

Analyse des données appliquée aux techniques d'enquête par sondage.



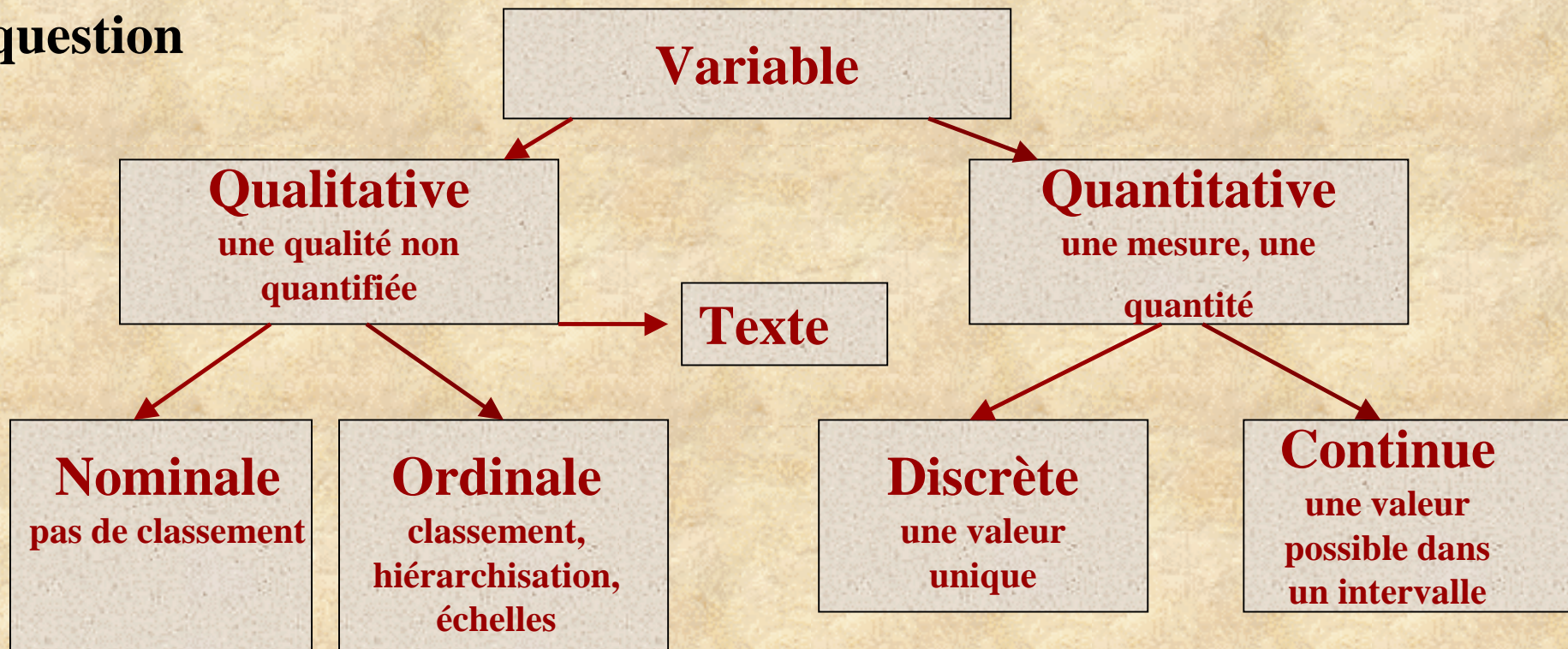
III. L'analyse des données

- 3.1 La notion de variable
- 3.2 Quels traitements développer ?
- 3.3 L'analyse des données de l'enquête



3.1 La notion de variables :

- Différencier question (*libellé*), variable (*titre et mode d'expression de la question*) et modalités (*réponses possibles*)
- Les différents types de variables cad de modes d'expression de la question



3.1 La notion de variables : *nature de l'information*

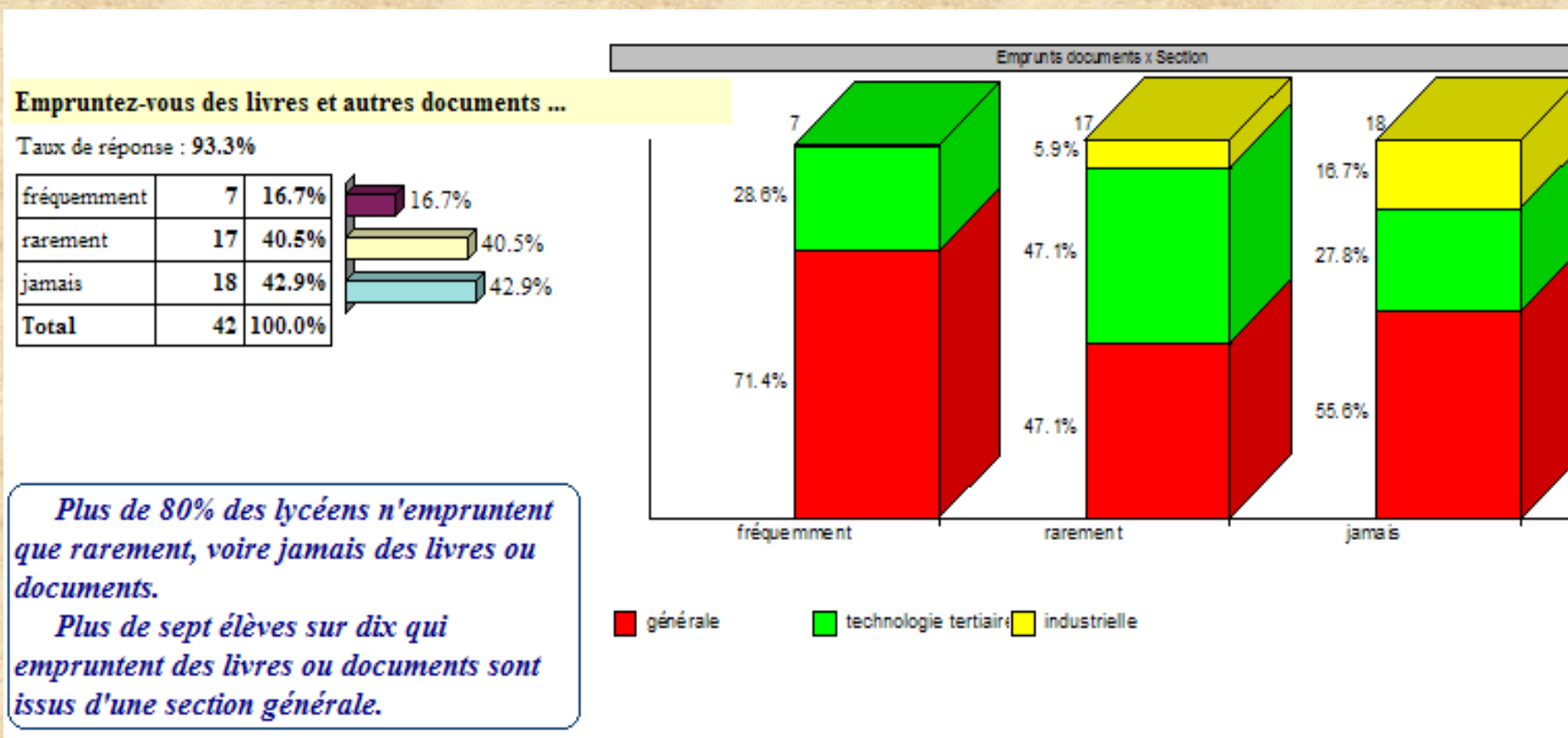
- Variables **quantitatives** ou numériques : (précision d'une **grandeur**) : âge, niveau dépense,...
- Variables **échelles** : organisation d'un ordre, **d'une graduation**
 - Satisfaction :
• Pas du satisfait, Peu satisfait, Assez satisfait, Très satisfait
- Variables **nominales** : définition d'un **état**
 - Genre :
Homme, Femme
- Variables **texte** : commentaires libres



