

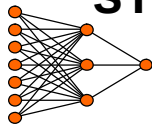
Stéphane Tufféry

Statisticien - Data Miner - Enseignant

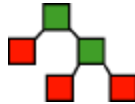


DATA MINING - SCORING

STATISTIQUE DÉCISIONNELLE



APPLICATION AU CRM



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

1

Plan du cours

- Qu'est-ce que le data mining ?
- À quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs à éviter
- Informatique décisionnelle et de gestion
- **La préparation des données**
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- Logiciels et consultants
- CNIL et limites légales du data mining
- Le web mining
- Le text mining

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

2

La préparation des données : Analyse exploratoire



Les différents formats de données

- Données *continues* (ou *d'échelle*)
 - dont les valeurs forment un sous-ensemble infini de **R** (exemple : salaire)
- Données *discrètes*
 - dont les valeurs forment un sous-ensemble fini ou infini de **N** (exemple : nombre d'enfants)
- Données *catégorielles* (ou *qualitatives*)
 - dont l'ensemble des valeurs est fini — ces valeurs sont numériques ou alphanumériques, mais quand elles sont numériques, ce ne sont que des codes et non des quantités (ex : PCS, no de département)
- Données *textuelles*
 - lettres de réclamation, rapports, dépêches AFP...



Précisions sur les formats

- Les données *continues* et *discrètes* sont des quantités :
 - on peut effectuer sur elles des opérations arithmétiques
 - elles sont ordonnées (on peut les comparer par la relation d'ordre $<$).
- Les données *catégorielles* ne sont pas des quantités
 - mais sont parfois ordonnées : on parle de données catégorielles **ordinales** (exemple : « faible, moyen, fort »)
 - données ordinales souvent traitées comme données discrètes
 - les données catégorielles **nominales** ne sont pas ordonnées
- Les données *textuelles* contiennent :
 - des abréviations
 - des fautes d'orthographe ou de syntaxe
 - des ambiguïtés (termes dont le sens dépend d'un contexte non facilement détectable automatiquement)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

5

Algorithmes et formats gérés

- La *régression linéaire* traite les variables continues
 - L'*analyse discriminante* traite les variables explicatives continues et les variables « cible » nominales
 - La *régression logistique* traite les variables explicatives continues ou binaires et les variables « cible » nominales ou ordinales
 - Les *réseaux de neurones* traitent de préférence les variables continues dans $[0,1]$
 - Certains *arbres de décision* (CHAID) traitent directement les variables discrètes et catégorielles mais discrétisent les variables continues
 - D'autres arbres (CART, C4.5, C5.0) peuvent aussi traiter directement les variables continues
- > *Tous les algorithmes n'admettent pas tous les types de données en entrée*

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

6

Changement de format

type de départ	type d'arrivée	opération	principe
continu	discret	discrétisation	découpage des valeurs en tranches
discret ou catégorique	continu	ACM	une Analyse des Correspondances Multiples fournit des facteurs continus à partir des données de départ

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

7

Pourquoi discrétiser ?

- Traiter simultanément des données quantitatives et qualitatives
- Appréhender des liaisons non linéaires (de degré >1) entre les variables continues
 - par une ACM, une régression logistique ou une analyse discriminante DISQUAL (Gilbert Saporta)
- Neutraliser les valeurs extrêmes
- Gérer les valeurs manquantes
- Renforce la robustesse d'un modèle logistique sur un faible nombre d'individus

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

8

Comment discrétiser ?

- Il faut garder en tête que :
 - il faut éviter d'avoir de grands écarts entre le nombre de modalités des différentes variables
 - un nombre convenable de modalités pour une variable discrète ou catégorielle tourne autour de 4 ou 5.
- pour les raisons que :
 - le poids d'une variable est proportionnel à son nombre de modalités
 - le poids d'une modalité est inversement proportionnel à son effectif
 - avoir peu de modalités fait perdre de l'information
 - avoir beaucoup de modalités implique des petits effectifs et une moindre lisibilité.

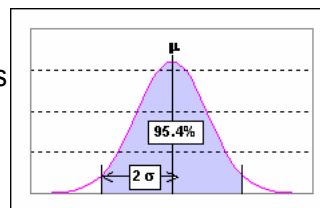
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

9

Analyse exploratoire des données

- Explorer la distribution des variables
- Vérifier la fiabilité des variables
 - valeurs incohérentes ou manquantes
 - => imputation ou suppression
- Détecter les valeurs extrêmes
 - voir si valeurs aberrantes à éliminer
- Tester la normalité des variables
- Tester l'homoscédasticité
- Détecter les liaisons entre variables
 - entre variables explicatives et à expliquer (bon)
 - entre variables explicatives entre elles (mauvais dans certaines méthodes : multicolinéarité)



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

10

Analyse exploratoire des données

- Variables continues
 - détecter la non-linéarité justifiant la discrétisation
 - transformer pour augmenter la normalité
- Variables discrètes
 - regrouper certaines modalités aux effectifs trop petits (poids trop grand)
- Créer des indicateurs pertinents à partir des données brutes (ratios X/Y ou $X(\text{période } t)/X(\text{période } t-1)$)
 - prendre l'avis des spécialistes du secteur étudié

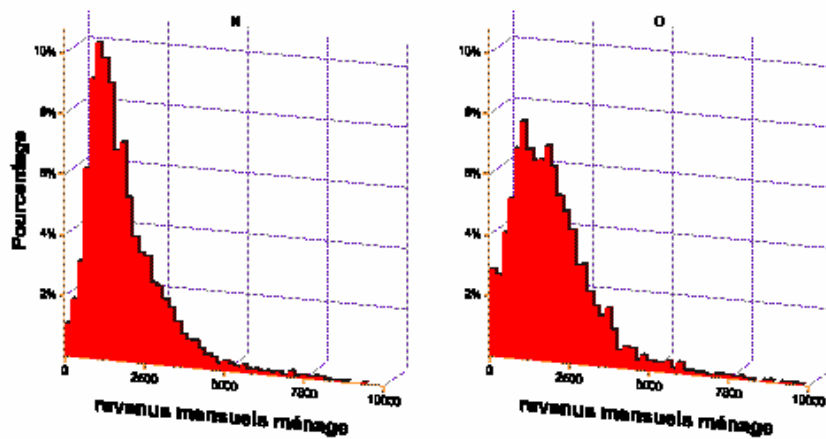


Caractéristiques de tendance centrale

- Mode
- Moyennes (arithmétique, géométrique, harmonique)
- Médiane
- Autres quantiles
- Découpage en quartiles et déciles souvent utilisé pour :
 - représentation graphique
 - discrétisation
 - croisement avec variable cible



Analyse exploratoire des données

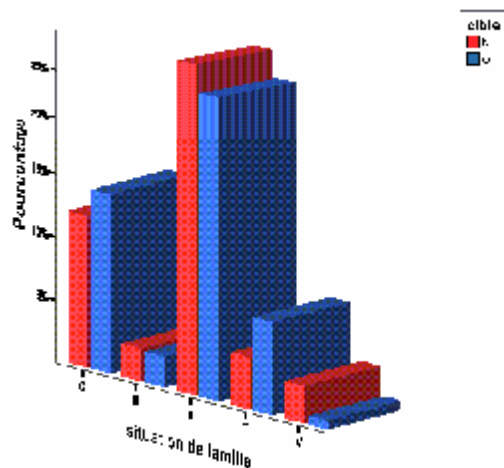


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

13

Analyse exploratoire des données

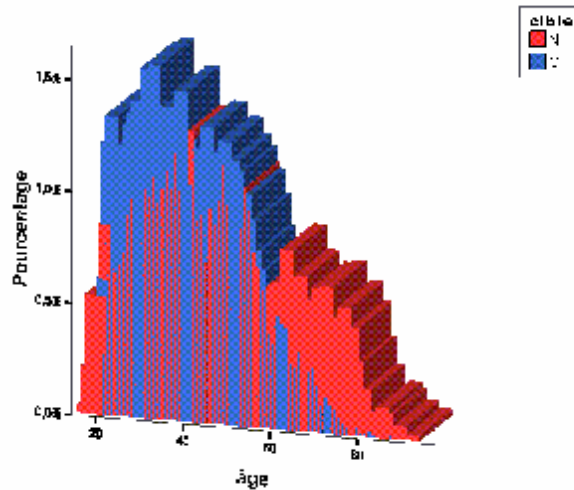


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

14

Analyse exploratoire des données

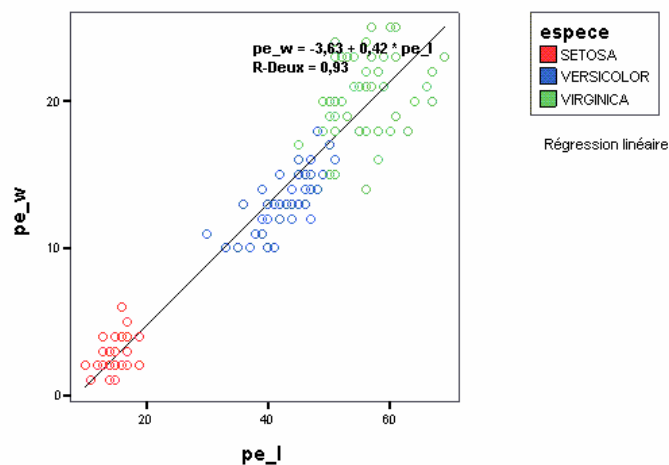


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

15

Analyse exploratoire des données



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

16

Caractéristiques de dispersion

- Étendue
- Écart interquartile $q_3 - q_1$
- Variance
 - H_0 = égalité des variances d'une variable dans plusieurs groupes : homoscedasticité (contraire : hétéroscedasticité)
 - test de Levene, de Bartlett ou de Fisher
 - $\text{proba} < 0,05 \Rightarrow$ hétéroscedasticité
- Écart-type
- Coefficient de variation
 - écart-type / moyenne
 - X dispersée si $\text{CV}(X) > 25 \%$
 - grandeur sans unité $\Rightarrow$ utile pour comparer la dispersion des variables

Test of Homogeneity of Variance

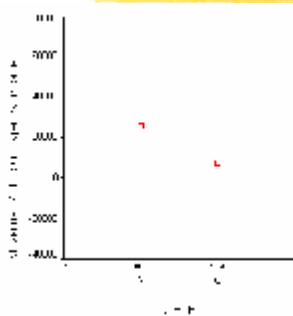
	Levene Statistic	df1	df2	Sig.
1996 Sales	6.708	2	387	.001

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

17

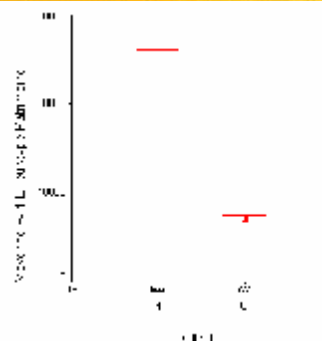
Homogénéité des variances



KO é

OK ->

Erreur-type =
écart-type de la
moyenne = écart-
type des
observations /
racine carrée de
l'effectif

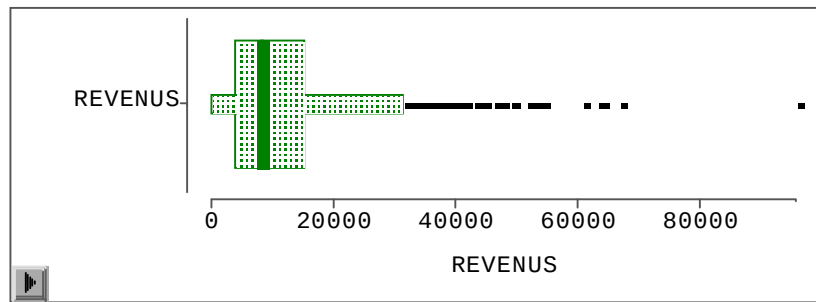


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

18

Boîte « à moustaches » (boxplot)



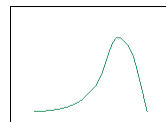
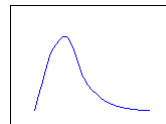
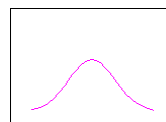
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

19

Caractéristiques de forme 1/2

- Coefficient d'asymétrie (« skewness »)
 - = 0 si la série de données est symétrique
 - > 0 si elle est allongée vers la droite
 - < 0 si elle est allongée vers la gauche
- Asymétrie positive fréquente dans les données économiques



