un tableau de contingence

Exemple

Ce que l'on vérifie ici :

$$\begin{pmatrix} 1 & 3 \\ 2 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1010100001 \\ 0000011010 \\ 0101000100 \end{pmatrix} \begin{pmatrix} 01 \\ 01 \\ 01 \\ 10 \\ 10 \\ 01 \\ 10 \\ 10 \\ 10 \\ 01 \end{pmatrix}$$

Le tableau C s'appelle tableau de contingence ; bien entendu, ${}^tC$, qui donne les mêmes informations, mais en permutant les lignes et les colonnes, est aussi un tableau de contingence.

Tableau de contingence : des effectifs

Exemple

Considérons une population de 90 *individus triée selon* 2 *critères :*

- *l'âge qui comporte* 3 *modalités : moins de* 30 *ans, entre* 30 *et* 50 *ans, plus de* 50 *ans.*
- *le diplôme le plus élevé détenu qui comporte aussi* 3 *modalités : BEPC, Baccalauréat (BAC), Licence ou plus*

Le tableau de contingence croisant les 2 *variables est le suivant :*

	BEPC	*BAC*	*Licence*	*Total*
Plus de 50 *ans*	*15*	*12*	*3*	*30*
Entre 30 *et* 50 *ans*	*10*	*18*	*4*	*32*
Moins de 30 *ans*	*15*	*5*	*8*	*28*
Total	*40*	*35*	*15*	*90*

Tableau de contingence ou des effectifs observés

où *Total* désigne les marges du tableau : il y a 30 individus de $>$ 50 ans dans la population et parmi lesquels 15 individus titulaires d'une licence ou plus.

Tableau de contingence : des fréquences

Exemple

A partir de ce tableau des effectifs observés, on peut calculer le tableau des fréquences relatives observées, en divisant chaque terme par n, l'effectif total, soit 90.

	BEPC	*BAC*	*Licence*	*Fréquence relative de la ligne*
Plus de 50 *ans*	*15/90*	*12/90*	*3/90*	*30/90*
Entre 30 *et* 50 *ans*	*10/90*	*18/90*	*4/90*	*32/90*
Moins de 30 *ans*	*15/90*	*5/90*	*8/90*	*28/90*
Fréquence relative de la colonne	*40/90*	*35/90*	*15/90*	*90/90*

Tableau des fréquences observées

Soit f_{ij} : fréquence relative des individus possédant à la fois la modalité d'ordre i de la première variable qualitative et la modalité d'ordre j de la second variable.
$f_{i\bullet}$: fréquence relative de la modalité i de la première variable qualitative.
$f_{\bullet j}$: fréquence relative de la modalité j de la seconde variable qualitative.

Tableau de contingence : des fréquences

Exemple

$$f_{13} = \frac{3}{90},\ f_{1\bullet} = \frac{30}{90}\ et\ f_{\bullet 2} = \frac{35}{90}.$$

Profils des lignes

Le tableau des profils des lignes indique pour une modalité *i* donnée de la première variable qualitative la proportion d'individus possédant une modalité *j* donnée de la seconde variable qualitative.

Exemple

	BEPC	*BAC*	*Licence*	*Total*
Plus de 50 *ans*	*15/30*	*12/30*	*3/30*	*1*
Entre 30 *et* 50 *ans*	*10/32*	*18/32*	*4/32*	*1*
Moins de 30 *ans*	*15/28*	*5/28*	*8/28*	*1*

Tableau des profils des lignes

Ainsi, $10\% = 3/30 \times 100\%$ des plus de 50 ans sont titulaires d'une licence.

Profils des colonnes

De façon symétrique, le tableau des profils des colonnes indique pour une modalité *j* donnée de la seconde variable qualitative la proportion d'individus possédant une modalité *i* donnée de la première variable qualitative.

Exemple

	BEPC	*BAC*	*Licence*
Plus de 50 *ans*	*15/40*	*12/35*	*3/15*
Entre 30 *et* 50 *ans*	*10/40*	*18/35*	*4/15*
Moins de 30 *ans*	*15/40*	*5/35*	*8/15*
Total	*1*	*1*	*1*

Tableau des profils des colonnes

Ainsi, $20\% = 3/15 \times 100\%$ des titulaires d'une licence ont plus de 50 ans.

