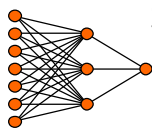


Stéphane Tufféry

Statisticien - Data Miner - Formateur



DATA MINING - SCORING



STATISTIQUE APPLIQUÉE

APPLICATION AU CRM



04/07/2004

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



1

Plan du cours

- Qu'est-ce que le data mining ?
- A 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 text mining
- Le web mining

04/07/2004

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



2

La préparation des données

04/07/2004

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



3

Les différents formats de données

- Données *continues* (ou *d'échelle*)
 - dont les valeurs forment un sous-ensemble infini de \mathbf{R} (exemple : salaire)
- Données *discrètes*
 - dont les valeurs forment un sous-ensemble fini ou infini de \mathbf{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...

04/07/2004

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



4

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)

04/07/2004

© 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*

04/07/2004

© 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

04/07/2004

© 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

04/07/2004

© 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.

04/07/2004

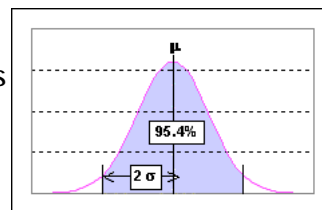
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



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)



04/07/2004

© 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

04/07/2004

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



11

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

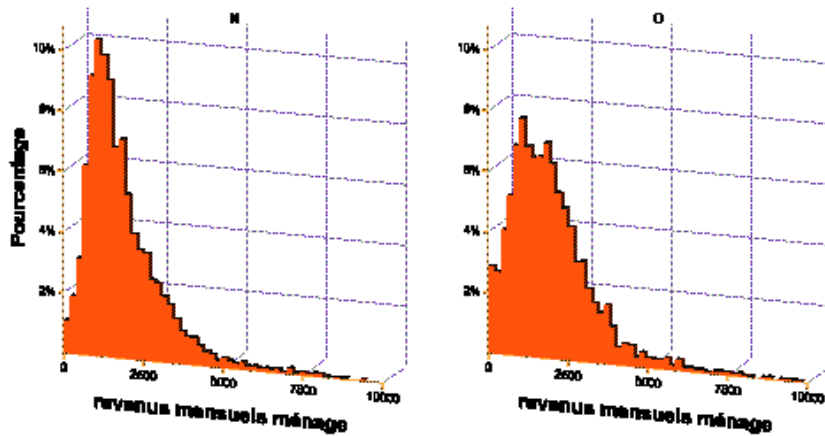
04/07/2004

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



12

Analyse exploratoire des données



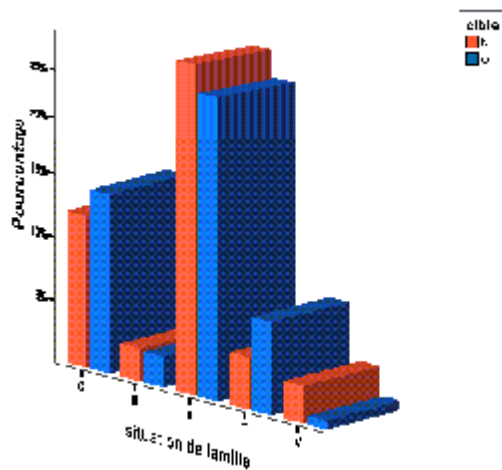
04/07/2004

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



13

Analyse exploratoire des données



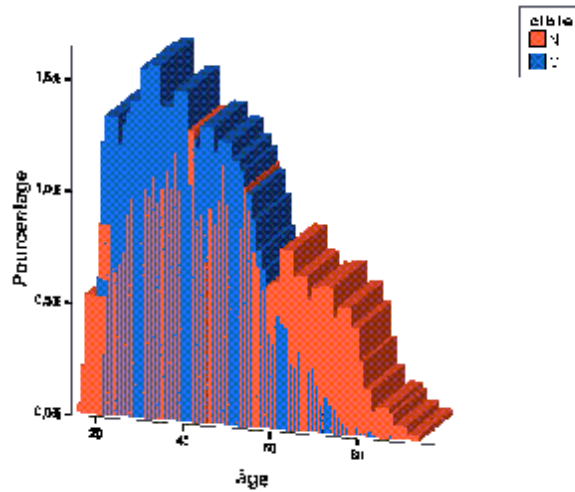
04/07/2004

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



14

Analyse exploratoire des données

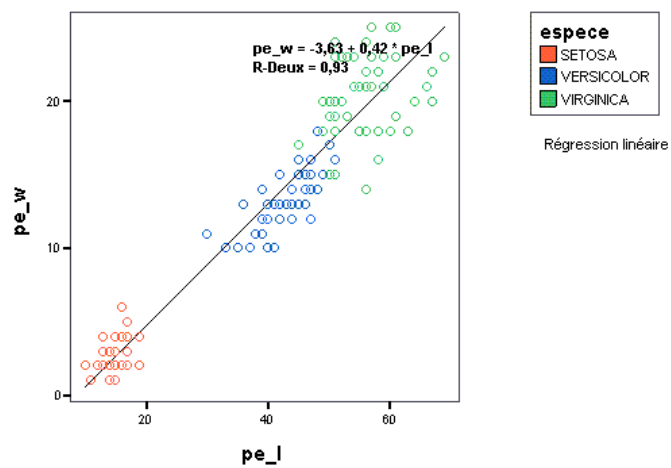


04/07/2004

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

15

Analyse exploratoire des données



04/07/2004

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

16