18/04/2005

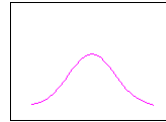
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

20

Caractéristiques de forme 2/2

- Coefficient d'aplatissement (« kurtosis »)

- = 3 si aplatie comme Gauss
- > 3 si plus concentrée que Gauss
- < 3 si plus aplatie que Gauss



- Kurtosis (loi uniforme sur $[0,1]$) = 1,8
- **On normalise souvent le kurtosis en soustrayant 3**

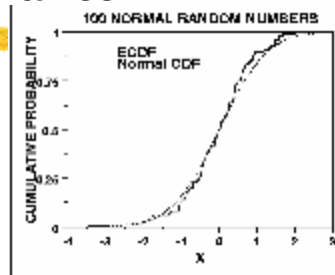
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

21

Tests de normalité

- Test de Kolmogorov-Smirnov
 - mesure l'écart maximum (en valeur absolue) entre la fonction de répartition de la variable testée et celle d'une variable normale
 - hypothèse nulle H_0 : les données suivent une distribution donnée (ici une distribution normale, mais le test de K-S s'applique + généralement à d'autres distributions continues)



- test plus sensible au centre de la distribution

Tests for Normality		
Test Statistic	Value	p-value
Shapiro-Wilk	0.985964	0.0000
Kolmogorov-Smirnov	0.047426	<.0100
Cramer-von Mises	0.329094	<.0050
Anderson-Darling	2.656450	<.0050

Invalidation de H_0 :
{distribution normale}

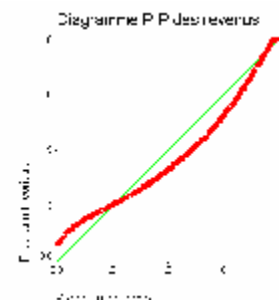
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

22

Tests de normalité

- Test d'Anderson-Darling
 - modifie Kolmogorov-Smirnov en donnant plus de poids aux queues de distribution
- Test de Lilliefors
 - perfectionne K-S quand on ne connaît pas la moyenne et la variance
 - car dans ce cas le test de K-S est conservateur si calculé avec la moyenne et la variance estimées sur l'échantillon
- Test de Shapiro-Wilk
 - mesure l'alignement sur la droite ci-contre correspondant à une distribution normale
 - le meilleur test sur de petits échantillons



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

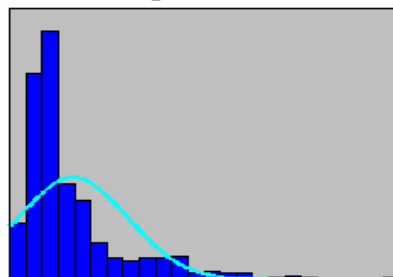
23

Normalisation

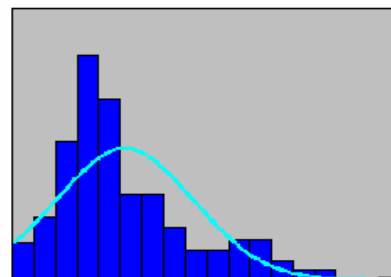
Frequencies Statistics

	Mean	Median	Std. Deviation	Skewness	Kurtosis
1996 Sales	\$371,893	\$307,500	\$171,311	2.112	5.247
Log of Sales	5.5367	5.4878	.1603	1.110	.791

Original Data



Log Transformed Data



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

24

Normalisation : transformations

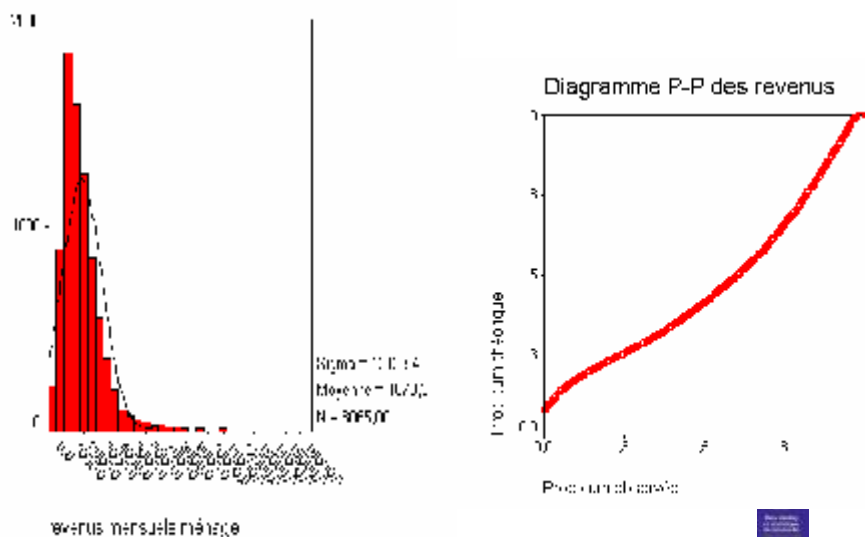
- Log (V)
 - transformation la plus courante pour corriger un coefficient d'asymétrie > 0
 - Si $V \geq 0$, on prend $\text{Log}(1 + V)$
- Racine carrée (V) si coefficient d'asymétrie > 0
- $-1/V$ ou $-1/V^2$ si coefficient d'asymétrie > 0
- V^2 ou V^3 si coefficient d'asymétrie < 0
- Arc sinus (V)
 - si V est un pourcentage compris entre 0 et 100
- Certains logiciels déterminent automatiquement la transformation la plus adaptée
 - en utilisant l'algorithme de Box et Cox ou la loi de Taylor

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

25

Normalisation : exemple des revenus

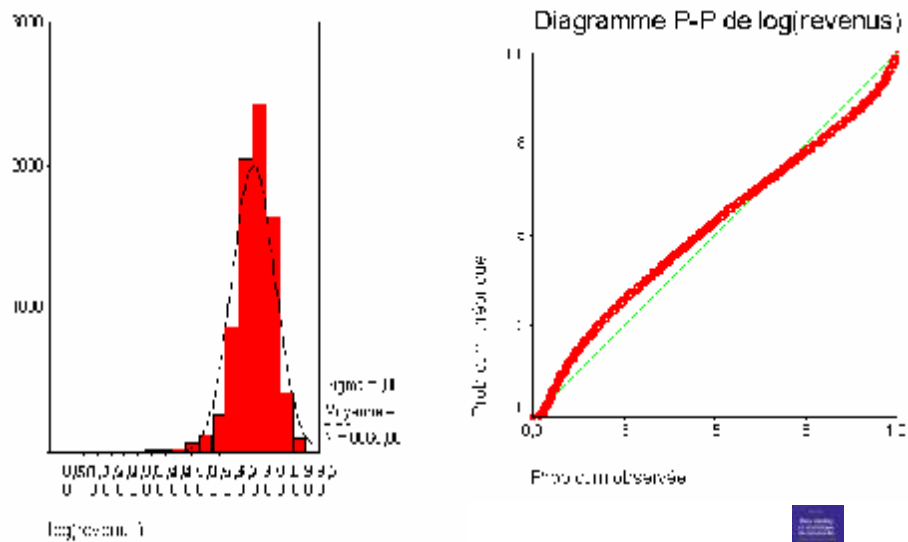


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

26

Normalisation par le logarithme

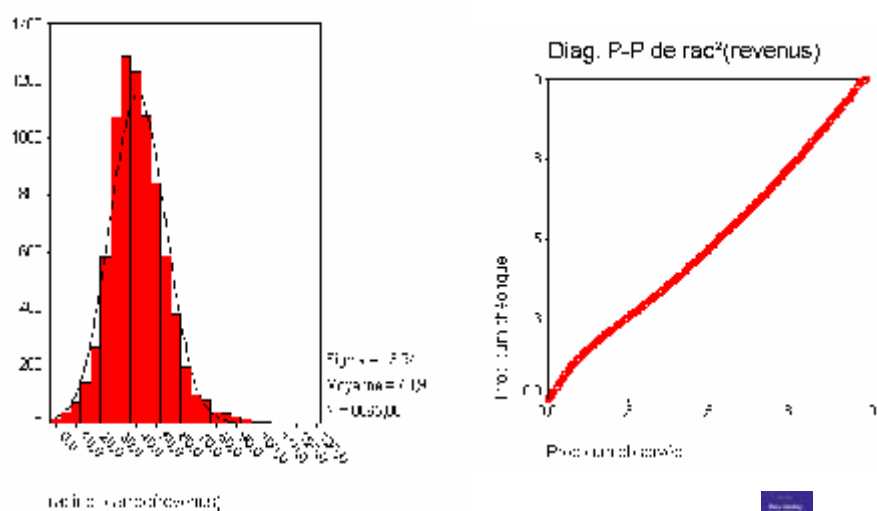


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

27

Normalisation par la racine carrée

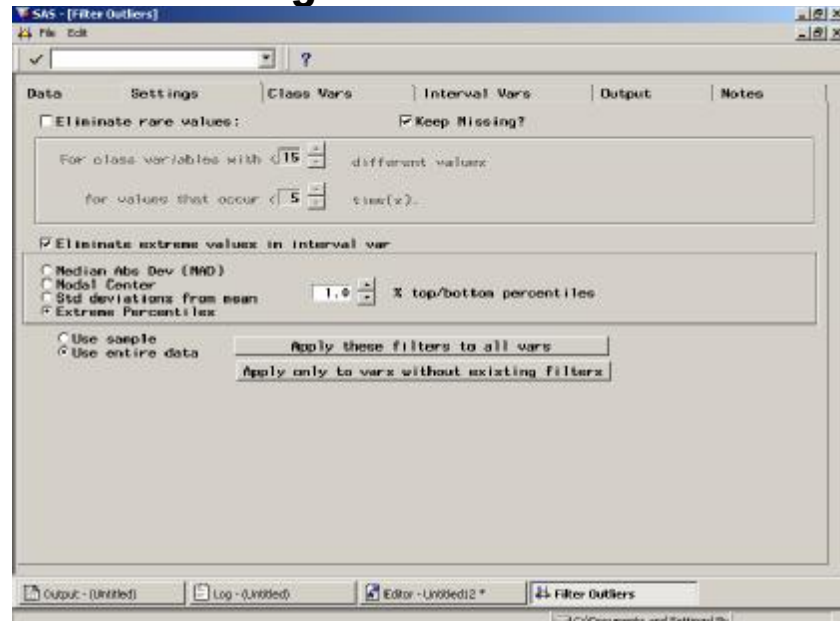


18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

28

Filtrage des extrêmes



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

29

Tableau de contingence

- Un **tableau de contingence** est le croisement de 2 variables catégorielles A et B : le coefficient x_{ij} du tableau = nb d'individus x tels que $A(x) = a_i$ et $B(x) = b_j$
- Le test du χ^2 permet de détecter une dépendance entre les deux variables
- La contribution au χ^2 de chaque cellule du tableau de contingence montre les liaisons entre modalités des 2 variables : soit sur-effectif, soit sous-effectif, soit équilibre
- S'il y a de nombreuses modalités, il est fastidieux de parcourir toutes les cellules
- S'il y a plus de 2 variables à croiser, c'est encore + ardu, voire impossible $\Rightarrow$ se tourner vers l'analyse des correspondances

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

30

Pièges des tableaux de contingence et paradoxe de Simpson 1/2

Tous clients				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	2 850	58	2 908	1,99%
<i>téléphone</i>	2 185	78	2 263	3,45%
<i>TOTAL</i>	5 035	136	5 171	2,63%
Hommes				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	2 341	24	2 365	1,01%
<i>téléphone</i>	515	3	518	0,58%
<i>TOTAL</i>	2 856	27	2 883	0,94%
Femmes				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	509	34	543	6,26%
<i>téléphone</i>	1 670	75	1 745	4,30%
<i>TOTAL</i>	2 179	109	2 288	4,76%

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

31

Pièges des tableaux de contingence et paradoxe de Simpson 2/2

Tous clients				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	1 400	100	1 500	6,67%
<i>téléphone</i>	1 375	125	1 500	8,33%
<i>TOTAL</i>	2 775	225	3 000	7,50%
Hommes				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	950	50	1 000	5,00%
<i>téléphone</i>	475	25	500	5,00%
<i>TOTAL</i>	1 425	75	1 500	5,00%
Femmes				
	<i>sans achat</i>	<i>avec achats</i>	<i>TOTAL</i>	<i>taux d'achat</i>
<i>courriel</i>	450	50	500	10,00%
<i>téléphone</i>	900	100	1 000	10,00%
<i>TOTAL</i>	1 350	150	1 500	10,00%