3.2 Quels traitements développer ?

Les objectifs de l'analyse de données

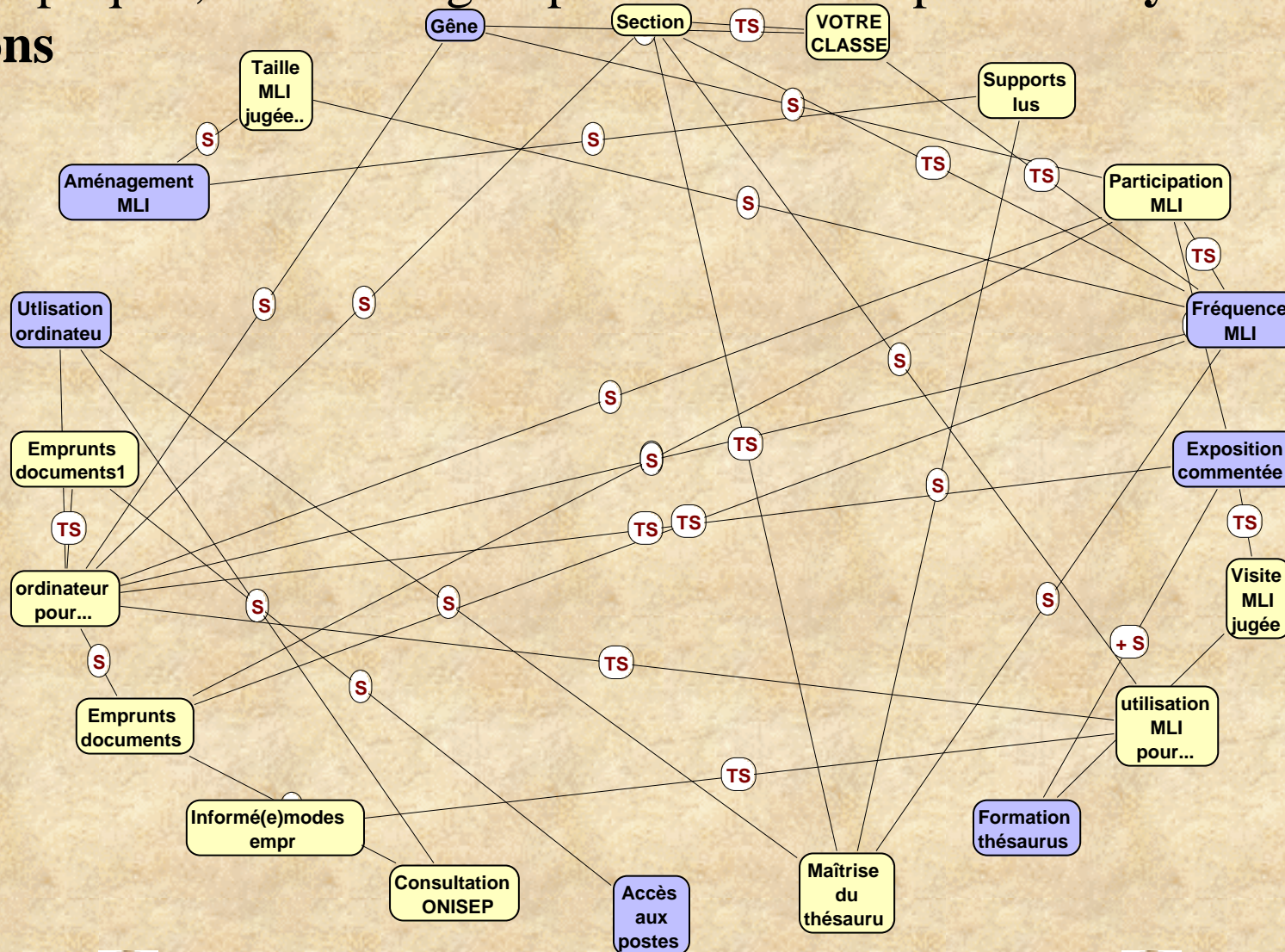
- Pour décrire, agréger, synthétiser : **Tableaux de Bord**