Caractéristiques de dispersion

- Étendue
- Écart interquartile $q3 - q1$
- Variance
 - 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

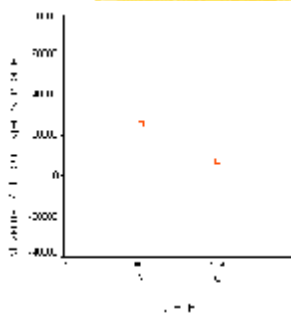
	Levene Statistic	df1	df2	Sig.
1996 Sales	6.708	2	387	.001

04/07/2004

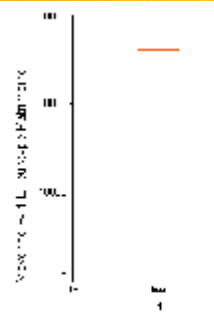
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

17

Homogénéité des variances



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



KO é

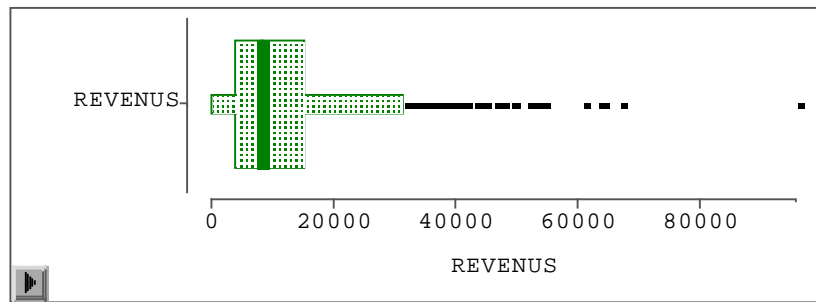
OK ->

04/07/2004

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

18

Boîte « à moustaches » (boxplot)



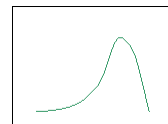
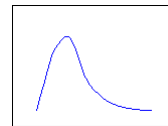
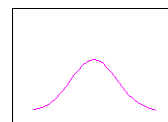
04/07/2004

© 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



04/07/2004

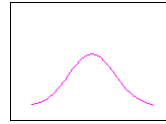
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

20

Caractéristiques de forme 2/2

- Coefficient d'aplatissement (« kurtosis »)

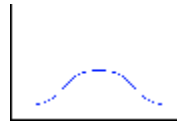
- = 3 si aplatie comme Gauss



- > 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

04/07/2004

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

21

Variables continues

Variable	Min	Max	Mean	Std Dev.	Missin	Skewness	Kurtosis
AGE	18	75	42.611	13.403	0%	0.4148	-0.659
NTEPLIGU	-3488	235255	8555.5	15338	0%	5.1065	50.707
PATRINDI	0	522307	17512	38404	0%	5.6887	48.771
NBPROD	1	20	3.369	1.164	0%	0.6739	-0.051
GOLDEMOY	-6499	90975	1723.1	3204.3	0%	12.859	311.25
NBJDEB	0	30	5.55	6.5228	0%	1.2235	0.6538
NBJDEP	0	27	1.8285	3.534	0%	3.0798	10.323
NTCRDCOM	511.54	22243	2649.9	1821.1	0%	3.5701	20.524
DEPENSES	3.3108	23314	2519.8	1384.5	0%	4.9079	44.173
NTPRELEV	0	3232.1	279.65	263.72	0%	2.3272	12.054
NTFINANC	0	1157.3	40.854	110.26	0%	4.7056	28.048
VAR_REVN	0.4513	303.15	49.57	40.767	0%	2.3623	7.2095
VAR_DEPS	3.7768	289.38	50.719	38.101	0%	2.9595	5.2955
REVENUS	297.68	15532	2142.5	1299	0%	2.6054	12.667
RATIOPSR	0	411.5	7.2979	18.973	0%	5.2763	152.37
RELATION	1	33	14.215	7.1463	0%	0.0227	-0.838
NTCREVO	0	20337	359.3	1117.1	0%	7.0285	81.517
NTDECAUT	0	12189	694.85	699.97	0%	5.7757	62.374
NBOPECRE	1	41	6.8465	3.5901	0%	1.3304	5.1824
NBOPECEB	2	111	33.599	15.87	0%	0.8725	0.9791
NTMTCRD	630.51	68176	3545.6	3279.2	0%	5.5453	89.24
AGECPTE	4	134	56.571	34.892	0%	0.4043	-0.944