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

32

Paradoxe de Simpson : explication

- Dans le dernier exemple :
 - les hommes ne répondent pas mieux au téléphone qu'au courriel
 - de même pour les femmes
 - et pourtant, le téléphone semble avoir globalement un meilleur taux d'achat
- Explication :
 - un individu pris au hasard ne répond pas mieux au téléphone
 - mais les femmes achètent plus et on a privilégié le téléphone pour les contacter (liaison positive entre les variables « sexe » et « canal de distribution »)



La préparation des données : Tests statistiques



Propriétés de la loi normale

- Elle est entièrement définie par sa moyenne μ et son écart-type σ et toute loi $N(\mu, \sigma)$ peut se ramener à une loi $N(0, 1)$
- Sa moyenne = sa médiane = son mode
- Loi limite des lois binomiales, de Poisson et du χ^2
- Loi très fréquente dans les phénomènes biologiques et médicaux
- Les observations sont distribuées symétriquement autour de la moyenne :
 - 68 % des observations se trouvent à une distance $\pm \sigma$ de μ
 - 95 % des observations se trouvent à une distance $\pm 2\sigma$ de μ
 - 99,8 % des observations se trouvent à une distance $\pm 3\sigma$ de μ
- Une transformation permet parfois de rendre normale une variable

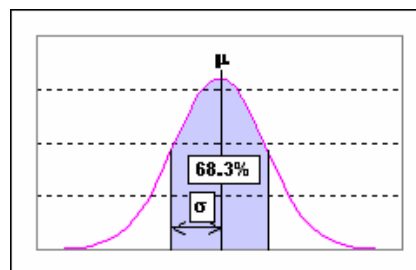
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

35

Situation de la loi normale

- Référence pour de nombreux indices
- Cadre de nombreux tests (t de Student, ANOVA, corrélation de Pearson)
- Hypothèse dans l'analyse discriminante de Fisher, dans la régression linéaire, etc.



- Non normalité moins gênante si les effectifs sont grands

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

36

Principe général d'un test d'hypothèse

- Quand on veut démontrer l'hypothèse H_1 que :
 - une moyenne mesurée dans un échantillon est significativement différente de la moyenne dans la population
 - significativement = ne résulte pas uniquement du hasard
 - des moyennes mesurées dans 2 échantillons sont significativement différentes
 - une variable ne suit pas une loi théorique donnée
 - deux variables sont significativement différentes
 - un échantillon n'est pas homogène mais est composé de plusieurs sous-populations
- ... on soumet l'hypothèse contraire H_0 à un test T qui doit être satisfait si H_0 est vraie
- ... puis on montre que T n'est pas satisfait $\Rightarrow H_0$ est faux
- Vocab. : H_0 : hypothèse nulle – H_1 : hypothèse alternative

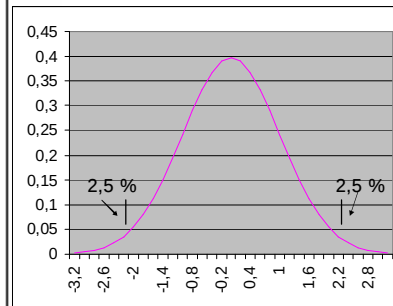
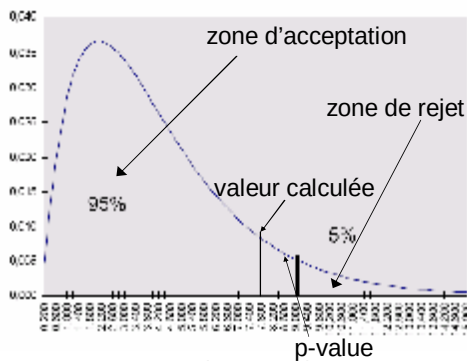
Exemples

- Égalité de moyennes dans 2 échantillons : test de Student
- Égalité de moyennes dans $k > 2$ échantillons : analyse de la variance
- Égalité de 2 variances : test de Fisher-Snedecor
- Égalité de 2 distributions : test de Kolmogorov-Smirnov
- Indépendance de 2 variables qualitatives : test du χ^2
 - ce test est non-paramétrique
 - mais non exact -> le test exact correspondant est le test de Fisher (ne pas confondre avec le test de Fisher-Snedecor)
 - voir plus loin ces notions de « paramétrique » et « exact »



Déroulement du test d'hypothèse 1/2

- À l'hypothèse nulle H_0 est associée une statistique, fonction des observations, qui suit une loi théorique connue si H_0 est vraie
 - exemple : si l'hypothèse nulle est ($H_0 : \mu = \mu_0$), alors $\frac{\bar{x} - m_0}{s/\sqrt{n}}$ suit une loi normale réduite



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

39

Déroulement du test d'hypothèse 2/2

- Dans la distribution de cette loi, on choisit une zone de rejet (unilatérale ou bilatérale), caractérisée par une probabilité α d'être dans cette zone
 - on choisit souvent $\alpha = 0,05$ (= 5 %)
 - le complémentaire est la zone d'acceptation (si $\alpha = 0,05$, il s'agit de la région autour de la moyenne où se trouvent 95 % des valeurs de la statistique)
- On mesure la valeur de la statistique sur l'échantillon et on compare cette valeur aux valeurs théoriques de la loi
- Si cette valeur mesurée tombe dans la zone de rejet, on rejette H_0
 - sinon, on ne la rejette pas

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

40

Niveau de signification (p-value)

- Niveau de signification = degré de signification = p-value = probabilité d'obtenir une statistique de test aussi extrême ($\geq$ ou $\leq$) que la valeur mesurée sur l'échantillon si H_0 est vraie
- Utilisation de la p-value :
 - $p\text{-value} \geq \alpha \Rightarrow$ ne pas rejeter H_0
 - $p\text{-value} < \alpha \Rightarrow$ rejeter H_0 (on considère qu'il est trop peu probable d'avoir une si faible p-value si H_0 est vraie, pour admettre que H_0 est vraie)
- Intérêt de la p-value :
 - elle a un sens absolu, qui ne dépend pas de la loi de probabilité et du nb de degrés de liberté



Risques d'erreur

- Les deux erreurs possibles dans un test sont :
 - le rejet d'une H_0 vraie (risque de 1^{ère} espèce, ou de type I)
 - probabilité de cette erreur = α
 - le non rejet d'une H_0 fausse (risque de 2^{de} espèce, ou de type II)
 - probabilité de cette erreur = β

		REALITE	REALITE
		H_0 vraie	H_0 fausse
DECISION	H_0 non rejetée	décision correcte ($1 - \alpha$)	risque β (type II)
DECISION	H_0 rejetée	risque α (type I)	décision correcte ($1 - \beta$)

- On ne peut réduire simultanément α et β



Puissance d'un test 1/2

- Puissance d'un test : $1 - \beta$
- Probabilité de rejeter H_0 si celle-ci est fausse
 - décision correcte
- Le risque β et la puissance $1 - \beta$ dépendent de :
 - la vraie valeur du paramètre de la population (plus elle est éloignée de la valeur testée, plus le risque β baisse)
 - l'écart-type σ de la population ($\sigma \uparrow \Rightarrow \beta \uparrow$)
 - le risque α choisi ($\alpha \uparrow \Rightarrow \beta \uparrow$)
 - la taille n de l'échantillon ($n \uparrow \Rightarrow \beta \downarrow$)
- Synonymes : test libéral = test puissant
- Antonymes : test conservateur $\neq$ test puissant

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

43

Puissance d'un test 2/2

- La puissance d'un test augmente avec la taille de l'échantillon
 - plus les observations sont nombreuses, plus on a d'éléments permettant de rejeter H_0 si elle est fausse
- Attention, avec des tests puissants, on rejette facilement H_0 dès que le nb d'observations augmente
 - ex : le test du χ^2
 - ex : les tests de normalité
- Remarque : les tests d'hypothèse s'appliquent bien à des hypothèses H_0 contraignantes (ex : $\mu = \mu_0$) car elles conduisent à des tests T précis $\Rightarrow$ les tests permettent de prouver qu'un échantillon est hétérogène ou n'a pas été constitué par un tirage au hasard, mais non l'inverse

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

44

Tests asymptotiques et exacts

- Test asymptotique
 - approximation valable quand les effectifs sont assez grands et les tableaux de données assez denses
 - ex : test du χ^2 (si effectifs théoriques ≥ 5)
- Test exact
 - utilisable sur des données éparses
 - calcul direct de probabilité
 - ... prenant en compte tous les cas de figure possibles
 - calcul pouvant être coûteux en temps machine
 - variante : approximation par la méthode de Monte-Carlo
 - ex : test de Fisher



Tests paramétriques et non-paramétriques

- Tests paramétriques
 - supposent que les variables suivent une loi particulière (normalité, homoscedasticité)
 - parfois plus puissants que des tests non-paramétriques, mais rarement beaucoup plus
 - ex : test de Student, ANOVA
- Tests non-paramétriques
 - ne supposent pas que les variables suivent une loi particulière
 - se fondent souvent sur les rangs des valeurs des variables plutôt que sur les valeurs elles-mêmes
 - peu sensibles aux valeurs aberrantes
 - à privilégier avec de petits effectifs (< 10)
 - par définition, les tests d'adéquation à une loi (ex : tests de normalité) sont non-paramétriques
 - ex : test de Wilcoxon, test de Kruskal-Wallis



Exemple du test du χ^2

- Test utilisé pour s'assurer que la distribution d'une variable suit une loi probabiliste donnée, en comparant la distribution observée (d'effectifs $\{O_i\}_i$) et la distribution théorique (d'effectifs $\{T_i\}_i$).
- Hypothèse H_0 : {effectifs observés = effectifs théoriques}
- Si H_0 est vraie, alors chaque élément $(O_i - T_i) / T_i$ tend vers une loi normale centrée réduite (d'après le théorème central-limite) lorsque l'effectif théorique T_i est assez grand (traditionnellement $n \geq 5$)
- Donc, si H_0 est vraie et si $T_i \geq 5$ pour tout i , la quantité $\sum_i (O_i - T_i)^2 / T_i$ suit une loi qui est une somme de p carrés de lois $N(0,1)$ indépendantes : une telle loi est dite « du χ^2 » à p degrés de liberté
- Si $T_i < 5$ pour au moins un i , préférer un test exact

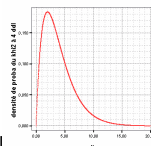
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

47

Le test du χ^2 appliqué à la liaison entre variables catégorielles

- Le test du χ^2 est souvent utilisé pour tester l'indépendance de deux variables catégorielles X et Y
- Si X et Y sont indépendantes (H_0), alors, pour tous i et j :
 - nombre d'individus tels que $\{X=i \text{ et } Y=j\} =$
 - nb d'individus tq $\{X=i\} \times \text{nb d'individus tq } \{Y=j\} \times 1/N$
- où N est le nombre total d'individus
- En notant O_{ij} le terme à gauche de l'égalité ci-dessus, et T_{ij} le terme de droite, le test d'indépendance de X et Y est le test du χ^2 appliqué à la statistique $\sum_{ij} (O_{ij} - T_{ij})^2 / T_{ij}$
- Nb de d° de liberté $p = (\text{nb lignes} - 1) \times (\text{nb colonnes} - 1)$



18/04/2005

© Stéphane Tufféry - I

J - <http://data.mining.free.fr>

48

Le test exact de Fisher

- Pour la loi du χ^2 à p degrés de liberté, on a : variance = $2p$, moyenne = p et mode = $p - 2$, et cette loi tend vers $N(p, \sqrt{2p})$ si $p > 30$
- Remarque : pas d'hypothèse sur la loi théorique suivie $\Rightarrow$ le test du χ^2 est non-paramétrique
- En revanche, ce n'est qu'asymptotiquement que $\sum_i (O_i - T_i)^2 / T_i$ suit une loi du χ^2 : si les effectifs théoriques sont faibles (< 5), il faut faire un calcul exact de probabilité
- Le test exact remplaçant le χ^2 pour un tableau de contingence $n \times n$ est appelé test de Fisher et fait appel à la loi hypergéométrique

Utilisation de la loi hypergéométrique

- Soient A et B deux variables à 2 modalités
- Elles ont un tableau de contingence 2×2 dont les effectifs seront notés a, b, c, d
- Si A et B sont indépendantes, la probabilité d'avoir un tableau (a, b, c, d) de marges fixées $a+c, b+d, a+b, c+d$, est donnée par la loi hypergéométrique

$$P(a, b, c, d) := \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!(a+b+c+d)!}$$

- Plus le tableau (a, b, c, d) s'éloigne de l'indépendance, plus la probabilité $P(a, b, c, d)$ est petite
- La p-value du test exact de Fisher = probabilité de la table observée (a, b, c, d) + probabilité de chaque table présentant un plus grand écart à l'indépendance que (a, b, c, d)

Loi hypergéométrique : exemple

- Soit le tableau suivant de marges (10,9,9,10) :

2	7
8	2

- La p-value du test unilatéral vaut :
 - $P(2,8,7,2) + P(1,9,8,1) + P(0,10,9,0) = 0,01754 + 0,00097 + 0,00001 = 0,01852$
- La p-value du test bilatéral vaut :
 - $P(2,8,7,2) + P(1,9,8,1) + P(0,10,9,0) + P(8,2,1,8) + P(9,1,0,9) = 0,01852 + 0,00438 + 0,00011 = 0,02301$

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	6,343 ^b	1	,01179		
Correction pour la continuité	4,237	1	,03955		
Test exact de Fisher				,02301	,01852
Nombre d'observations valides	19				

a. Calculé uniquement pour un tableau 2x2

b. 3 cellules (75,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 4,26.