3.2 Quels traitements développer ?

Les objectifs de l'analyse de données

- Pour expliquer, cibler des groupes de variables proches: **Systemes de Relations**



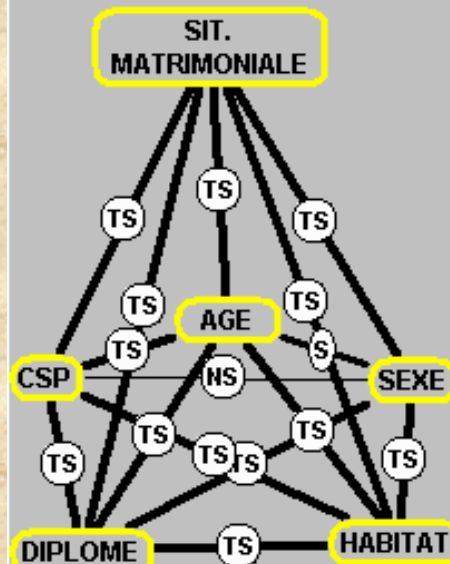
- Pour expliquer, cibler des groupes de variables proches: **Systemes de Relations-** *autre exemple (tiré de Sphinx développement)*

Le rêve des français. Echantillon total : 993 observations

2

Tableaux croisés : analyse bi variée

Graphes des relations entre les variables socio-démographique



Chi2, Anova : TS <= 1% < S <= 5% < PS <= 15%
 Corr. : TS >= 0.8 > S >= 0.6 > PS >= 0.4 > NS

*Il faut examiner 15 tableaux pour savoir comment les caractéristiques de cette population se combinent entre elles...
 L'analyse multivariée qui suit permet une représentation beaucoup plus synthétique.*

CSP x Diplôme

	Aucun	CEP BEPC	CAP BEP	Bac	Bac+2 BTS	Bac+4 DESS	Refus	Autres	Total
Agriculteurs	8.5%	48.9%	14.9%	17.0%	8.5%	0.0%	0.0%	2.1%	100.0%
Commerçant, artisan	14.9%	27.7%	36.2%	14.9%	4.3%	2.1%	0.0%	0.0%	100.0%
Cadre.Prof.Intell. Sup.	3.9%	8.6%	10.5%	13.2%	20.4%	41.4%	0.0%	2.0%	100.0%
Prof.Intermédiaires	1.3%	9.0%	12.8%	19.2%	28.2%	29.5%	0.0%	0.0%	100.0%
Employés	10.8%	19.5%	35.7%	17.0%	11.6%	5.1%	0.0%	0.4%	100.0%
Ouvriers	15.9%	35.6%	37.9%	8.3%	2.3%	0.0%	0.0%	0.0%	100.0%
Retraités	12.5%	42.0%	19.9%	8.0%	6.8%	8.5%	0.6%	1.7%	100.0%
Inactifs, Autre	11.9%	16.7%	7.1%	22.6%	19.0%	22.6%	0.0%	0.0%	100.0%
Total	10.2%	24.7%	24.2%	14.2%	12.3%	13.6%	0.1%	0.8%	100.0%

p = <0.1% ; chi2 = 372.04 ; ddl = 49 (TS)

Age x Sexe

	AGE
Homme	39.92
Femme	42.38
Total	41.28

p = 1.9% ; F = 5.41 (S)

Habitat x Age

	AGE
- de 2 000	43.37
2 000 - 20 000	40.32
20 000 - 100 000	41.64
+ de 100 000	37.59
Région parisienne	41.51
Total	41.28

p = 0.5% ; F = 3.73 (TS)

Age x Sexe

	AGE
Aucun	45.87
CEP BEPC	50.46
CAP BEP	37.75
Bac	34.50
Bac+2 BTS	34.39
Bac+4 DESS	39.53
Refus	57.00
Autres	59.88
Total	41.28

p = <0.1% ; F = 25.20 (TS)

Age x CSP

	AGE
Agriculteurs	43.32
Commerçant, artisan	39.77
Cadre.Prof.Intell. Sup.	39.20
Prof.Intermédiaires	35.49
Employés	35.13
Ouvriers	36.46
Retraités	63.05
Inactifs, Autre	32.37
Total	41.28