04/07/2004

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

22

Après suppression des extrêmes

Name	Min	Max	Mean	Std Dev.	Missing %	Skewness	Kurt
AGE	19	74	42.556	13.161	0%	0.3864	***
NTEPLIGU	-1000	60000	7329.8	11201	0%	2.5778	***
PATRINDI	0	150000	15457	27086	0%	2.9761	***
MBPROD	1	20	9.4525	4.2905	0%	0.5348	***
SOLDENDY	-1000	10000	1676.5	2056.7	0%	2.0899	***
MBJDEB	0	30	5.5965	6.5537	0%	1.2478	***
MBJDEP	0	24	1.8345	3.4863	0%	2.7908	***
NTCRDCOM	611.6	10000	2584.8	1684.9	0%	2.1741	***
DEFENSES	191.07	10000	2532.2	1652.9	0%	2.1424	***
NTPRELEV	0	3232.1	280.53	277.36	0%	2.7267	***
NTFINANC	0	700	41.343	97.095	0%	3.3358	***
VAR_REVN	0.5726	330.84	49.713	42.807	0%	2.682	***
VAR_DEPS	6.9548	295.76	59.856	38.394	0%	2.0821	***
REVENUS	437.14	22232	2213.8	1503.1	0%	4.1567	***
RATIOPSR	0	60	6.1717	10.345	0%	2.336	***
RELATION	1	33	14.291	7.0285	0%	0.0175	***
NTCREVO	0	20337	357.03	1143.8	0%	7.4958	***
NTDECAUT	0	10674	722.92	706.29	0%	4.4652	***
MBOPECRE	1	28	6.8195	3.4168	0%	0.3397	***
MBOPEDEB	1	30	34.282	16.047	0%	0.7363	***
NTWTCRD	701.8	79269	3661.2	3820.3	0%	8.3589	***
AGECPTE	4	134	56.597	33.955	0%	0.3922	***

04/07/2004

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

23

Filtrage des extrêmes

Eliminate rare values: Keep Missing?

For class variables with different values
for value that occur time(s).

Eliminate extreme values in interval var

Median Abs Dev (MAD)
 Model Center
 Std deviations from mean % top/bottom percentiles
 Extreme Percentiles

Use sample
 Use entire data

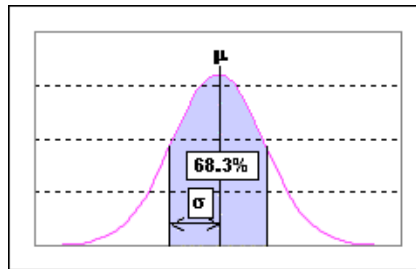
04/07/2004

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

24

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

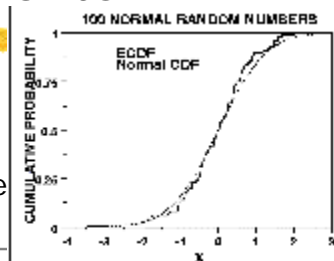
04/07/2004

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

25

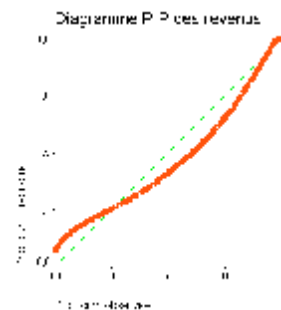
Tests de normalité

- 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 gaussienne



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

- Test de Lilliefors
 - perfectionne Kolmogorov-Smirnov
- Test de Shapiro-Wilk (petits effectifs)
 - mesure l'alignement sur la droite