Indépendance de deux caractères qualitatifs
Le test de Khi-deux : `chisq.test`

Supposons que les deux variables qualitatives *diplôme* et *âge* soient indépendantes (ceci interdit des affirmations telles que : il est âgé, donc il est peu probable qu'il possède un diplôme élevé...).
Sous cette hypothèse d'indépendance, la fréquence théorique des individus possédant à la fois la modalité i de la première variable et la modalité j de la seconde variable est égale $f_{i\bullet} f_{\bullet j}$ et on peut construire le tableau des fréquences théoriques suivant :

Exemple

	BEPC	*BAC*	*Licence*	*Fréquence relative de la ligne*
Plus de 50 *ans*	*1200/8100*	*1050/8100*	*450/8100*	*30/90*
Entre 30 *et* 50 *ans*	*1280/8100*	*1120/8100*	*480/8100*	*32/90*
Moins de 30 *ans*	*1120/8100*	*980/8100*	*420/8100*	*28/90*
Fréquence relative de la colonne	*40/90*	*35/90*	*15/90*	*90/90*

Tableau des fréquences théoriques

Le tableau des effectifs théoriques est alors obtenu en multipliant les fréquences théoriques par l'effectif total, soit 90.

Exemple

	BEPC	*BAC*	*Licence*	*Fréquence relative de la ligne*
Plus de 50 ans	*13.33*	*11.66*	*5.00*	*30*
Entre 30 et 50 ans	*14.22*	*12.44*	*5.33*	*32*
Moins de 30 ans	*12.44*	*10.88*	*4.66*	*28*
Fréquence relative de la colonne	*40*	*35*	*15*	*90*

Tableau des effectifs théoriques

Indépendance de deux caractères qualitatifs
Le test de Khi-deux : `chisq.test`

L'écart entre le tableau des effectifs théoriques et le tableau des effectifs observés permet de détecter *les écarts à l'hypothèse d'indépendance* : ici, on met en évidence, par exmple, la sur-représentation des licenciés chez les moins de 30 ans (il y'en a 8 dans la réalité, pour un effectif théorique de 4.66).

Pour mesurer l'écart à l'indépendance, une distance entre le tableau des effectifs observés et le tableau des effectifs théoriques est définie. Cette distance est notée χ^2.

Si on désigne par O_{ij} (resp. T_{ij}) l'élément situé à l'intersection de la ligne i et de la colonne j du tableau des effectifs observés (resp. théoriques), alors :

Proposition

$$\chi^2 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \frac{(O_{ij} - T_{ij})^2}{T_{ij}},$$

où p_1 et p_2 sont respectivement le nombre de lignes et de colonnes des tableaux.

Indépendance de deux caractères qualitatifs
Le test de Khi-deux : `qchisq`

Exemple

Dans l'exemple ci-dessus : $p_1 = p_2 = 3$ *et* $\chi^2 = 11.18$ *et les références sont* $\texttt{qchisq}(0.975, 3, 3) = 16.52135$ *et* $\texttt{qchisq}(0.25, 3, 3) = 2.826002$.

Le χ^2 peut être calculé en comparant le tableau des fréquences observées au tableau des fréquences théoriques.

Proposition

$$\chi^2 = n \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}.$$

Remarque

Il faut noter que l'approche descriptive utilisée ici pour présenter les coefficients de mesure entre plusieurs ensembles de variables (coefficient de corrélation et khi-deux) peut être complétée par une approche utilisant les outils de l'inférence statistique.

Tableau de Burt

Exemple

Considérons 3 variables qualitatives, les 2 variables précédentes X et Y, et une troisième variable Z, décrite elle aussi par un tableau disjonctif complet

$$X = \begin{pmatrix} 100 \\ 001 \\ 100 \\ 001 \\ 100 \\ 010 \\ 010 \\ 001 \\ 010 \\ 100 \end{pmatrix} \quad Y = \begin{pmatrix} 01 \\ 01 \\ 01 \\ 10 \\ 10 \\ 01 \\ 10 \\ 10 \\ 10 \\ 01 \end{pmatrix} \quad Z = \begin{pmatrix} 010 \\ 100 \\ 100 \\ 010 \\ 100 \\ 001 \\ 001 \\ 010 \\ 100 \\ 100 \end{pmatrix}$$

Le tableau de Burt B est la tableau croisant les modalités des 3 variables qualitatives :

$$B = {}^t(X, Y, Z)\,(X, Y, Z).$$

Exemple

Numériquement, ici

$$\begin{pmatrix}
4 & 0 & 0 & 1 & 3 & 3 & 1 & 0 \\
0 & 3 & 0 & 2 & 1 & 1 & 0 & 2 \\
0 & 0 & 3 & 2 & 1 & 1 & 0 & 2 \\
\\
1 & 2 & 2 & 5 & 0 & 2 & 2 & 1 \\
3 & 1 & 1 & 0 & 5 & 3 & 1 & 1 \\
\\
3 & 1 & 1 & 2 & 3 & 5 & 0 & 0 \\
1 & 0 & 2 & 2 & 1 & 0 & 3 & 0 \\
0 & 2 & 0 & 1 & 1 & 0 & 0 & 2
\end{pmatrix}$$

Tableau de Burt

Définition

Un tableau de Burt est un tableau croisant les modalités de plusieurs variables qualitatives.