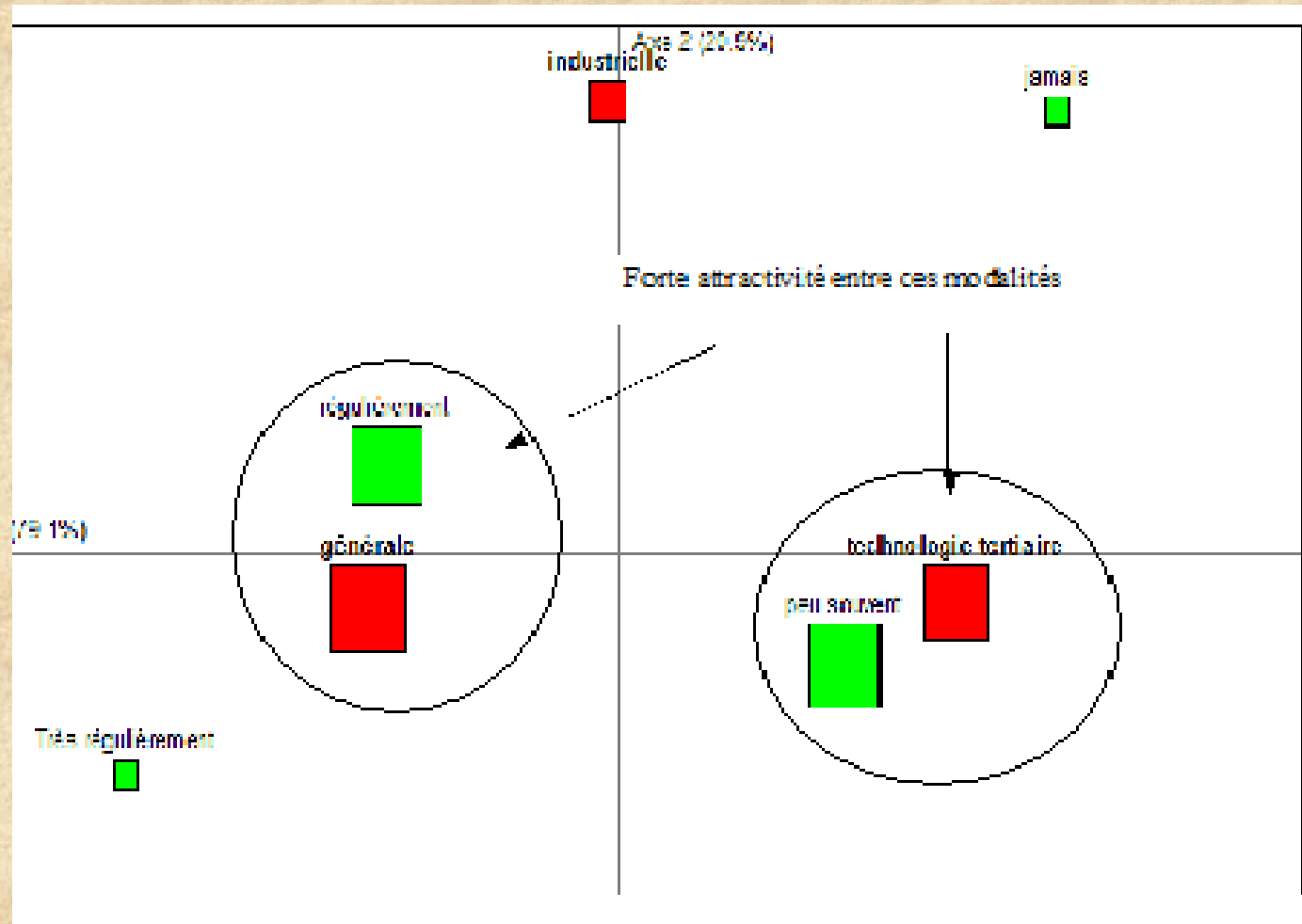
p = <0.1% ; F = 91.08 (TS)



3.2 Quels traitements développer ?

Les objectifs de l'analyse de données

- Pour regrouper, organiser, segmenter : **Typologies et arbres de décision**



3.2 Quels traitements développer ?

Les différents niveaux de l'analyse de données

- Analyse univariée ou « analyse à plat » → *Tableaux à plat*

Pour décrire les résultats d'une variable à la fois

- Analyse bivariée ou « analyse croisée » → *Tableaux croisés ou tableaux de contingence*

Pour mettre en relation deux variables afin d'expliquer, de préciser une analyse

- Analyse multivariée des données» → *Cartes factorielles*

Pour analyser simultanément plus de deux variables pour dresser des typologies, synthétiser



3.3 Les différents niveaux de l'analyse des données :

3.31 *l'analyse univariée*

- Décrire les caractéristiques d'une seule variable à la fois
il y a 25% de lycéens
L'âge moyen des élèves est de 17,9 ans
- Variable nominale ou échelle :
calcul des effectifs, pourcentages et intervalle de confiance.
 - Variable numérique ou échelle :
calcul des moyennes écart-type, mise en classes



3.3 Les différents niveaux de l'analyse des données

l'analyse à plat des variables nominales

Question à réponse unique

Vous êtes dans une section...

Taux de réponse : 100.0%

générale	23	51.1%	51.1%
technologie tertiaire	17	37.8%	37.8%
industrielle	5	11.1%	11.1%
Total	45	100.0%	

La somme des pourcentages est égale à 100

Question à réponses multiples non ordonnées ou ordonnées

Pour quelle(s) matière(s) vous rendez-vous à la MLI ?

Taux de réponse : 82.2%

Somme des pourcentages différente de 100 du fait des réponses multiples et des suppressions.

français	14	31.1%	31.1%
histoire-geo	6	13.3%	13.3%
ECJS	4	8.9%	8.9%
SES	2	4.4%	4.4%
SVT	0	0.0%	0.0%
TPE	2	4.4%	4.4%
PPCP	1	2.2%	2.2%
mathématiques	1	2.2%	2.2%
physique-chimie	0	0.0%	0.0%
langues	1	2.2%	2.2%
exposé	0	0.0%	0.0%
toutes les matieres	4	8.9%	8.9%
faire devoirs	1	2.2%	2.2%
heures de soutien	1	2.2%	2.2%
Total	45		