04/07/2004

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

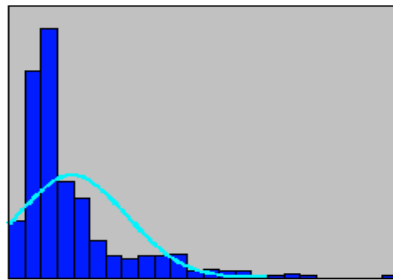
26

Normalisation

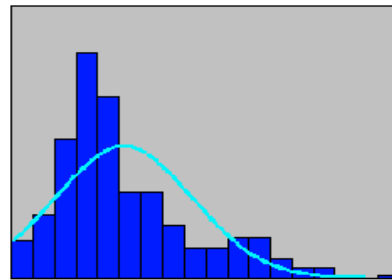
Frequencies Statistics

	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

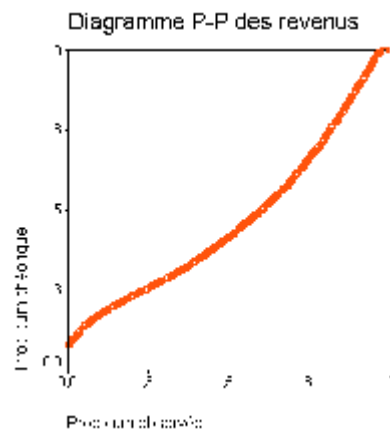
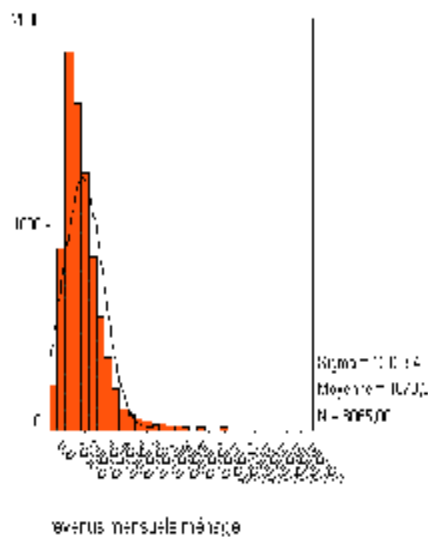