Propriétés

- La diagonale principale du tableau de Burt est constituée par les effectifs des modalités, et on retrouve dans le tableau de Burt tous les tableaux de contingence croisant les variables qualitatives deux à deux.
- Le tableau de Burt est symétrique (par construction) ; la somme des éléments d'une ligne (et, par symétrie, la somme des éléments d'une colonne) est égale à l'effectif relatif de la modalité correspondante multipliée par le nombre de variables qualitatives (ici 3), puisque la modalité est décrite autant de fois qu'il y a de variables qualitatives.
- Le tableau de Burt décrit toutes les relations entre les variables qualitatives prises deux à deux.

Transformation d'une variable quantitative en une variable qualitative par découpage en classes

Une variable quantitative peut être transformée en une variable qualitative. Il suffit de créer des classes de telle manière que chaque individu appartienne à une classe et une seule. Ainsi, si on considère la variable *revenu annuel*, on peut définir des classes ou des tranches de revenu de la manière suivante :

- classe 1 : revenu annuel inférieur à 1200 Euros.
- classe 2 : revenu annuel compris entre 1200 Euros et 1500 Euros.
- ...
- classe p : revenu annuel supérieur à 22000 Euros.

Chaque individu appartient à une tranche de revenu et une seule et on a donc créé une variable qualitative.

Transformation d'une variable quantitative en une variable qualitative par découpage en classes

Remarque

- *A priori, cette transformation semble peu intéressante puisqu'elle occasionne une perte d'information : le revenu d'un individu donné est connu avec moins de précision puisqu'on ne connaît plus son montant exact, mais seulement la classe à laquelle appartient ce revenu.*
- *Mais, à partir de deux variables quantitatives découpées en tranches, il peut être déterminé le tableau de contingence croisant les classes des deux variables qualitatives ainsi constituées. Et ce tableau fournit des informations plus intéressantes sur la liaison entre les variables que le simple calcul du coefficient de corrélation entre les variables quantitatives d'origine.*
- *Et ce d'autant plus que ces informations pourront être traitées notamment par une technique d'analyse de tableaux de contingence, l'Analyse Factorielle des Correspondances, étudiée au chapitre* 4*.*

Distances dans l'espace des individus

Considérons le tableau de données suivant :

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

Un individu décrit par une ligne du tableau X prend p valeurs, une pour chaque variable, et peut donc être représenté par un point dans $\mathbb{R}^p$, l'espace des individus.

Lorsque $\mathbb{R}^p$ est muni du produit scalaire usuel, si on considère 2 individus i et i' dont les coordonnées sont données respectivement par les 2 vecteurs $(x_{i1}, \dots, x_{ij}, \dots, x_{ip})$ et $(x_{i'1}, \dots, x_{i'j}, \dots, x_{i'p})$, alors la distance entre les individus i et i' notée $d(i, i')$ est donnée par

$$d^2(i, i') = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Distances dans l'espace des individus

Mais $\mathbb{R}^p$ peut aussi être muni d'un autre produit scalaire, c'est-à-dire qu'il est possible de définir d'autres façons de calculer la distance entre 2 points.

En particulier, il est possible d'associer à chaque variable X_j un nombre positif ν_j et la distance $d(i, i')$ est alors

$$d^2(i, i') = \sum_{j=1}^{p} \nu_j (x_{ij} - x_{i'j})^2$$

Distances dans l'espace des individus

- Ainsi, les variables associées à des coefficient ν_j ont-elles une importance plus élevée dans la définition de la distance que celles qui ont un poids faible.
- Le choix d'un système de coefficients n'est pas neutre : la distance entre 2 points étant la mesure de la *ressemblance* entre les individus correspondants (2 individus *ressemblants* ayant une faible distance entre eux), cette distance doit être définie en fonction de l'idée que l'on se fait de cette *ressemblance* entre 2 points.
- Contrairement à l'espace des variables où un produit scalaire particulier s'impose car il permet des interprétations géométriques simples des concepts statistiques, dans l'espace des individus le choix d'un produit scalaire (donc d'une distance) est toujours arbitraire.