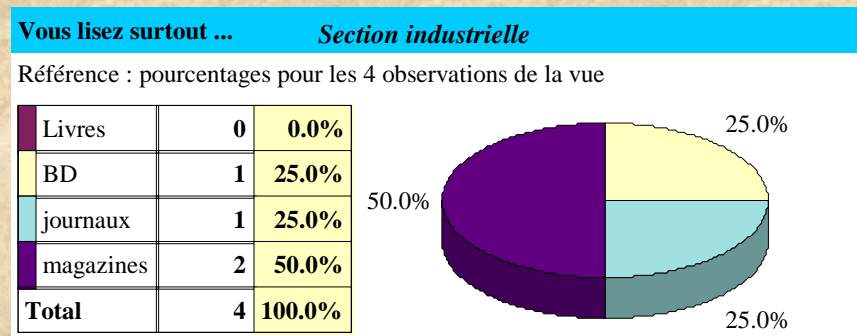
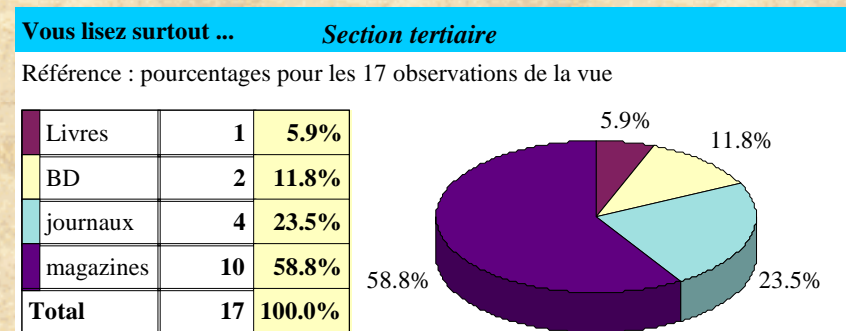
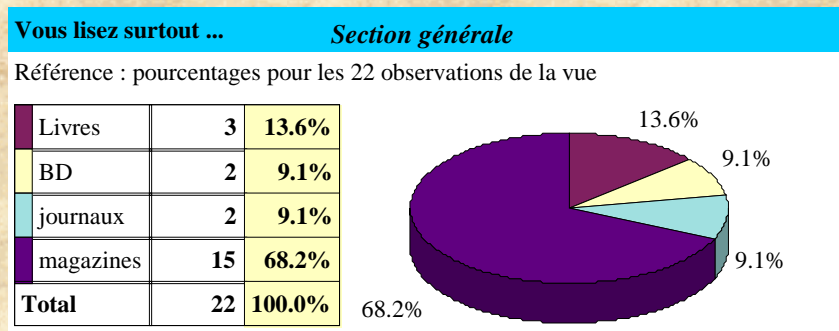
Pourcentages calculés par rapport au nombre d'observations : la somme est supérieure à 100



3.3 Les différents niveaux de l'analyse des données

l'analyse à plat stratifiée

Pour présenter les résultats d'une variable en stratifiant l'échantillon (*ici les habitudes de lecture des lycéens selon leur section d'appartenance*)



3.3 Les différents niveaux de l'analyse des données

l'analyse à plat des variables nominales

Se rend surtout la A quel moment de la journée vous y rendez-vous le plus souvent ?

Se rend surtout la	Nb. cit.	Fréq.
▷ matin	2	4.2%
▷ entre 12h-14h	18	37.5%
▷ 14h-16h	7	14.6%
▷ après 16h	6	12.5%
▷ heures libres	15	31.3%
TOTAL CIT.	48	100%

La différence avec la répartition de référence est très significative. $\chi^2 = 18.46$, ddl = 5, 1-p = 99.76%.
Le χ^2 est calculé avec des effectifs théoriques égaux pour chaque modalité.
Le tableau est construit sur 45 observations.
Les pourcentages sont calculés par rapport au nombre de citations.

- **Test du χ^2** : le χ^2 est calculé comme la somme des carrés des écarts aux effectifs théoriques (l'effectif théorique est la valeur de la case si la répartition était équilibrée).



3.3 Les différents niveaux de l'analyse des données

l'analyse à plat des variables nominales

- Le test du χ^2 est la certitude exprimée en pourcentage de la dépendance des deux variables.
- Selon la valeur de cette certitude, on dira que l'écart est très significatif ($1-p > 99\%$), significatif ($99\% > 1-p > 95\%$), peu significatif ($95\% > 1-p > 85\%$), non significatif ($1-p < 85\%$). Cette certitude est notée " $1-p$ ", p étant donc le risque de se tromper, qui est souvent utilisé comme référence.
- Les cases qui sont le plus importantes dans le calcul du χ^2 sont encadrées (jusqu'à concurrence de 60%). Si l'effectif est inférieur à l'effectif théorique, l'encadrement est en rouge sinon en bleu.



3.3 Les différents niveaux de l'analyse des données

l'analyse à plat des variables nominales

Se rend surtout la A quel moment de la journée vous y rendez-vous le plus souvent ?

Se rend surtout la	Nb. cit.	Intervalles de confiance
▷ matin	2	0.0% < 4.2 < 9.8%
▷ entre 12h-14h	18	23.8% < 37.5 < 51.2%
▷ 14h-16h	7	4.6% < 14.6 < 24.6%
▷ après 16h	6	3.1% < 12.5 < 21.9%
▷ heures libres	15	18.1% < 31.3 < 44.4%
TOTAL CIT.	48	

L'intervalle de confiance à 95% est donné pour chaque modalité.

Le tableau est construit sur 45 observations.

Les pourcentages sont calculés par rapport au nombre de citations.

Intervalles de confiance : affiche l'intervalle de confiance de chacune des modalités.

Ceux-ci tiennent compte du taux d'erreur qui dépend notamment de la taille de l'échantillon.