04/07/2004

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

27

Normalisation : exemple des revenus

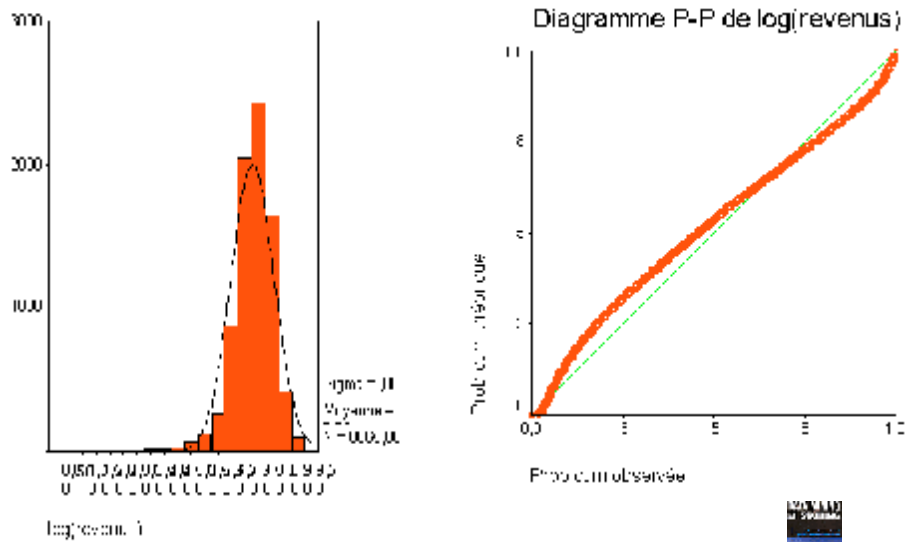


04/07/2004

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

28

Normalisation par le logarithme

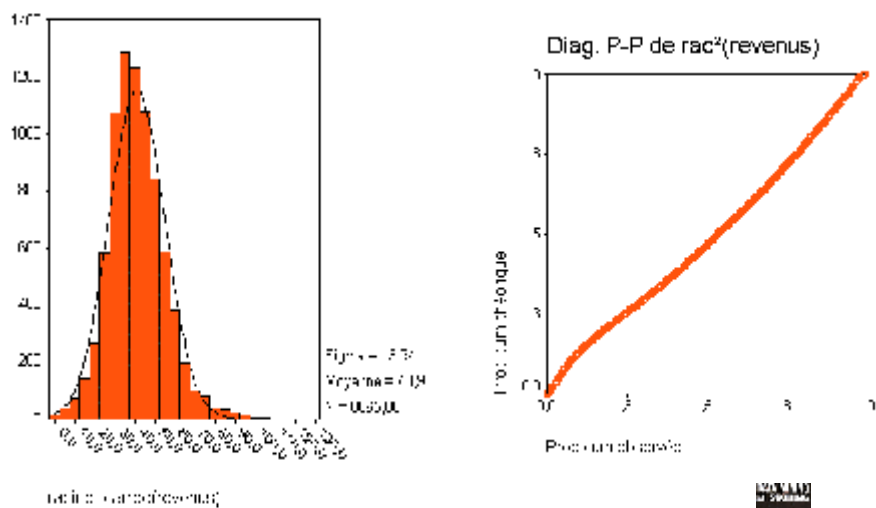


04/07/2004

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

29

Normalisation par la racine carrée

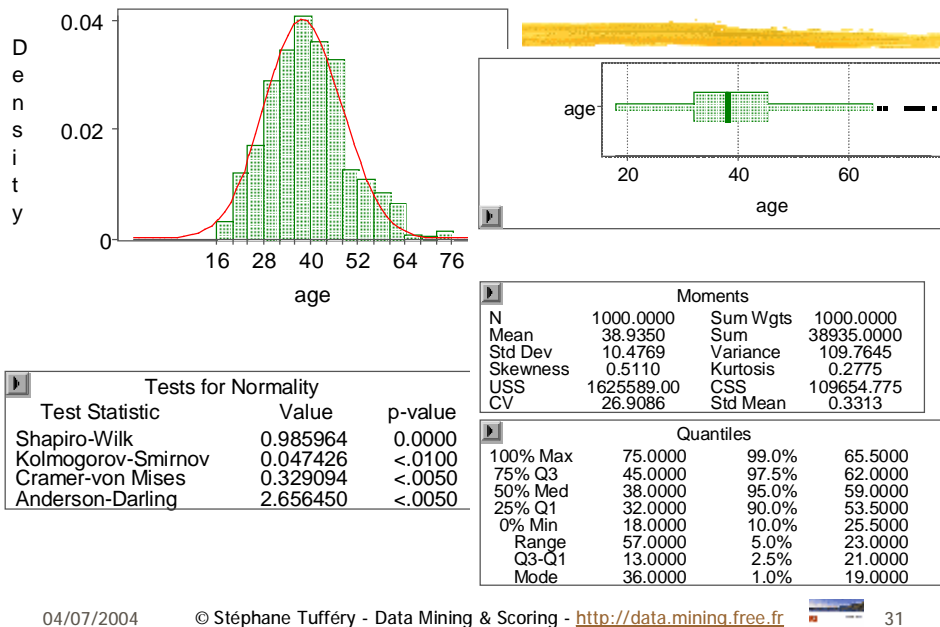


04/07/2004

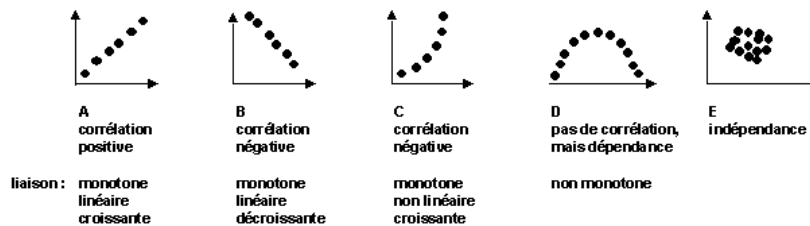
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

30

Exploration sous INSIGHT



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

Coefficients de Pearson et Spearman

- Rho de Spearman plus général car calculé sur les rangs des valeurs et non les valeurs elles-mêmes
- 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



corrélation négative
apparemment positive

04/07/2004

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

33

Multicolinéarité

- 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)

	Valeur propre	Indice de conditionnement	Proportions de la variance						
			(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	7,088	.97	.00	.00	.00	.76	.00	.24

04/07/2004

34

Généralisation à plus de 2 variables

- Analyse canonique de la corrélation linéaire
 - non plus entre 2 variables continues ou binaires
 - mais entre $n (\geq 2)$ ensembles $\{U_i\}, \{V_i\}...$ de variables continues ou binaires
 - on cherche les combinaisons linéaires qui maximisent la corrélation entre $\sum_i \lambda_i U_i, \sum_i \mu_i V_i...$
 - proc CANCELL de SAS – proc OVERALS de SPSS
- Analyse canonique de la corrélation non linéaire
 - entre $n (\geq 2)$ ensembles de variables quelconques
 - permet la détection d'effets non linéaires
 - proc CANCELL de SAS – proc OVERALS de SPSS

04/07/2004

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



35

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 – hétéroscedasticité	test T de Welch	Welch - ANOVA
non normalité – hétéroscedasticité (**)	Wilcoxon – Mann – Whitney	Kruskal – Wallis
non normalité – hétéroscedasticité (**)	test de la médiane	test de la médiane
non normalité – hétéroscedasticité (**)		test de Jonckheere-Terpstra (échantillons ordonnés)

moins puissant

(*) Ces test supportent mieux le défaut de normalité que l'hétéroscedasticité.
 (**) 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 => on détecte à tort des différences significatives (au seuil de 95 % : dans 27 % des cas pour 4 échantillons égales)

04/07/2004

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



36

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 => MANOVA
 - si m variables nominales indépendantes => 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