18/04/2005

.fr

51

Loi hypergéométrique : exemple (suite)

- À noter que sur cet exemple, au seuil α de 2 %, le test du χ^2 rejette l'hypothèse H_0 , tandis que le test exact de Fisher ne rejette pas l'hypothèse d'indépendance
 - on retrouve ici la puissance du test du χ^2 qui lui permet d'être (trop ?) sensible à des écarts faibles à l'indépendance
 - on retrouve le même phénomène avec d'autres tests comme celui de Wilcoxon, dont la p-value exacte est > p-value asymptotique
- Avantage du test exact : existe en version unilatérale, contrairement à χ^2
- Correction de continuité de Yates
 - utile pour de petits effectifs, où le test du χ^2 corrigé se rapproche du test exact de Fisher
 - se rapproche du χ^2 asymptotique quand les effectifs

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

52

Loi hypergéométrique : cadre général

- On procède à un tirage sans remise (= simultané) de n boules dans une urne contenant n_1 boules gagnantes et n_2 boules perdantes
- La loi hypergéométrique de paramètres (n, n_1, n_2) est la loi que suit la variable aléatoire « nb de boules gagnantes tirées »
- Lien avec la proba $P(a, b, c, d)$ du tableau de contingence à marges fixées

boules gagnantes	boules perdantes	
a	b	$a+b=n$
c	d	$c+d$
$a+c = n_1$	$b+d = n_2$	$n_1 + n_2$

- Proba de tirer « a » boules gagnantes = proba d'avoir la configuration (a,b,c,d)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

53

χ^2 : Attention aux effectifs 1/2

	segm 1	segm 2	ensemble
Fréquences observées			
A	55	45	100
B	20	30	50
total	75	75	150
Fréquences attendues si la variable est indépendante du segment			
A	50	50	100
B	25	25	50
total	75	75	150
chi deux	0,08326454		

Dans la population de 150 individus, il y a 66,66% d'individus tq A

Dans le segment 1, il y a 73,33% d'individus tq A

=> le test du chi deux indique que cet écart n'est pas significatif (proba > 0,05)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

54

χ^2 : Attention aux effectifs 2/2

	segm 1	segm 2	ensemble
Fréquences observées			
A	550	450	1000
B	200	300	500
total	750	750	1500
Fréquences attendues si la variable est indépendante du segment			
A	500	500	1000
B	250	250	500
total	750	750	1500
chi deux	4,3205E-08		

Dans la population de 1500 individus, il y a 66,66% d'individus tq A

Dans le segment 1, il y a 73,33% d'individus tq A

⇒ le test du chi deux indique que cet écart est significatif (proba < 0,05)

> Quand la taille de la population augmente, le moindre écart devient significatif aux seuils usuels

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

55

V de Cramer

- V de Cramer = $\sqrt{\frac{C^2}{C_{\max}^2}}$
 - mesure directement l'intensité de la liaison de 2 variables catégorielles, sans avoir recours à une table du χ^2
 - en intégrant le nombre de degrés de liberté, par l'intermédiaire de $\chi^2_{\max}$
 - $\chi^2_{\max} = \text{effectif} \times [\min(\text{nb lignes}, \text{nb colonnes}) - 1]$

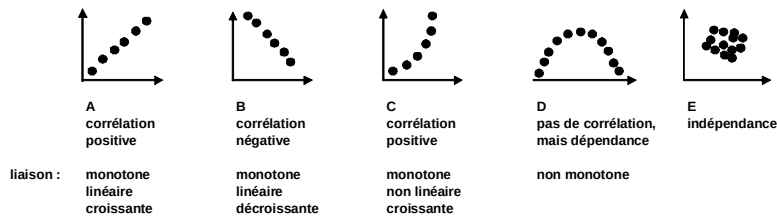
valeur de V	intensité de la relation
0	nulle
0 – 0,2	faible
0,2 – 0,4	moyenne
0,4 – 0,7	forte
0,7 – 1	très forte
1	parfaite

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

56

Liaison entre 2 variables continues : Coeff. de corrélation linéaire (Pearson)



La **liaison est nulle** si le coefficient de corrélation = 0 (nuage de points circulaire ou parallèle à un des 2 axes)

La **liaison est parfaite** si le coefficient de corrélation = +1 ou -1 (nuage de points rectiligne)

La **liaison est forte** si le coefficient de corrélation > +0,8 ou < -0,8 (nuage de points elliptique et allongé)

Mais une **liaison non linéaire** (par ex : quadratique) et surtout **non monotone** n'est pas mesurable par le coefficient de corrélation

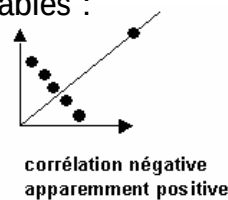
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

57

Coefficients de Pearson et Spearman

- Rho de Spearman plus général car calculé sur les rangs des valeurs et non les valeurs elles-mêmes
 - c'est un test non paramétrique (contrairement à Pearson)
- Préférer le rho de Spearman si les variables :
 - ne suivent pas une loi normale
 - ont des valeurs extrêmes
 - ne sont pas continues mais ordinales
- ou pour détecter des liaisons monotones non linéaires
- Comparer r de Pearson et ρ de Spearman :
 - $r > \rho \Rightarrow$ présence de valeurs extrêmes
 - $\rho > r \Rightarrow$ liaison non linéaire non détectée par Pearson
 - ex : $x = 1, 2, 3, \dots$ et $y = e^1, e^2, e^3, \dots$



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

58

Multicolinéarité (corrélation multiple)

- Certaines techniques (ADL, régression logistique) sont sensibles à la colinéarité des variables explicatives
- Il ne suffit pas de vérifier les variables 2 à 2
- Tolérance d'une variable = proportion de la variance non expliquée par les autres variables - doit être $> 0,1$
 - VIF (variable inflation factor) = $1 / \text{tolérance}$
- Indices de conditionnement de la matrice des corrélations
 - multicolinéarité modérée (forte) si des indices $\eta_k > 10$ (30)

		Indice de conditionnement		Proportions de la variance						
	Valeur propre									
		(cste)	var 1	var 2	var 3	var 4	var 5	var 6		
1	3.268	1,000	,01	,00	,03	,02	,01	,01	,02	
2	1.022	1.788	,00	,56	,01	,02	,00	,33	,00	
3	,976	1.830	,00	,42	,00	,10	,00	,42	,01	
4	,811	2.008	,00	,02	,07	,81	,00	,14	,00	
5	,636	2.266	,01	,00	,78	,04	,02	,09	,00	
6	,221	3.842	,01	,00	,11	,01	,20	,00	,73	
7	,065	7.099	,97	,00	,00	,00	,76	,00	,24	



UNIVERSITÉ
DE BURGUNDIE
DIJON

18/04/2005

59

18/04/2005

3.fr

59

Généralisation à des groupes de variables 1/2

- Analyse de corrélation canonique (linéaire)
 - non plus entre $n (\geq 2)$ ensembles composés chacun de 1 variable continue
 - mais entre $n (\geq 2)$ ensembles $\{U_i\}, \{V_j\}...$ de plusieurs variables continues ou binaires
 - on cherche les combinaisons linéaires (variables canoniques) maximisant la corrélation entre $\sum_i \lambda_i U_i, \sum_j \mu_j V_j...$
 - proc CANCORR de SAS (si $n = 2$) – proc OVERALS de SPSS
 - origine : Harold Hotelling (1936)
 - généralise la régression multiple
 - $n = 2$ et l'un des ensembles de variables est composé d'une seule variable

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

60

Généralisation à plus de 2 variables

- Analyse de corrélation canonique (linéaire)
 - généralise aussi l'analyse discriminante linéaire
 - $n = 2$ et l'un des ensembles de variables est composé des indicatrices de la partition à discriminer
 - généralise aussi l'analyse factorielle des correspondances
 - $n = 2$ et chaque ensemble de variables est composé des indicatrices d'une variable catégorielle
- Analyse de corrélation canonique (non linéaire)
 - entre $n (\geq 2)$ ensembles de variables quelconques
 - permet la détection d'effets non linéaires
 - proc OVERALS de SPSS

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

61

Liaison entre 1 variable continue et 1 variable catégorielle

lois suivies	2 échantillons	3 échantillons et plus (***)
normalité – homoscedasticité (*)	test T de Student	ANOVA
normalité – hétéroscédasticité	test T de Welch	Welch - ANOVA
non normalité – hétéroscédasticité (**)	Wilcoxon – Mann – Whitney	Kruskal – Wallis
non normalité – hétéroscédasticité (**)	test de la médiane	test de la médiane
non normalité – hétéroscédasticité (**)		test de Jonckheere-Terpstra (échantillons ordonnés)

moins puissant

(*) Ces tests supportent mieux la non-normalité que l'hétéroscédasticité.
 (**) Ces tests travaillant sur les rangs et non sur les valeurs elles-mêmes, ils sont plus robustes et s'appliquent également à des variables ordinales
 (***) ne pas comparer toutes les paires par des tests T $\Rightarrow$ on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 moyennes égales)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

62

Test ANOVA à 1 facteur

- Test d'égalité de la moyenne d'une variable continue Y dans k (≥ 2) groupes (définis par les modalités d'une variable nominale)
 - si plusieurs variables continues dépendantes $\Rightarrow$ MANOVA
 - si m variables nominales indépendantes $\Rightarrow$ ANOVA à m facteurs
- Généralise le test de Student quand $k > 2$
- Ne teste que l'égalité de toutes les moyennes, sans dire le cas échéant lesquelles diffèrent
- Exemples :
 - comparer les productivités de plusieurs usines
 - comparer les rendements de plusieurs champs
 - comparer les effets de plusieurs engrais

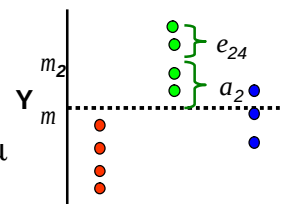
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

63

ANOVA à 1 facteur : modèle général

- $Y_{ij} = \mu + \alpha_i + e_{ij}$
- Y_{ij} = valeur de l'obs. j dans le groupe i
- μ = moyenne générale de Y
- α_i = moyenne de Y dans le groupe i $- \mu$
- e_{ij} = valeur résiduelle
 - distribution normale dans tous les groupes (hypothèse la moins importante pour la qualité du test)
 - moyenne = 0
 - variance égale dans tous les groupes (homoscédasticité)
 - indépendance $\forall i, j$
 - une observation ne doit pas dépendre des autres du groupe
 - les observations d'un groupe ne doivent pas dépendre de celles des autres groupes (cas d'un même individu présent plusieurs fois – cas de la comparaison de traitements)



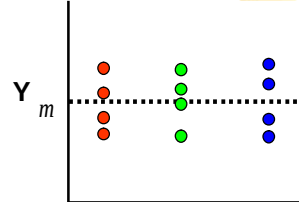
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

64

Hypothèses de l'ANOVA

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
 - les moyennes sont toutes égales
 - $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$