3.3 Les différents niveaux de l'analyse des données

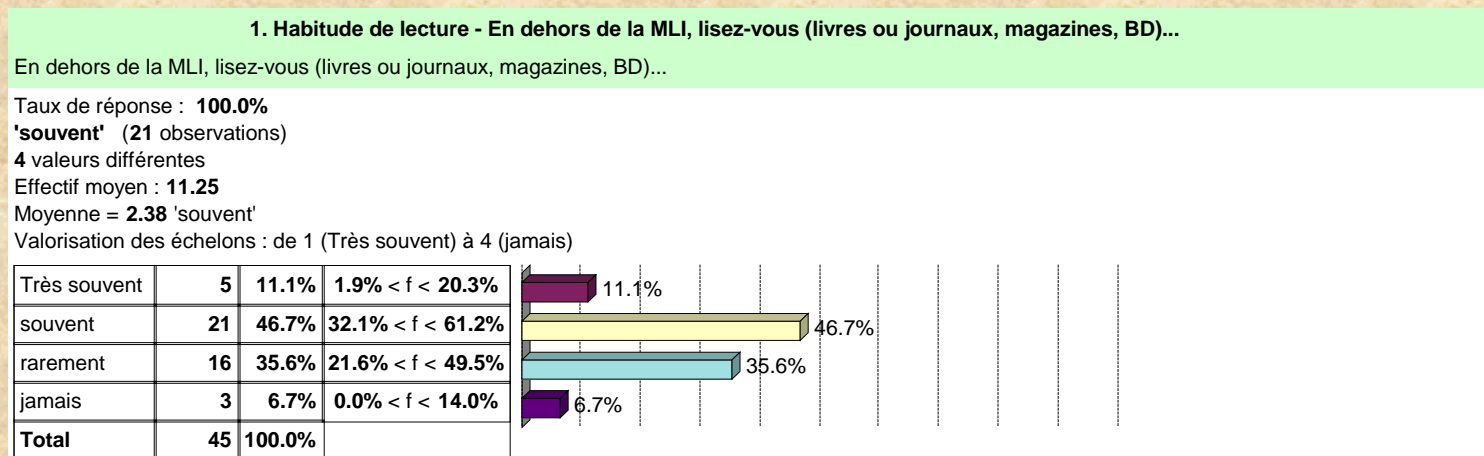
l'analyse à plat des échelles

Les questions "échelle" ont la particularité de pouvoir être traitées comme des questions fermées ou numériques. En effet, à chaque échelon correspond un nombre, de 1 à n.

1. En dehors de la MLI, lisez-vous (livres ou journaux, magazines, BD)...

- 1.Très souvent 2.souvent
 3.rarement 4.jamais

On peut analyser les données comme des questions fermées avec un tableau de fréquence et donc calculer des paramètres statistiques de position et de dispersion; évaluer le degré de précision des résultats

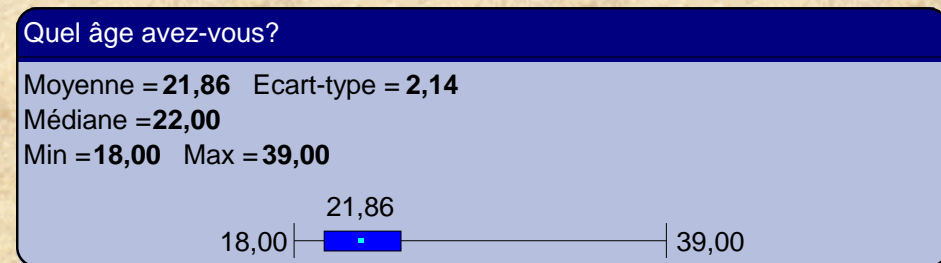


3.3 Les différents niveaux de l'analyse des données

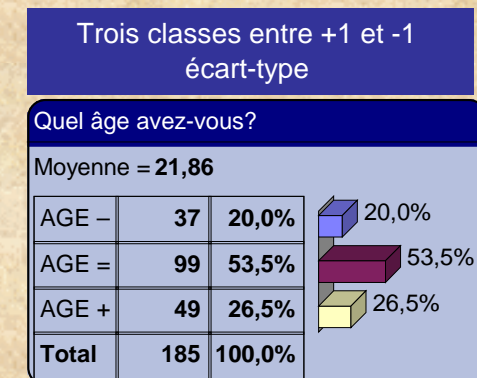
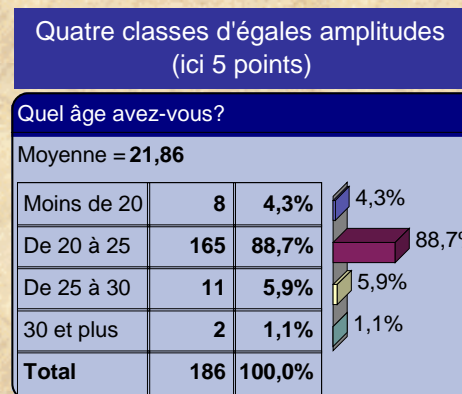
l'analyse à plat des variables quantitatives

Le traitement à plat d'une question numérique peut être présenté de différentes manières : *(exemple tiré de Sphinx développement)*

On peut se contenter des principaux paramètres de position et de dispersion (moyenne et écart-type en général) et de leur représentation graphique.



On préfère en général une mise en classes et celle-ci peut obéir à plusieurs logiques.



3.3 Les différents niveaux de l'analyse des données

Tris à plat des questions-textes (exemple tiré de Sphinx développement)

Pour dépouiller une question texte, il est possible de faire apparaître le lexique, la liste des mots les plus cités. On a exclu ici les mots outils (mots grammaticaux sans contenu). La liste a été limitée aux 18 mots les plus cités.

vie	122	10,9%
Réussir	81	7,2%
travail	55	4,9%
argent	45	4,0%
Gagner	44	3,9%
profiter	38	3,4%
famille	36	3,2%
bon	26	2,3%
amis	24	2,1%
Trouver	22	2,0%
Fonder	18	1,6%
personnelle	16	1,4%
garder	16	1,4%
faire	16	1,4%
célibataire	14	1,2%
activités	14	1,2%
job	11	1,0%
...	523	46,7%
Total	1121	100,0%

Le cas échéant, on peut présenter la liste des réponses, si leur diversité n'est pas trop grande. Ici, les réponses sont très diversifiées.

Réussir sa vie professionnelle et sa vie personnelle	2	1,1%
Gagner beaucoup d'argent	2	1,1%
Réussir sa vie professionnelle et sa vie personnelle, c'est réussir sa vie	1	0,5%
réussir sa vie, c'est d'abord réussir sa vie affective	1	0,5%
...	180	96,8%
Total	186	100,0%



3.3 Les différents niveaux de l'analyse des données :



3.32 l'analyse biva



3.32 L'analyse bivariée

- Il s'agit de comparer les réponses à chaque modalité de la variable A en fonction de des réponses à la question B.

Exemple : voir les comportements de lecture des lycéens en fonction de leur section.

- Pour étudier les relations entre 2 variables :
entre une variable explicative et une variable expliquée

Si..la section du lycéen... **alors**, les comportements de lecture...