04/07/2004

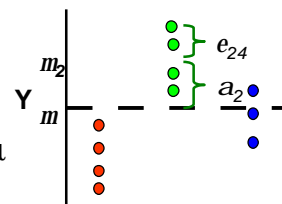
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



37

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 - μ
- 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
 - 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)



04/07/2004

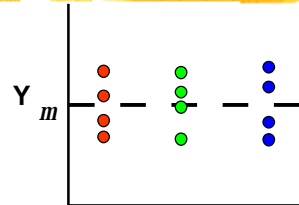
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



38

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)

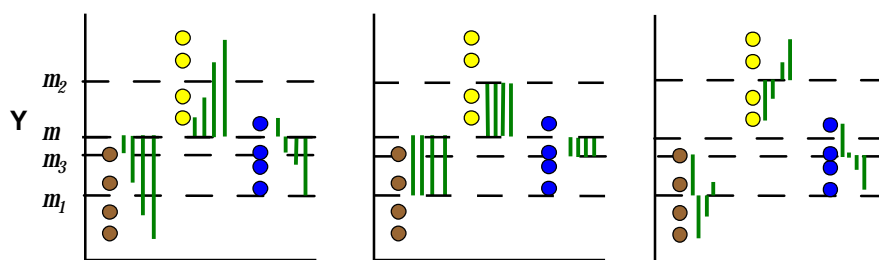
04/07/2004

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



39

Répartition de la somme des carrés



SC Totale

SC Modèle
(interclasse)

SC Erreur
(intraclasse)

- Groupe 1
- Groupe 2
- Groupe 3

04/07/2004

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



40

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

$\eta^2 = \rho^2 + \text{non-linéarité}$ (ρ = coefficient de Pearson)

04/07/2004

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

41

Principe du test ANOVA

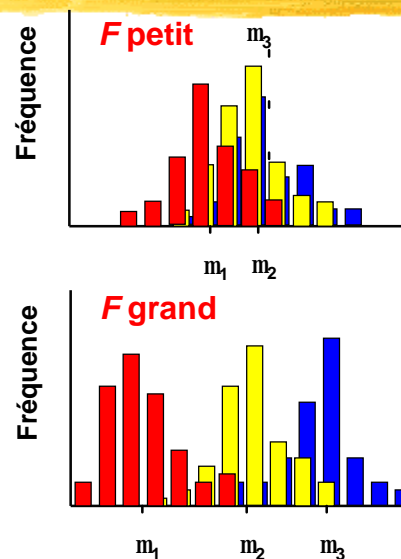
- On appelle « analyse de la variance » ce qui est en fait un test d'égalité de la moyenne, en raison de la façon de réaliser ce test, qui consiste à décomposer la variance de la variable continue Y en 2 parties :
 - ce qui peut être attribué aux différences entre groupes (variance inter-classe)
 - ce 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

04/07/2004

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

42

Illustration du test ANOVA



04/07/2004

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



43

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 est :

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

- 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 $U = 16 + 10 - 13 = 13$

04/07/2004

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



44

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 libertés
 - sinon, regarder valeurs critiques dans une table
- Presque aussi puissant qu'ANOVA sur son propre terrain

04/07/2004

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

45

Liaison entre 2 variables catégorielles : Test du Chi-2

Test utilisé pour s'assurer que la distribution d'une variable suit une loi probabiliste donnée, en comparant la distribution observée (d'effectifs $\{O_i\}$) et la distribution théorique (d'effectifs $\{T_i\}$).

Si l'hypothèse $(H_0) = \{O_i = T_i, \text{ pour tout } i = 1, \dots, n\}$ est vraie, alors la variable

$$\chi^2 = \sum (O_i - T_i)^2 / T_i$$

doit suivre une loi probabiliste appelée loi du χ^2 .

La table de la loi du χ^2 (à $n - 1$ degrés de liberté) fournit la valeur $\chi^2_{a,cal}$ telle que χ^2 ait une probabilité $< \alpha \%$ de dépasser $\chi^2_{a,cal}$. On prend généralement $\alpha = 5 \%$. Si la valeur mesurée de χ^2 est pourtant $> \chi^2_{a,cal}$, cette probabilité $< \alpha \%$ est considérée comme trop faible pour que l'hypothèse (H_0) puisse être considérée comme valide : elle est donc rejetée.

Le test du χ^2 n'est bien sûr utilisable que si $T_i > 0$ pour tout i

Il n'est fiable que si $T_i \geq 5$ pour tout i .

04/07/2004

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

46

Liaison entre variables catégorielles : Chi-2

Le test du χ^2 est souvent utilisé pour tester l'indépendance de deux variables X et Y.