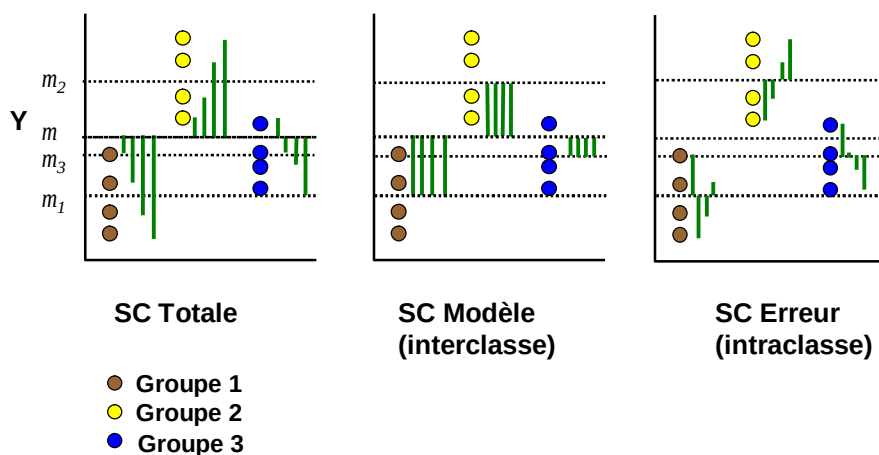
- H_1 : les moyennes ne sont pas toutes égales
 - au moins une moyenne est différente
 - ne signifie pas : $\mu_1 \neq \mu_2 \neq \dots \neq \mu_k$
 - pour déterminer quelles moyennes diffèrent significativement :
 - test de Bonferroni
 - test de Scheffé (plus puissant)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

65

Répartition de la somme des carrés



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

66

Tableau ANOVA et statistique F

Source de variation	Somme des carrés (SC)	Degrés de liberté (dl)	Carré moyen (CM)	F
Totale	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2$	$n - 1$	SC/dl	
Inter-classe	$\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2$	$k - 1$	SC/dl	$\frac{CM_{interclasse}}{CM_{intraclasse}}$
Intra-classe	$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$n - k$	SC/dl	

$CM_{inter}/CM_{intra} = F$ à comparer au F d'une loi de Fisher de ddl (k-1, n-k)
 $\eta^2 = SC_{interclasse} / SC_{totale}$ = proportion de la variance expliquée

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

67

Principe du test ANOVA

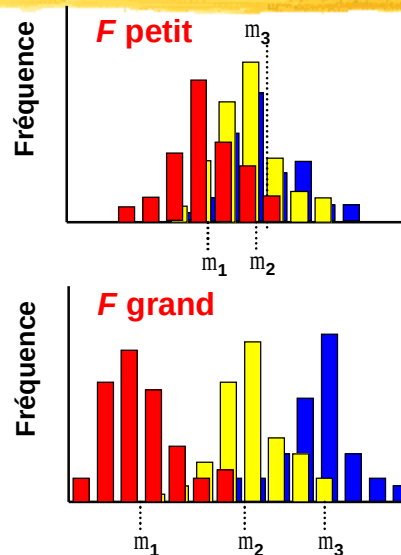
- On appelle « analyse de la variance » ce qui est en fait un test d'égalité de la moyenne, en raison de la façon de réaliser ce test, qui consiste à décomposer la variance de la variable continue Y en 2 parties :
 - ce qui peut être attribué aux différences entre groupes (variance inter-classe)
 - ce qui peut être attribué aux variations aléatoires (variance intra-classe, appelée « erreur »)
- Si CM_{inter}/CM_{intra} est grand, c.a.d. si les variations aléatoires sont faibles par rapport à l'effet des différences entre classes, on peut rejeter H_0
- Cela se produit quand CM_{inter}/CM_{intra} dépasse la valeur critique de la loi de Fisher au niveau α avec k-1 et n-k degrés de liberté

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

68

Illustration du test ANOVA



18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

69

Test non-paramétrique de Wilcoxon-Mann-Whitney

- Utilisé pour $k = 2$ groupes, d'effectifs n_1 et n_2
 - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soit R_i = somme des rangs des observations du groupe i
- La statistique du test comparée à une valeur théorique est:

$$U = \min \left\{ n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \right\}$$

- Avec les observations des 2 groupes G_1 et G_2 :

G_1 : 0.45 0.65 1.06 0.97

G_2 : 1.44 2.21 0.50 1.63

- On obtient les rangs

G_1 : 1 3 5 4 G_2 : 6 8 2 7

- D'où $R_1 = 13$, $R_2 = 23$ et $U = 16 + 10 - 23 = 3$

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

70

Test non-paramétrique de Kruskal-Wallis

- Utilisé pour $k \geq 2$ groupes
 - quand les hypothèses de normalité et d'égalité des variances ne sont pas satisfaites
- Soient N = nb d'observations, n_i l'effectif du groupe i et R_i la somme des rangs des observations du groupe i
- La statistique du test est :
$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$
- Correctif à apporter en cas d'égalités de rangs
- Si les effectifs sont grands ou si $k > 6$, H tend vers χ^2 à $k-1$ d° de liberté
 - sinon, regarder valeurs critiques dans une table
- Presque aussi puissant qu'ANOVA sur son propre terrain

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

71

Tests non-paramétriques sur SAS

- La PROC NPAR1WAY de SAS permet d'effectuer le test de Kruskal-Wallis et de Wilcoxon-Mann-Whitney (si $k = 2$)
 - **PROC NPAR1WAY WILCOXON** data=table correct=no;
 - **class** a; /* variable de groupe */
 - **var** x; /* variable quantitative */
 - **exact**; /* test exact facultatif */
 - **run**;
- Autres options que WILCOXON :
 - ANOVA : anova classique
 - EDF : tests de Kolmogorov-Smirnov et Cramer-von Mises, et, si 2 niveaux de classification seulement, statistique de Kuiper
 - MEDIAN : test de la médiane

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

72

Résultats des tests avec option WILCOXON

Wilcoxon Two-Sample Test	
Statistic (S)	27.000 0
Normal Approximation	
Z	1.7146
One-Sided Pr > Z	0.0432
Two-Sided Pr > Z	0.0864
t Approximation	
One-Sided Pr > Z	0.0624
Two-Sided Pr > Z	0.1248
Exact Test	
One-Sided Pr >= S	0.0556
Two-Sided Pr >= S - Mean	0.1111

Wilcoxon Scores (Rank Sums) for Variable x Classified by Variable a					
a	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	5	18.0	25.0	4.082483	3.600
2	4	27.0	20.0	4.082483	6.750

Kruskal-Wallis Test	
Chi-Square	2.9400
DF	1
Pr > Chi-Square	0.0864

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

73



La préparation des données : Analyse factorielle des correspondances

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

74



L'analyse des correspondances

- **L'analyse factorielle des correspondances (AFC)** (Guttman, 1941 – Benzécri, 1973) offre une visualisation en 2 dimensions des tableaux de contingence :
 - deux modalités liées positivement (sureffectif) sont proches
 - deux modalités liées négativement (sous-effectif) sont opposées
 - on utilise une métrique (du χ^2) pondérant ces sur- ou sous-effectifs par l'inverse de la fréquence globale d'une modalité (non croisée avec la modalité de l'autre variable)
 - les + fortes oppositions sont sur l'axe horizontal
 - les modalités non liées aux autres sont au centre
- **L'analyse des correspondances multiples (ACM)** s'applique à plus de 2 variables catégorielles
 - en utilisant le tableau de contingence multiple (tableau de Burt)
 - ou le tableau disjonctif

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

75



Tableaux utilisés en ACM

- **Tableau de contingence multiple (= tableau de Burt)**
 - les lignes et les colonnes correspondent aux modalités des variables
 - croisement des 2 mêmes modalités : nb d'individus possédant cette modalité
 - croisement de 2 modalités différentes, appartenant à des variables différentes : nb d'individus possédant à la fois la 1^{ère} et la 2^{de} modalité
- **Tableau disjonctif**
 - une ligne par individu
 - une colonne par modalité x_{jk} de variable X_j
 - croisement de la i^e ligne et de la colonne correspondant à la modalité x_{jk} : 1 si $X_j(\text{i^e individu}) = x_{jk}$, et 0 sinon

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

76



Intérêt de l'AFC et l'ACM 1/2

- L'AFC et l'ACM s'appuient sur un instrument puissant de détection des groupes d'individus et des individus isolés : l'œil. La perception visuelle est inégalable pour la reconnaissance des formes. Examiner un plan factoriel est :
 - + efficace que d'examiner tous les plans (x,y) des variables d'origine
 - + rapide que dépouiller tous les tableaux de contingence
- Représentation simultanée des individus et des modalités
 - sur un même plan
- Possibilité de visualiser certaines variables (« supplémentaires ») sans les prendre en compte dans le calcul des correspondances :
 - variables qu'on veut lier aux variables actives mais pas lier entre elles
 - ou variables qu'on veut expliquer par les variables actives

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

77

Intérêt de l'AFC et l'ACM 2/2

- Détection des liaisons entre les variables
 - même non linéaires (liaisons de $d^{\circ} > 1$)
 - pas forcément visibles sur un tableau de contingence
- Filtrage des fluctuations aléatoires des données
 - en remplaçant les variables d'origine par les 1^{ers} axes factoriels
 - utile avant une classification ou un réseau de neurones
- Transformation de var. qualitatives en var. quantitatives
 - en remplaçant les variables qualitatives par les coordonnées de leurs modalités sur les axes factoriels
 - transformation utilisée dans l'analyse discriminante DISQUAL
- Permet de traiter simultanément les variables quantitatives (en les discrétisant) et qualitatives

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

78

AFC : signification des valeurs propres

- L'inertie du nuage de points mesure la dispersion du nuage et la liaison entre les 2 variables. On a :
 - inertie x effectif = χ^2 du tableau
 - inertie = somme des valeurs propres $\neq 0$
- Si on a k valeurs propres proches de 1 => le nuage de points (resp. tableau) est scindé en k+1 groupes (blocs)
- Si toutes les valeurs propres sont ≈ 1 , chaque modalité de V_1 correspond (presque) à une seule modalité de V_2
- Si V_1 et V_2 sont partiellement corrélées (« vont dans le même sens »), le tableau peut être réordonné en un tableau avec des 0 partout sauf dans une bande entourant la diagonale, et le nuage de points a une forme parabolique dite « en fer à cheval » (effet Guttman)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

79

Représentation graphique 1/2

- Propriété de l'analyse des correspondances : on peut superposer individus et modalités sur le même graphique
 - le point représentant une modalité est le barycentre des individus possédant cette modalité
- Le centre correspond au profil moyen pour les 2 variables
- Une modalité A est loin du centre si la distribution des modalités de l'autre variable est très différente dans l'ensemble des individus vérifiant A et dans l'ensemble de tous les individus

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

80

Représentation graphique 2/2

- Proximité de 2 modalités A_1 et A_2 de la même variable
 - si l'autre variable a la même distribution restreinte à A_1 et A_2
- Proximité de 2 modalités A et B de 2 variables différentes
 - si sureffectif de l'intersection $A \times B$
 - NB : ceci est souvent vrai mais pas toujours
- Avant d'interpréter la proximité de 2 modalités, s'assurer qu'elles sont bien représentées sur le plan factoriel



Interprétation des résultats 1/2

- Sélectionner les axes factoriels dont la valeur propre :
 - est $\neq$ de 0 et 1 (critère obligatoire)
 - est $> 1/p$ (p = nb de variables)
 - correspond à un coude dans le diagramme des valeurs propres
 - on dépasse rarement les 5 premiers axes
- Coordonnée d'une modalité sur un axe
 - on s'intéresse parfois aux modalités dont la coordonnée $>$ racine carrée (valeur propre de l'axe)
- Contribution d'une modalité à un axe
 - % de l'inertie de la modalité dans l'inertie de l'axe
 - on explique un axe par ses modalités à forte contribution



Interprétation des résultats 2/2

- Cosinus² (qualité de la projection sur un axe)
 - % pris par l'axe dans la dispersion de la modalité
 - plus $\cos^2$ est proche de 1, meilleure est la représentation d'une modalité sur un axe
 - il faut n'apprécier la proximité de 2 modalités sur un axe que si elles ont un $\cos^2$ assez grand sur cet axe
 - la somme des $\cos^2$ sur l'ensemble des axes vaut 1
- Variables supplémentaires
 - ne contribuent pas aux axes
 - mais leurs $\cos^2$ peuvent être analysés