Quand.. 'Insatisfait'... **alors**, 'Ne revient pas'

- Pour étudier le degré de convergence qui relie
deux variables quantitatives (*coefficient de corrélation*)
- Pour étudier le degré de significativité d'une relation entre deux variables
(*par le test du kish2 ou à partir d'une analyse de variance avec le test F de Fisher*)

**Le choix des traitements et tests statistiques entre deux variables
dépendra de la nature de ces dernières**



3.32 L'analyse bivariée

Variable expliquée

Variable explicative

<p>Si ↗ ↘ Alors</p>	<p>V2 Nominale</p>	<p>V2 Numérique</p>
<p>V1 Nominale</p>	<p>Dépendance <i>(ex: sexe et types de lecture)</i> Traitements : tris croisés et AFC Test statistique : test χ^2 – test de discrimination (belson)</p>	<p>Comparaison <i>(ex : sexe et nombre de livres lus)</i> Traitements : tableaux de moyennes et analyse de la variance Tests statistiques test de Fisher</p>
<p>V1 Numérique</p>	<p>Comparaison <i>(ex : sexe et nombre de livres lus)</i> Traitements : tableaux de moyennes Tests statistiques : analyse de la variance et test de Fisher</p>	<p>Corrélation <i>(ex : nombre livres lus et âge)</i> Traitements : régression Test statistique : corrélation</p>



3.32 / L'analyse bivariée : croiser deux variables nominales

- Le tableau de résultats d'un tri croisé est appelé tableau de contingence (étudier les contingences cad les relations et le contenu des relations entre deux variables.

en effectifs

Supports lus	Livres	BD	journaux	magazines	TOTAL
générale	3	2	2	15	22
technologie tertiaire	1	2	4	10	17
industrielle	0	1	1	2	4
TOTAL	4	5	7	27	43

Ces deux modes de représentation des données ne permettent pas une lecture croisée des données; il convient donc de retraiter ces données

en % total

Supports lus	Livres	BD	journaux	magazines	TOTAL
générale	6.7%	4.4%	4.4%	33.3%	48.9%
technologie tertiaire	2.2%	4.4%	8.9%	22.2%	37.8%
industrielle	0.0%	2.2%	2.2%	4.4%	8.9%
TOTAL	8.9%	11.1%	15.6%	60.0%	



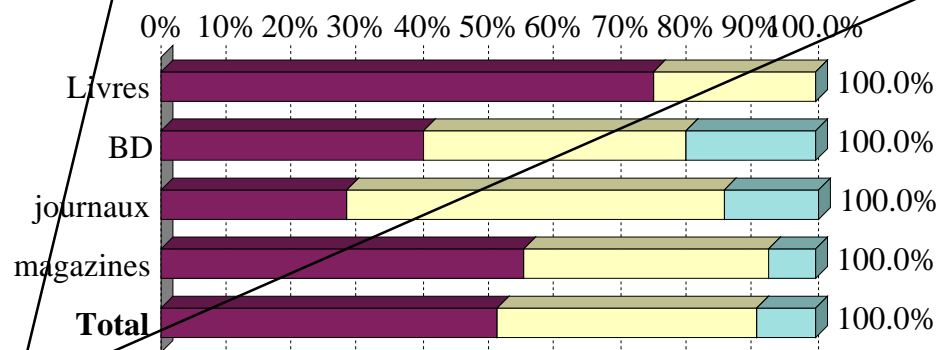
3.32 / L'analyse bivariée : croiser deux variables nominales

Pour permettre une lecture croisée des données, on les retranscrit en % lignes ou en % colonnes (« fréquences marginales lignes ou colonnes) :

fréquences marginales lignes

Support de lecture privilégié en fonction de la section

	générale	technologie tertiaire	industrielle	Total
Livres	75.0%	25.0%	0.0%	100.0%
BD	40.0%	40.0%	20.0%	100.0%
journaux	28.6%	57.1%	14.3%	100.0%
magazines	55.6%	37.0%	7.4%	100.0%
Total	51.2%	39.5%	9.3%	100.0%



Exemple de lecture : 55,6% des magazines sont lus par les lycéens issus d'une section générale.

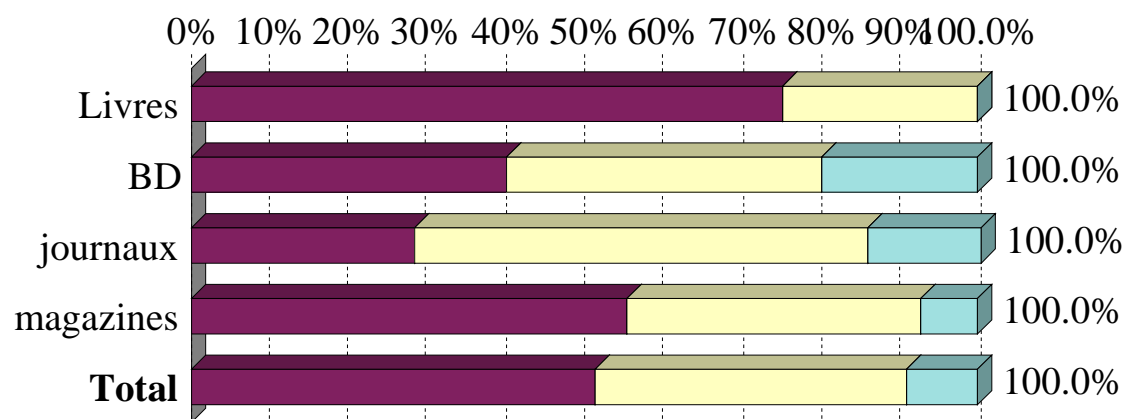


3.32 / L'analyse bivariée : croiser deux variables nominales

fréquences marginales colonnes

Support de lecture et section

	générale	technologie tertiaire	industrielle	Total
Livres	13.6%	5.9%	0.0%	9.3%
BD	9.1%	11.8%	25.0%	11.6%
journaux	9.1%	23.5%	25.0%	16.3%
magazines	68.2%	58.8%	50.0%	62.8%
Total	100.0%	100.0%	100.0%	100.0%



Exemple de lecture :

Sur l'ensemble des lycéens en section générale, 68,2% privilégient les magazines comme support de lecture premier.