Si X et Y sont indépendantes, alors, pour tous i et j :

$$\begin{aligned} & \text{nombre d'individus tels que } \{X=i \text{ et } Y=j\} \\ & = \\ & \text{nb d'individus tels que } \{X=i\} \times \text{nb d'individus tels que } \{Y=j\} \times 1/N \end{aligned}$$

où N est le nombre total d'individus.

En notant O_{ij} le terme à gauche de l'égalité, et T_{ij} le terme de droite, le test d'indépendance de X et Y est le test du χ^2 appliqué à la variable

$$\chi^2 = \sum (O_{ij} - T_{ij})^2 / T_{ij}$$

04/07/2004

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



47

Chi-2 : attention aux effectifs

	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)

04/07/2004

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



48

Chi-2 : attention aux effectifs

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.



04/07/2004

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

49

V de Cramer

- V de Cramer = racine carrée (χ^2 / χ^2_{\max})
 - mesure directement l'intensité de la liaison de 2 variables catégorielles, sans avoir recours à une table du chi-2
 - en intégrant le nombre de degrés de liberté, par l'intermédiaire de χ^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



04/07/2004

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

50

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 chi-2 permet de détecter une dépendance entre les deux variables
- La contribution au chi-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

04/07/2004

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



51

Tableau de contingence : attention aux pièges

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%
TOTAL	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%
TOTAL	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%
TOTAL	2 179	109	2 288	4,76%

04/07/2004

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



52

Analyse des correspondances

- L'analyse factorielle des correspondances 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
 - ou le tableau disjonctif

04/07/2004

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



53

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 $X_j(i^e \text{ individu}) = x_{jk}$, et 0 sinon

04/07/2004

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



54

Analyse des correspondances

- L'ACM s'appuie 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
- L'ACM permet 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

04/07/2004

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



55

Interprétation des sorties

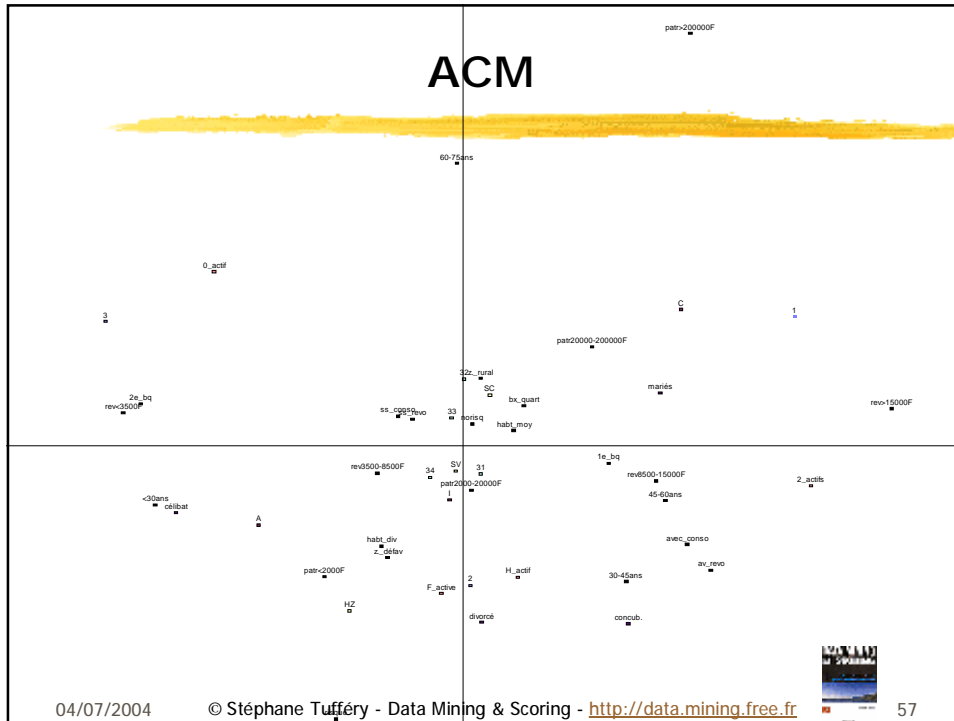
- Coordonnée d'une modalité sur un axe
 - on s'intéresse parfois aux modalités dont la coordonnée > racine carrée (valeur propre de l'axe)
- Contribution d'une modalité à un axe
 - % de l'inertie de la modalité dans l'inertie de l'axe
- Cosinus² : qualité de la représentation d'une modalité sur un axe
 - % pris par l'axe dans la dispersion de la modalité
 - plus \cos^2 est petit, meilleure est la représentation

04/07/2004

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



56



Analyse en composantes principales

- A 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.