Les 2 types de fichiers traités par la proc CORRESP de SAS 1/2

- Type 1 : le fichier contient des données « brutes »
 - format le plus courant dans les bases de données
 - chaque ligne du fichier correspond à un individu
 - chaque colonne correspond à une variable catégorielle caractérisant les individus
 - chaque intersection « ligne x colonne » contient la modalité de la variable pour l'individu
 - utiliser l'instruction « TABLES »
 - qui crée un tableau de contingence, de Burt ou disjonctif selon les options
 - puis effectue l'AFC ou l'ACM



Les 2 types de fichiers traités par la proc CORRESP de SAS 2/2

- Type 2 : le fichier contient des données « matricielles »
 - les données sont déjà sous forme d'un tableau de contingence, de Burt ou disjonctif
 - chaque ligne du fichier correspond à une ligne du tableau
 - chaque colonne correspond à une modalité A_i spécifiée dans l'instruction « VAR »
 - chaque intersection « ligne x colonne » contient l'effectif d'une cellule du tableau
 - cette intersection est une donnée numérique ≥ 0
 - si < 0 ou manquante $\Rightarrow$ la ligne n'entre pas dans l'analyse
 - utiliser l'instruction « VAR »
 - qui effectue directement l'AFC ou l'ACM
 - les lignes Y sont spécifiées par l'instruction ID Y
 - sauf pour le tableau de Burt qui est symétrique

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

85

Traitement des données matricielles

- VAR $A_1 \dots A_m$;
- ID Y;
 - avec un tableau de Burt $\Rightarrow$ utiliser l'option MCA et spécifier le nb de variables par NVAR = n
 - avec un tableau disjonctif ou de contingence, on ne spécifie pas d'option (ni MCA ni BINARY)
 - pour un tableau disjonctif ou de contingence, les lignes sont identifiées par la variable Y
 - Y = variable en ligne pour un tableau de contingence
 - Y = label pour un tableau disjonctif
 - pas d'instruction ID Y possible pour un tableau de Burt
 - car dans ce cas les lignes et les colonnes sont les mêmes et sont toutes spécifiées dans l'instruction VAR
- Variables supplémentaires
 - à énumérer dans l'instruction SUPPLEMENTARY
 - à énumérer aussi dans l'instruction VAR

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

86

Exemple

Nom	Sexe	État
M. Dupont	Homme	Marié
Mme Dupont	Femme	Marié
M. Durand	Homme	Célibataire
M. Dubois	Homme	Marié
Mme Dubois	Femme	Marié
Mme Martin	Femme	Veuf

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

87

Exemple de syntaxe SAS sur tableau de Burt (donné)

- `proc corresp data=sasuser.tableau mca nvars=2 out=sortie all;`
- `var homme femme marié célibataire veuf;`
- `run;`

	Femme	Homme	Marié	Célibataire	Veuf
Femme	3	0	2	0	1
Homme	0	3	2	1	0
Marié	2	2	4	0	0
Célibataire	0	1	0	1	0
Veuf	1	0	0	0	1

- NVARS = 2 indique que les 5 modalités du tableau de Burt proviennent ici de 2 variables
- Attention : les modalités doivent être listées dans le même ordre que dans le fichier

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

88

Contenu du fichier en sortie

- Le fichier OUT contient des résultats (inertie, coordonnées sur les axes factoriels, $\cos^2$...) sur :
 - les modalités de variables suivant la commande VAR (_type_ = « VAR » dans le fichier)
 - les lignes correspondant à l'instruction ID (_type_ = « OBS »), sauf pour le tableau de Burt qui n'a pas d'enregistrements de _type_ = « OBS »
 - seul le tableau disjonctif (ID LABEL) permet d'obtenir des résultats sur les individus
 - mais les résultats sur les modalités (_type_ = « VAR ») sont les mêmes avec un tableau disjonctif ou de Burt
 - les coordonnées contenues dans le fichier OUT permettent de représenter sur un plan factoriel les modalités et éventuellement (si tableau disjonctif) les individus

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

89



Exemple de syntaxe SAS sur tableau disjonctif complet (donné)

- **proc** **corresp** **data**=sasuser.tableau
out=sortie **all**;
- **var** homme femme marié célibataire veuf;
- **id** label;
- **run**;

	Femme	Homme	Marié	Célibataire	Veuf
1	0	1	1	0	0
2	1	0	1	0	0
3	0	1	0	1	0
4	0	1	1	0	0
5	1	0	1	0	0
6	1	0	0	0	1

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

90



Exemple de syntaxe SAS sur tableau de contingence (donné)

- **proc** **corresp** **data**=sasuser.tableau
out=sortie **all**;
- **var** marié célibataire veuf;
- **id** sexe;
- **run**;

	marié	célibataire	veuf
femme	2	0	1
homme	2	1	0

Traitement des données brutes

- TABLES $X_1 \dots X_m, Y_1 \dots Y_n$;
 - crée et analyse le tableau de contingence simple des variables $X_1 \dots X_m$ (en lignes) par $Y_1 \dots Y_n$ (en colonnes)
 - ne pas oublier la virgule séparant les variables !
 - cas particulier de l'AFC : TABLES X , Y
- TABLES $X_1 \dots X_m$;
 - crée et analyse le tableau de Burt de $X_1 \dots X_m$ (option « MCA » : multiple correspondence analysis)
 - ou crée et analyse le tableau disjonctif (option « BINARY »)
- Variables X_i et Y_i catégorielles
- Variables supplémentaires
 - à énumérer dans l'instruction SUPPLEMENTARY
 - à énumérer aussi dans l'instruction TABLES

Exemple de syntaxe SAS sur tableau de Burt ou disjonctif (créé)

- `proc corresp data=sasuser.enquete mca
dimens=5 out=sortie all;`
- `tables sexe age frequence rayon
satisfaction;`
- `supplementary satisfaction;`
- `run;`
- `%plotit(data=sortie,datatype=corresp,
plotvars=dim2 dim1)`

nb d'axes
demandés (2
par défaut)

Fichier contenant pour les modalités en colonnes (_type_ = « var ») et éventuellement en lignes (_type_ = « obs ») du tableau étudié :

- leur inertie
- leurs coordonnées sur les axes
- leurs contributions aux axes
- leurs cos²

mca : tableau de Burt
binary : tableau disjonctif

all : tout afficher - short : affichage minimal
observed : affichage du tableau analysé

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

93

Contenu du fichier en sortie

- Le fichier OUT contient des résultats (inertie, coordonnées sur les axes factoriels, cos²...) sur :
 - les modalités de variables suivant la commande TABLES, à l'exception de celles à gauche de la virgule le cas échéant (_type_ = « VAR » dans le fichier)
 - les individus en cas de tableau disjonctif (option BINARY) (_type_ = « OBS » dans le fichier)
 - les modalités des variables en ligne (à gauche de la virgule) en cas de tableau de contingence (_type_ = « OBS »)
 - pas d'enregistrement de _type_ = « OBS » en cas de tableau de Burt (option MCA)
 - mais les résultats sur les modalités (_type_ = « VAR ») sont les mêmes avec un tableau disjonctif ou de Burt
 - les coordonnées contenues dans le fichier OUT permettent de représenter sur un plan factoriel les modalités et éventuellement (si tableau disjonctif) les individus

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

94

Diagrammes des valeurs propres

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	
					2 4 6 8 10 ----- ----- ----- ----- -----
0.56420	0.31833	759.10	8.49	8.49	*****
0.56043	0.31408	748.97	8.38	16.86	*****
0.53721	0.28860	688.21	7.70	24.56	*****
0.52911	0.27996	667.61	7.47	32.03	*****
0.52401	0.27459	654.80	7.32	39.35	*****
0.51099	0.26111	622.66	6.96	46.31	*****
0.50151	0.25151	599.77	6.71	53.02	*****
0.49404	0.24408	582.04	6.51	59.53	*****
0.49052	0.24061	573.77	6.42	65.94	*****
0.48516	0.23538	561.31	6.28	72.22	*****
0.47999	0.23039	549.40	6.14	78.36	*****
0.47222	0.22299	531.75	5.95	84.31	*****
0.46499	0.21622	515.61	5.77	90.08	*****
0.44178	0.19517	465.41	5.20	95.28	*****
0.42071	0.17700	422.09	4.72	100.00	*****
Total	3.75000	8942.50	100.00		
Degrees of Freedom = 324					5

Coordonnées des modalités sur les axes

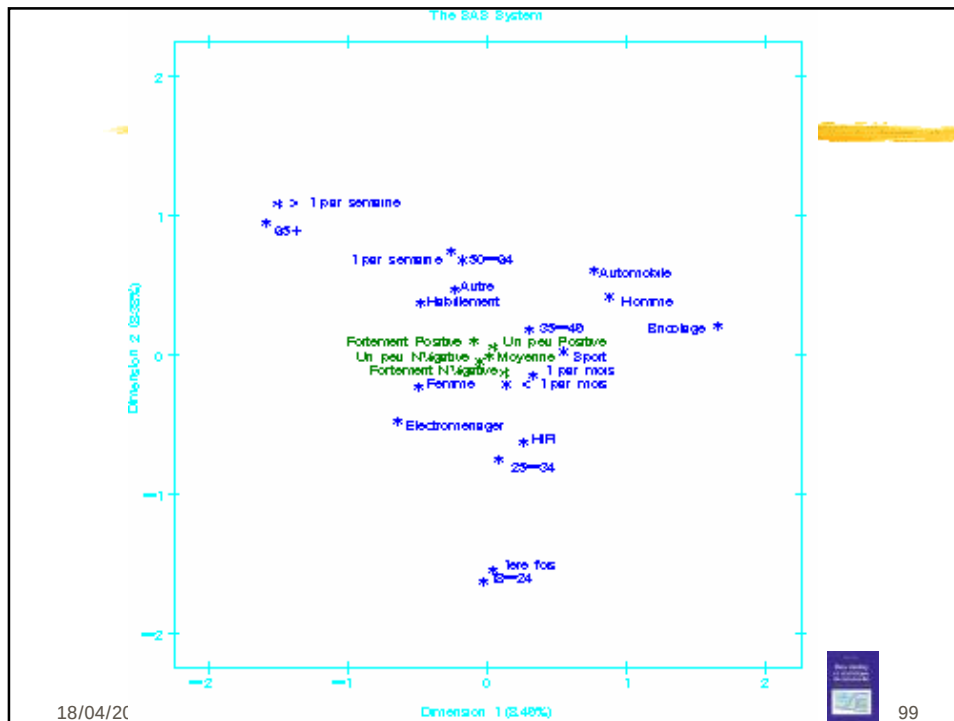
Column Coordinates					
	Dim1	Dim2	Dim3	Dim4	Dim5
Femme	-0.4925	-0.2347	-0.1592	-0.1355	-0.1049
Homme	0.8789	0.4189	0.2842	0.2418	0.1872
18-24	-0.0312	-1.6313	0.0627	1.0440	1.5625
25-34	0.0824	-0.7573	0.6018	-0.1037	-0.7778
35-49	0.2980	0.1770	-0.5109	-0.0344	-0.2147
50-64	-0.1814	0.6814	0.2733	-0.5462	0.6680
65+	-1.5907	0.9484	-0.0621	1.6670	-0.6844
1 par mois	0.3271	-0.1452	-0.7581	0.1949	-0.0720
1 par semaine	-0.2645	0.7421	-0.0498	-0.8008	0.4807
1ere fois	0.0447	-1.5486	0.3944	0.1858	0.6363
< 1 par mois	0.1341	-0.2131	0.7872	0.0211	-0.6013
> 1 par semaine	-1.5009	1.0865	0.5446	1.8132	0.1506
Automobile	0.7695	0.6039	-0.1511	0.6080	-0.2946
Autre	-0.2313	0.4674	0.5169	-0.6090	-1.1817
Bricolage	1.6517	0.2073	-0.5205	-0.1879	0.4933
Electromenager	-0.6415	-0.4791	-0.0306	-0.7177	0.5409
HIFI	0.2628	-0.6295	-0.7936	0.0237	-0.6002
Habillement	-0.4837	0.3713	-0.4197	0.9038	0.2133
Sport	0.5439	0.0242	1.7960	0.5286	0.6428