Méthodes et logiciels d'analyse de données
Les logiciels existant

Les techniques exposées dans ce chapitre permettent de mesurer l'intensité de la relation entre deux ensembles contenant chacun un certain nombre de variables.

Les méthodes d'analyse de données prolongent ces techniques de statistique descriptive. La mise en oeuvre des ces méthodes nécessite des calculs importants, qu'il n'est pas possible d'effectuer sans avoir recours à des logociels de statistique. Parmi les nombreux logiciels existant, les plus utilisés sont :

- SPAD : CISIA, 1 Avenue Herbillon, 94160 Saint-Mandé.
- SAS : SAS Institute, Domaine de Grégy, BP. 5, 77166 Evru Grégy-sur-Yerres.
- SPSS : Conceptel, 16 Rue d'Ouessant, 75015 Paris.
- STATITCF : ITCF, Boigneville, 91720 Maisse.
- LE SPHINX : Le Sphinx, 7 Rue Blaise Pascal, 74600 Seynod.

Les logiciels existant

- ADDAD : LADDAD, 22 Rue Charcot, 75013 Paris.
- BMDP : Statistical Software Ldt, Cork Technology Park, Model Farm Roed, Cork, Ireland.
- Python
- Le logiciel R reprend les fonctionnalités de Splus. Ce logiciel est free : http ://pbli.univ-lyon1.fr/R

- Il est à noter que quelques différences (concernant notamment les calculs effectués et les sorties graphiques) existent entre les logiciels anglo-saxons et les logiciels français, différences qui correspondent à des différences de conception des méthodes d'analyse des données. Le logiciel utilisé pour les calculs effectués dans ce ce fascicule est `R`.
- Les différents logiciels permettent sans difficultés de saisir des données ou de traiter des données stockées sous des formes diverses (ASCII, EXCEL, DBASE, ...).
- De fait, la principale difficulté rencontrée lors de l'utilisation d'un logiciel est le choix de la procédure appropriée pour traiter les données dont on dispose, c'est-à-dire le choix de la méthode statistique à retenir.

Méthodes et logiciels d'analyse de données

- L'ANALYSE EN COMPOSANTE PRINCIPALE : a pour objectif la description de grands ensembles de variables quantitatives observées pour les mêmes individus, c'est-à-dire synthétise l'information contenue dans un tableau dont le nombre d'observations et/ou le nombre de variables est élevé ; cette technique est décrite au chapitre 3.
- L'ANALYSE CANONIQUE : permet de saisir l'essentiel des relations linéaires existant entre deux ensembles de variables décrites pour les mêmes individus. Le chapitre 4, qui peut être lu indépendamment des autres chapitres, décrit cette méthode.
- L'ANALYSE FACTORIELLE DES CORRESPONDANCES : décrit les relations linéaires existant entre deux variables qualitatives, c'est-à-dire, en d'autres termes, met en évidence les éléments essentiels d'un tableau de contingence.
- L'ANALYSE FACTORIELLE DES CORRESPONDANCES MULTIPLES : décrit les relations linéaires existant entre plusieurs variables qualitatives. C'est une extension de l'analyse factorielle des correspondances à plus de deux ensembles de variables qualitatives. Elle s'applique à des tableaux de Burt, ...

Méthodes et logiciels d'analyse de données

- L'ANALYSE FACTORIELLE DES DONNÉES MIXTES : décrit les relations linéaires existant entre plusieurs variables qualitatives et/ou qualitatives.
- L'ANALYSE FACTORIELLE DISCRIMINANTE : examine les liens existant entre un ensemble de variables quantitatives et une variable qualitative, c'est-à-dire relie l'appartenance d'un individu à une classe donnée aux valeurs qu'il prend pour les variables quantitatives.
- CLASSIFICATION HIÉRARCHIQUE : Les méthodes de *classification automatique* ne nécessitant pas d'apprentissage ont un intérêt important lorsque les données sont complètement inconnues. Elles permettent ainsi de dégager des classes qui ne sont pas évidentes a priori. Les deux principales méthodes développées sont la méthode des centres mobiles (apparentée à la méthode des *k-means* ou des nuées dynamiques (comme un cas particulier)) et la *classification hiérarchique ascendante* ou *descendante*. Nous pouvons également citer les approches fondées sur les graphes et hypergraphes, ...