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
- L'ACP sur les indicatrices de variables nominales conduit aux mêmes résultats que l'ACM sur ces var. nominales

04/07/2004

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



59

Métrique dans l'espace des individus

- Métrique euclidienne
 - $d(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2$
- Métrique « 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 métrique, 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

04/07/2004

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



60

Composantes principales

- 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 »

$$\text{cov}(XY) = \sigma_X \sigma_Y r_{XY}$$

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

04/07/2004

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

61

Composantes principales

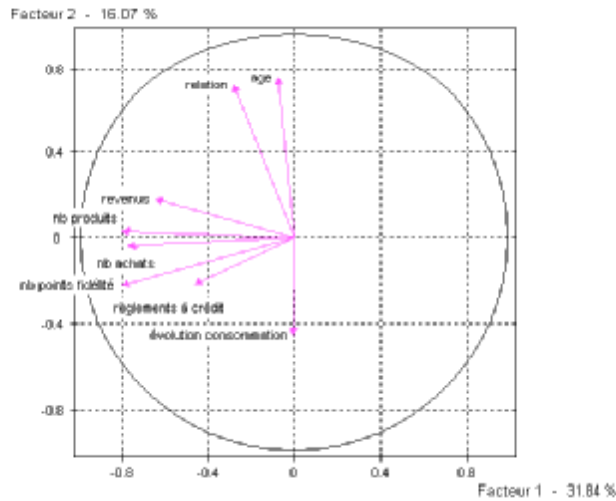
- Si on a des données hétérogènes, avec des ordres de grandeur différents, diagonaliser la matrice des corrélations
- Sinon, diagonaliser la matrice des covariances
- Les composantes principales sont classées par variance décroissante
- Variance d'une composante principale = valeur propre correspondante 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

04/07/2004

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

62

ACP



04/07/2004

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



63

Pièges à éviter

- Ne pas superposer l'espace des individus et celui des variables
- La proximité de 2 variables n'a de sens que si elles sont proches du cercle des corrélations
- Le 1er plan principal n'est pas le seul intéressant
- Eviter qu'un individu (ou un petit groupe d'individus) ait une trop forte contribution aux 1ères CP
- *Critère de Kaiser* : sur des données centrées réduites, conserver les CP correspondant aux valeurs propres > 1
- Calculer les valeurs cumulées successives $\lambda_1/\sum_i \lambda_i$, $(\lambda_1 + \lambda_2)/\sum_i \lambda_i \dots$ pour voir quelle proportion de la somme des variances $\sum_i \lambda_i$ est restituée par les p premières composantes
- Effet taille : toutes les var. sont corrélées positivement entre elles \Rightarrow toutes du même côté d'un axe factoriel

04/07/2004

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



64

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

04/07/2004

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



65

Variantes de l'ACP

- Rappel sur le but de l'ACP standard
 - maximiser la variance restituée sur le 1^{er} facteur
 - $\hat{\sigma}$ 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

04/07/2004

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



66

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 = + forte corrélation des facteurs avec les variables
 - mais interprétation plus difficile
 - ACP oblimin
 - ACP promax (+ rapide => utilisé sur de gros volumes)

04/07/2004

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



67

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

04/07/2004

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



68

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 facteur
 - les facteurs sont facilement interprétables
- Variante de l'ACP la plus utilisée

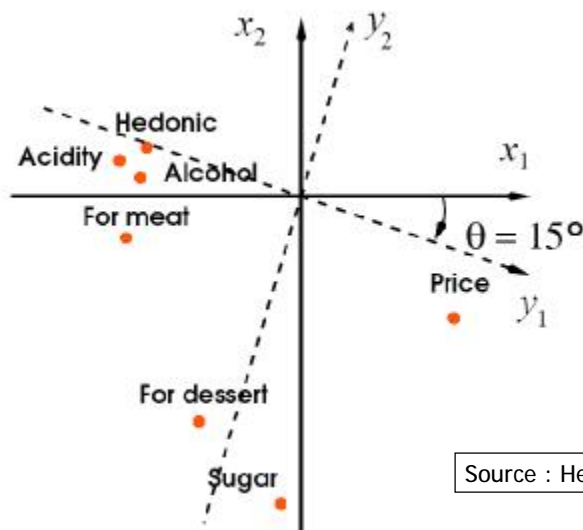
04/07/2004

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



69

Exemple de rotation VARIMAX



Source : Hervé Abdi

04/07/2004

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



70

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

04/07/2004

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



71

Le travail des données : Échantillonnage

04/07/2004

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



72

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

04/07/2004

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



73

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

04/07/2004

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



74

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 » !)



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 :

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

Donc l'intervalle

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

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.



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_a 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_a par $t_{a/2}$, 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_a 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 :

$$\left[m - t_a \frac{S}{\sqrt{n}}, m + t_a \frac{S}{\sqrt{n}} \right].$$



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.