Contribution des modalités aux axes

Partial Contributions to Inertia for the Column Points					
	Dim1	Dim2	Dim3	Dim4	Dim5
Femme	0.1221	0.0281	0.0141	0.0105	0.0064
Homme	0.2179	0.0502	0.0251	0.0188	0.0115
18-24	0.0001	0.1674	0.0003	0.0769	0.1757
25-34	0.0012	0.0996	0.0685	0.0021	0.1202
35-49	0.0276	0.0099	0.0894	0.0004	0.0166
50-64	0.0065	0.0933	0.0163	0.0673	0.1026
65+	0.1093	0.0394	0.0002	0.1364	0.0234
1 par mois	0.0290	0.0058	0.1720	0.0117	0.0016
1 par semaine	0.0134	0.1070	0.0005	0.1397	0.0513
1ere fois	0.0001	0.1705	0.0120	0.0028	0.0329
< 1 par mois	0.0037	0.0095	0.1411	0.0001	0.0866
> 1 par semaine	0.1034	0.0549	0.0150	0.1715	0.0012
Automobile	0.0384	0.0239	0.0016	0.0272	0.0065
Autre	0.0061	0.0254	0.0338	0.0484	0.1857
Bricolage	0.1730	0.0028	0.0190	0.0025	0.0179
Electromenager	0.0811	0.0458	0.0002	0.1154	0.0668
HIFI	0.0077	0.0450	0.0778	0.0001	0.0468
Habillement	0.0360	0.0215	0.0299	0.1429	0.0081
Sport	0.0236	0.0000	0.2833	0.0253	0.0381

Qualité des projections sur les axes

Squared Cosines for the Column Points					
	Dim1	Dim2	Dim3	Dim4	Dim5
Femme	0.4328	0.0983	0.0452	0.0328	0.0196
Homme	0.4328	0.0983	0.0452	0.0328	0.0196
18-24	0.0001	0.2284	0.0003	0.0935	0.2095
25-34	0.0019	0.1601	0.1011	0.0030	0.1689
35-49	0.0580	0.0205	0.1706	0.0008	0.0301
50-64	0.0111	0.1569	0.0252	0.1008	0.1508
65+	0.1472	0.0523	0.0002	0.1617	0.0272
1 par mois	0.0564	0.0111	0.3032	0.0200	0.0027
1 par semaine	0.0226	0.1777	0.0008	0.2069	0.0746
1ere fois	0.0002	0.2353	0.0153	0.0034	0.0397
< 1 par mois	0.0064	0.0162	0.2210	0.0002	0.1290
> 1 par semaine	0.1398	0.0732	0.0184	0.2040	0.0014
Automobile	0.0532	0.0328	0.0021	0.0332	0.0078
Autre	0.0092	0.0374	0.0457	0.0634	0.2388
Bricolage	0.2397	0.0038	0.0238	0.0031	0.0214
Electromenager	0.1378	0.0769	0.0003	0.1725	0.0980
HIFI	0.0115	0.0659	0.1047	0.0001	0.0599
Habillement	0.0570	0.0336	0.0429	0.1990	0.0111
Sport	0.0334	0.0001	0.3639	0.0315	0.0466



Autres fonctionnalités de la proc CORRESP

- Gestion des valeurs manquantes (MISSING)
- Standardisation des variables (PROFILE)
- Création de croisements de modalités de variables (CROSS)
 - exemple : age * situation de famille
- Pondérations d'observations (WEIGHT)
 - impossible à utiliser avec VAR et MCA (le fichier en entrée est un tableau de Burt)
 - une observation avec un poids < 0 est une observation supplémentaire

La préparation des données : Analyse en composantes principales

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



101

But de l'ACP (analyse en composantes principales)

- À partir de n variables initiales continues, construire m ($\leq n$) autres variables, appelées *composantes principales*, combinaisons linéaires des variables initiales, tq :
 - les CP sont ordonnées selon l'information (variance) qu'elles restituent, la 1ère étant celle qui restitue le plus d'information
 - on sait quelle part d'information restitue chaque CP, et des critères permettent de décider combien de CP il est pertinent de conserver
 - les CP sont des vecteurs indépendants, c'est-à-dire des variables non corrélées entre elles
 - on a une inégalité stricte $m < n$ s'il existe des relations linéaires entre les variables initiales
- Origine : Karl Pearson (1901) – Harold Hotelling (1933)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



102

Intérêt de l'ACP

- Représentation assez fidèle des individus d'une population en 2 ou 3 dimensions
- Localisation des grandes masses d'individus
- Détection des individus exceptionnels et d'éventuels groupes isolés d'individus
- Détection des liaisons entre les variables
- Outil de réduction des dimensions d'un problème
 - diminuer le nombre de variables étudiées sans perdre beaucoup d'information
 - utile avant un réseau de neurones ou une classification
- L'ACP sur les indicatrices de variables nominales conduit aux mêmes résultats que l'ACM sur ces var. nominales
 - > ACM $\hat{=}$ ACP quand les variables sont binaires

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

103

Métrie dans l'espace des individus

- Métrie euclidienne
 - $d(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2$
- Métrie « inverse des variances »
 - $d(x,y) = ((x_1 - y_1) / \sigma_1)^2 + ((x_2 - y_2) / \sigma_2)^2 + \dots + ((x_n - y_n) / \sigma_n)^2$
 - avec σ_i = écart-type de la $i^{\text{ème}}$ variable
- Avec cette nouvelle métrie, qui revient à réduire les variables, la distance entre deux individus ne dépend plus de l'unité de mesure, et les variables les plus dispersées ne sont pas avantagées

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

104

Obtention des composantes principales 1/2

- Les composantes principales : obtenues en exprimant les variables initiales, non selon les axes d'origine de l'espace des individus, mais selon de nouveaux axes, les *axes principaux*, qui sont les vecteurs propres de la matrice
 - des covariances, lorsque c'est la métrique euclidienne qui a été choisie dans l'espace des individus
 - des corrélations lorsque les unités de mesure ne sont pas les mêmes pour toutes les variables et que l'on a choisi la métrique « inverse des variances »

$$\sigma(XY) = \sigma_X \sigma_Y r_{XY}$$

$$M_{corr} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ \dots & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}$$

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

105

Obtention des composantes principales 2/2

- Si on a des données hétérogènes, avec des ordres de grandeur différents, diagonaliser la matrice des corrélations
 - sinon, diagonaliser la matrice des covariances
- Les composantes principales sont classées par variance décroissante
- Variance d'une composante principale = valeur propre correspondante de la matrice diagonalisée
- Trace de cette matrice = Σ valeurs propres = Σ variances des composantes principales = Σ variances des variables d'origine = inertie du nuage d'individus
 - = nombre de variables pour la matrice des corrélations

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

106

Exemple de syntaxe SAS

- **proc princomp** data=sasuser.etude_de_cas out=coord outstat=stat;
- **var** age anciennete revenus nbproduits nbachats nbpoints evolconsom utilcredit;
- **weight** poids;
- **run**;
- **proc transpose** data=stat out=sortie;run;
- **proc plot** data=sortie; plot prin2*prin1=_name_ \$ _name_;run;
- Si la variable de l'instruction WEIGHT est ≤ 0 pour un individu, il n'est pas pris en compte dans les analyses, mais ses composantes principales sont calculées, ce qui permet de le représenter sur les graphiques comme point supplémentaire (illustratif)

pour utiliser la matrice des covariances : ajouter l'option « COV »

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

107

Contenu du fichier OUTSTAT

Obs	_TYPE_	_NAME_	age	anciennete	revenus	nbproduits	nbachats	nbpoints	evolconsom	utilcredit
1	MEAN		38.34	12.24	1882.08	9.23	38.99	1.51	1.14	32.88
2	STD		9.51	7.04	1529.57	4.43	22.66	0.88	0.32	112.92
3	N		5942.00	5942.00	5942.00	5942.00	5942.00	5942.00	5942.00	5942.00
4	CORR	age	1.00	0.38	0.09	0.06	-0.09	-0.08	-0.11	0.01
5	CORR	anciennete	0.38	1.00	0.16	0.22	0.15	0.08	-0.16	0.02
6	CORR	revenus	0.09	0.16	1.00	0.45	0.58	0.36	-0.04	0.16
7	CORR	nbproduits	0.06	0.22	0.45	1.00	0.52	0.62	0.06	0.21
8	CORR	nbachats	-0.09	0.15	0.58	0.52	1.00	0.54	0.03	0.15
9	CORR	nbpoints	-0.08	0.08	0.36	0.62	0.54	1.00	0.12	0.28
10	CORR	evolconsom	-0.11	-0.16	-0.04	0.06	0.03	0.12	1.00	0.00
11	CORR	utilcredit	0.01	0.02	0.16	0.21	0.15	0.28	0.00	1.00
12	EIGENVAL		2.70	1.48	0.94	0.93	0.67	0.57	0.39	0.32
13	SCORE	Prin1	0.04	0.18	0.44	0.50	0.49	0.48	0.03	0.23
14	SCORE	Prin2	0.64	0.61	0.08	0.00	-0.09	-0.19	-0.41	-0.08
15	SCORE	Prin3	0.26	0.19	-0.04	0.12	0.00	0.03	0.77	-0.54
16	SCORE	Prin4	0.31	0.06	-0.27	0.02	-0.30	0.11	0.38	0.75
17	SCORE	Prin5	0.33	-0.39	0.65	-0.31	0.13	-0.38	0.18	0.15
18	SCORE	Prin6	-0.49	0.62	0.12	-0.34	0.22	-0.31	0.23	0.22
19	SCORE	Prin7	0.26	-0.05	-0.32	-0.62	0.54	0.39	-0.03	-0.02
20	SCORE	Prin8	0.11	-0.14	-0.43	0.39	0.55	-0.57	0.01	0.10

Valeurs propres

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.69628418	1.21286011	0.3370	0.3370
2	1.48342407	0.54475789	0.1854	0.5225
3	0.93866619	0.00785658	0.1173	0.6398
4	0.93080961	0.26102963	0.1164	0.7561
5	0.66977998	0.10079635	0.0837	0.8399
6	0.56898363	0.17882698	0.0711	0.9110
7	0.39015665	0.06826097	0.0488	0.9598
8	0.32189568		0.0402	1.0000

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



109

Coordonnées des variables sur les axes (= coefficients de corrélation)

		Eigenvectors							
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
age	âge	0.035722	0.637571	0.258608	0.314857	0.332034	-.485766	0.260887	0.109258
anciennete	ancienneté client	0.182484	0.605982	0.187981	0.057860	-.389761	0.622209	-.049299	-.138971
revenus	revenus du client	0.443030	0.080052	-.040509	-.273237	0.651810	0.117833	-.318435	-.425293
nbproduits	nb produits	0.495244	0.001202	0.117938	0.022297	-.306377	-.337970	-.616976	0.389326
nbachats	nb achats	0.491426	-.093575	0.003426	-.304492	0.130701	0.220890	0.542087	0.545237
nbpoints	nb points fidélité	0.478177	-.191761	0.030312	0.110554	-.381114	-.308847	0.389827	-.573440
evokonsom	évolution consommation	0.029868	-.408820	0.770733	0.384360	0.184274	0.234905	-.032310	0.005507
utilcredit	règlements à crédit	0.229862	-.084950	-.535980	0.754779	0.151093	0.222308	-.024057	0.100762

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



110

Interprétation des résultats

- *Critère de Kaiser* : avec la matrice des corrélations (= sur des données centrées réduites), conserver les axes correspondant aux valeurs propres > 1
- Calculer les valeurs cumulées successives $\lambda_1/\Sigma\lambda_i$, $(\lambda_1+\lambda_2)/\Sigma\lambda_i...$ pour voir quelle proportion de la somme des variances $\Sigma\lambda_i$ (c.a.d quelle part de l'inertie totale) est restituée par les p premiers axes principaux
 - plus les variables sont nombreuses et moins elles sont corrélées, plus est faible la part d'inertie restituée par les p premiers axes
- Avec la matrice des corrélations, la coordonnée d'une variable V sur un axe P = le coefficient de corrélation de V et P . Une variable n'est bien représentée sur un plan principal (= 2 axes principaux) que si ses coordonnées sont élevées $\Rightarrow$ si elle est proche du cercle des corrélations

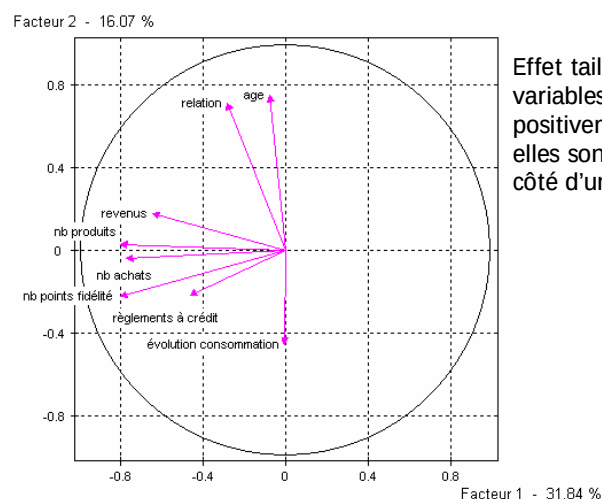
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



111

ACP avec SPAD



Effet taille : si toutes les variables sont corrélées positivement entre elles $\Rightarrow$ elles sont toutes du même côté d'un axe factoriel

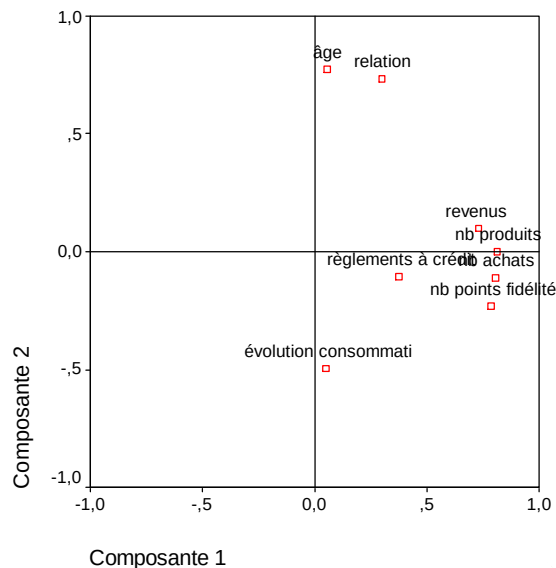
18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



112

ACP avec SPSS



18/04/2005

l.fr

113

Pièges à éviter

- La proximité de 2 variables est mesurée par le cosinus de l'angle qu'elles forment avec l'origine dans $\mathbf{R}^n$
 - dans $\mathbf{R}^2$, cette proximité n'a de sens que si les variables sont bien représentées sur le plan factoriel, i.e. sont proches du cercle des corrélations
- Le 1^{er} plan principal n'est pas toujours le seul intéressant
- Ne pas superposer l'espace des individus et celui des variables
- Eviter qu'un seul individu (ou un petit groupe d'individus) ait une trop forte contribution aux 1^{ers} axes (> 25 %)
 - si c'est le cas : lui affecter un poids = 0 (et un poids = 1 aux autres individus) pour le transformer en point illustratif grâce à l'instruction WEIGHT

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

114

Variables supplémentaires

- Variables ne servant pas à la construction des axes principaux, mais représentées sur les plans principaux
- Variables qualitatives ou quantitatives
- Objectif :
 - variables que l'on veut lier aux variables actives mais pas lier entre elles
 - des variables que l'on veut expliquer par les variables actives
 - variables que l'on veut utiliser pour conforter l'interprétation des axes sans faire appel à des variables ayant servi à les déterminer

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



115

Variantes de l'ACP

- Rappel sur le but de l'ACP standard
 - maximiser la variance restituée sur le 1^{er} facteur
 - **6** maximiser sa valeur propre = somme des carrés des coefficients de corrélation des variables avec ce facteur
- Inconvénient
 - les variables vont toutes + ou - dans le sens du 1^{er} facteur
- Rotation des facteurs :
 - on pivote les facteurs, en respectant ou non l'orthogonalité, en remplaçant le critère ci-dessus par un autre (selon les méthodes)
 - le but est de faciliter l'interprétation
 - la variance expliquée totale ne change pas après rotation (le sous-espace de projection est le même)
 - mais sa répartition change

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



116

Les différentes rotations possibles

- Rotation orthogonale : facteurs non corrélés
 - meilleure interprétation des facteurs
 - ACP varimax
 - ACP quartimax
 - ACP equamax : compromis entre varimax et quartimax
- Rotation oblique : facteurs corrélés
 - les facteurs ne sont plus orthogonaux
 - valeurs propres + fortes $\Rightarrow$ + forte corrélation des facteurs avec les variables
 - mais interprétation plus difficile
 - ACP oblimin
 - ACP promax (+ rapide $\Rightarrow$ utilisé sur de gros volumes)
- Dans SAS/STAT : proc FACTOR
 - + générale, mais + complexe et - rapide que PRINCOMP

ACP varimax

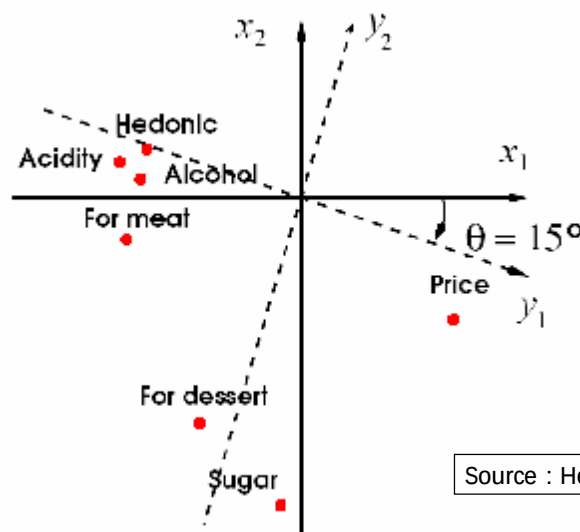
- Pour chaque facteur
 - on calcule ses coefficients de corrélation avec l'ensemble des variables
 - puis on calcule la variance de ces coefficients de corrélation
 - on pivote le facteur de façon à maximiser cette variance
 - différence avec l'ACP standard, où on maximise la somme des carrés des coefficients de corrélation et non leur variance
- Chaque facteur est fortement corrélé à quelques variables et faiblement corrélé aux autres



ACP varimax

- Chaque variable est identifiée à un (ou un petit nombre) facteur
 - les variables ne vont plus seulement dans le sens du 1^{er} axe
 - elles sont bien séparées par axe
 - les axes sont facilement interprétables
- Variante de l'ACP la plus utilisée

Exemple de rotation VARIMAX



Source : Hervé Abdi

ACP quartimax

- Pour chaque variable
 - on calcule ses coefficients de corrélation avec l'ensemble des facteurs
 - puis on calcule la variance de ces coefficients
 - on pivote les facteurs de façon à maximiser cette variance
- Chaque variable est fortement corrélée à quelques facteurs et faiblement corrélée aux autres
- Minimisation du nombre de facteurs nécessaires pour expliquer chaque variable
- Variante utilisée dans la classification VARCLUS de variables (implémentée dans SAS/STAT)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



121

Autre variante : l'ACP sur variables catégorielles

- Transforme de façon optimale les variables qualitatives en variables numériques
 - de façon à optimiser les propriétés de la matrice des corrélations ou des covariances
 - notamment maximiser l'importance des 1^{ères} valeurs propres
- Dans SAS/STAT : proc PRINQUAL
- Par rapport à l'analyse des correspondances (proc CORRESP) :
 - (=) on traite des variables qualitatives
 - ($\neq$) on travaille sur la matrice des corrélations et non sur les effectifs de tableaux de contingence
 - ($\neq$) il s'agit d'une ACP sur des var. qualitatives transformées chacune en 1 var. numérique, et non, comme l'ACM, d'une ACP sur les k indicatrices de chaque variable qualitative

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



122

La préparation des données : Échantillonnage



L'échantillonnage des données 1

- Étape incontournable de plusieurs techniques
 - notamment la prédiction et le classement, dont la plupart des algorithmes mettent en œuvre un *échantillon d'apprentissage* et un *échantillon de test*
 - panel de consommateurs
- Néanmoins il est parfois déconseillé d'effectuer toute une étude sur un échantillon seulement :
 - recherche de typologie de fraudes ou de segments étroits à forte valeur ajoutée
- Dans tous les cas, l'échantillonnage est une opération délicate, qui nécessite une bonne connaissance de la population étudiée



L'échantillonnage des données 2

- Possible à condition :
 - de réussir à constituer un échantillon non biaisé, dont les observations peuvent être extrapolées à l'ensemble de la population
 - d'avoir un nombre minimum d'individus dans l'échantillon
 - de ne pas rechercher de phénomènes trop rares
- Types d'échantillonnage aléatoire :
 - simple
 - systématique
 - stratifié
 - par grappes
 - par étapes

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



125

Exemple d'échantillonnage

- Échantillon de clients, numérotés *aaaffffnn*
 - *aaa* = no agence (de 1 à 999)
 - *ffff* = no de foyer dans l'agence (de 1 à 9999)
 - *nn* = rang du client dans le foyer (1=H, 2=F, autres = enfants)
- Échantillonnage simple : tirage aléatoire du no de client
- Échantillonnage systématique : 1er no de client tiré aléatoirement, puis $no+k$, $no+2k$, etc. (NB : si $k = 100$!)
- Échantillonnage stratifié : répartition des clients en tranches d'âge, puis no de client tiré au sort
- Échantillonnage par grappes : tirage aléatoire de l'initiale du nom de famille, puis recensement (NB : si l'initiale = « D » ou « L » !)

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



126

Taille d'échantillon (pour un taux)

Quand un événement se produit dans une population avec une probabilité p (exemple : 80 % des clients sont satisfaits $\Rightarrow p = 0,8$), cette probabilité peut être estimée à partir d'un échantillon de taille n de cette population. Cette probabilité p est estimée par la fréquence $f = k/n$ de survenance de l'événement dans l'échantillon. Comme la variable k suit une loi binomiale $B(n,p)$ de moyenne $\mu = n.p$ et de variance $\sigma^2 = n.p.(1-p)$, la fréquence f suit une loi binomiale de moyenne $= p$ et de variance $= p.(1-p)/n$. On sait que lorsque n est grand, la loi binomiale tend vers une loi normale de paramètres (μ, σ) . Sachant que 95 % des valeurs d'une loi normale (μ, σ) se trouvent dans l'intervalle $[\mu - 1,96\sigma, \mu + 1,96\sigma]$, la fréquence f a une probabilité de 95 % de se trouver dans l'intervalle de confiance :

$$[p - 1,96 \sqrt{\frac{p(1-p)}{n}}, p + 1,96 \sqrt{\frac{p(1-p)}{n}}]$$

Donc l'intervalle

$$[f - 1,96 \sqrt{\frac{f(1-f)}{n}}, f + 1,96 \sqrt{\frac{f(1-f)}{n}}]$$

a une probabilité proche de 95 % de contenir la vraie valeur de p .

On dit que l'intervalle ci-dessus est l'intervalle de confiance au seuil de risque de 5 % (le plus fréquemment utilisé). Au seuil de risque de 1 %, il faudrait remplacer la constante 1,96 ci-dessus par 2,5758.

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

127

Taille d'échantillon (pour une moyenne)

Si l'on veut estimer la moyenne m d'une variable X dans la population entière, à partir des valeurs mesurées $x_1, \dots, x_n$ dans un échantillon de n individus, voici comment il faut procéder.

On calcule la moyenne $m = \frac{1}{n} \sum_i x_i$, puis l'écart-type d'échantillon σ selon la formule :

$$\sigma = \sqrt{\frac{\sum_i (x_i - m)^2}{n-1}}$$

Ensuite, si $n \leq 30$, on va lire un paramètre t_α dans la table de la distribution de Student à $n-1$ degrés de liberté, en se fixant un seuil de risque α (généralement $\alpha = 0,05$, c'est-à-dire 5 %). Si le test est bilatéral, il y a 2 zones de rejet, chacune avec une probabilité $\alpha/2$, soit α au total. Dans ce cas, on remplace t_α par $t_{\alpha/2}$, et on regarde donc généralement $t_{0,025}$.

Si $n > 30$, c'est même plus simple, la loi de Student est approchée par la loi normale centrée réduite, et on va chercher t_α dans cette table ; en particulier, $t_{0,025} = 1,96$.

Enfin, on peut conclure qu'au seuil de risque α , la moyenne m est dans l'intervalle de confiance :

$$[m - t_{\alpha/2} \frac{\sigma}{\sqrt{n}}, m + t_{\alpha/2} \frac{\sigma}{\sqrt{n}}].$$

18/04/2005

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

128

Retour au data mining

- Les considérations précédentes peuvent être utilisées pour interpréter une classification.
- Pour chaque segment :
 - et chaque variable continue : on compare sa moyenne dans le segment à sa moyenne générale
 - et chaque variable catégorielle : on compare la proportion de chaque modalité dans le segment à sa proportion dans la population entière.
- On peut ainsi caractériser chaque segment par les variables qui le singularisent le + de la population entière.