3.32 / L'analyse bivariée : croiser deux variables nominales pour vérifier leur degré de dépendance – le test du chi2

Le test du Chi-deux indique si la relation entre les deux variables est significative.

Quels sont pour vous dans la liste suivante les trois principaux points à prendre en considération pour l'achat d'une automobile ?

Sexe de l'interviewé

	Vitesse	Confort	Sécurité	Consommation	Prix	Publicité	Distributeur	S.A.V. Entretien	Total
Homme	9.2%	5.9%	3.9%	34.6%	30.1%	2.6%	3.9%	9.8%	100.0%
Femme	29.4%	17.4%	14.9%	13.4%	5.0%	12.4%	4.0%	3.5%	100.0%

$p = <0.1\%$; $\chi^2 = 104.51$; $ddl = 7$ (TS)

Les couples de modalités en bleu (rose) sont sur-représentés (sous-représentés)

Les cases colorées nous montrent les informations essentielles cad les Chi-deux partiels les plus forts.

- en bleu, les sur-représentations
- en rose, les sous-représentations

Dans cet exemple, certains critères sont liés au genre : pour les hommes la consommation et le prix sont déterminants, alors que pour les femmes interrogées le critère premier est la vitesse...



3.32 / L'analyse bivariée : croiser deux variables nominales pour vérifier leur degré de dépendance –

Explication du test du chi2 (illustration tirée de Sphinx développement)

MARQUE SEXE	Français	Etranger	TOTAL
Homme	37	20	57
Femme	33	47	80
TOTAL	70	67	137

Comparer les effectifs observés à la référence de l'équi-répartition.

37 hommes achètent français

ils devraient être :

$$70 \times (57 / 137) = 30$$

*Le test du chi2
mesure l'écart à une répartition de référence
et évalue son importance*



3.32 / L'analyse bivariée : croiser deux variables nominales pour évaluer un caractère explicatif – le test de Belson

- **Pourquoi ?** : Quand on dispose de plusieurs facteurs explicatifs d'un phénomène observé, pour ces différents facteurs on peut se demander quel est celui qui dispose de la force explicative la plus importante cad le facteur discriminant.
- **Comment ?** : Par le critère de Belson (exclusivement sur des variables dichotomiques cad ne prenant que deux états possibles).
- **Exemple** : On se pose la question de savoir parmi les deux variables explicatives suivantes (sexe et PCS) quelle est celle qui détermine le mieux la satisfaction liée aux conditions de la mise en place des 35 heures dans une grande entreprise ? (Tableaux page suivante)



3.32 / L'analyse bivariée : croiser deux variables nominales pour évaluer un caractère explicatif – le test de Belson

En % global	Satisfait	non satisfait	Total
Masculin	50%	5%	55%
Féminin	30%	15%	45%
Total	80%	20%	100%

En % global	Satisfait	non satisfait	Total
PCS-	60%	3%	63%
PCS+	20%	17%	37%
Total	80%	20%	100%

Mode d'analyse :

Etape 1 : Comparer la situation réelle à la situation de référence

Une situation de référence correspond à ce que serait la réalité s'il n'y avait aucune relation entre la variable à expliquer et chacune des deux variables. Il faut alors mesurer la distance qui sépare la situation réelle de celle de référence. La meilleure variable explicative sera celle la plus éloignée de l'indépendance totale (la distance est élevée au carré pour obtenir des chiffres positifs)

L'indépendance totale suppose que :

	Satisfait	non satisfait	Total
Masculin	44%	11%	55%
Féminin	36%	9%	45%
Total	80%	20%	100%

	Satisfait	non satisfait	Total
PCS-	50.4%	12.6%	63%
PSC+	29.6%	7.4%	37%
Total	80%	20%	100%



3.32 / L'analyse bivariée : croiser deux variables nominales pour évaluer un caractère explicatif – le test de Belson

Etape 2 : Mesure de l'écart entre population réelle et théorique par le critère de Belson cad en mettant au carré la valeur absolue de l'écart déterminé dans chaque tableau :

$$\text{Sexe } d = 6^2 = 36$$

$$\text{PCS } d = 9.6 = 92.16$$

Donc $d_{\text{PCS}} > d_{\text{sexe}}$ soit la variable PCS est la plus éloignée de l'indépendance cad la plus près de la dépendance : *La PCS est une meilleure variable explicative que le sexe pour la satisfaction des conditions de la mise en place des 35heures*

Limite du critère de Belson :

Il s'adapte à une situation où les tableaux de contingence sont de taille réduite et les variables dichotomiques.

Dans les autres cas, prendre le khi 2



3.32 / L'analyse bivariée : croiser deux variables nominales *L'Analyse Factorielle des Correspondances (AFC)*

	PCS	CADRE SUPERIEUR/PROFESSION LIBERALE	COMMERCANT/ARTISAN	ETUDIANT	EMPLOYE/OUVRIER	CADRE MOYEN	RETRAITE	INACTIF	TOTAL
LOGEMENT									
LOUEUR		6.1%	3.0%	6.1%	45.5%	6.1%	36.4%	0.0%	100%
HOTEL		0.0%	0.0%	11.1%	33.3%	22.2%	0.0%	22.2%	100%
RESIDENCE SECONDAIRE		0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	25.0%	100%
CAMPING		0.0%	0.0%	31.6%	42.1%	5.3%	5.3%	10.5%	100%
AMIS/FAMILLE		2.8%	0.0%	19.4%	36.1%	8.3%	8.3%	11.1%	100%
TOTAL		2.9%	1.0%	16.3%	40.4%	8.7%	15.4%	8.7%	100%

La dépendance est très significative. $\chi^2 = 64.07$, ddl = 24, $1-p = >99.99\%$.

Les cases encadrées en bleu (rose) sont celles pour lesquelles l'effectif réel est nettement supérieur (inférieur) à l'effectif théorique.

Attention, 21 (60.0%) cases ont un effectif théorique inférieur à 5, les règles du χ^2 ne sont pas réellement applicables.

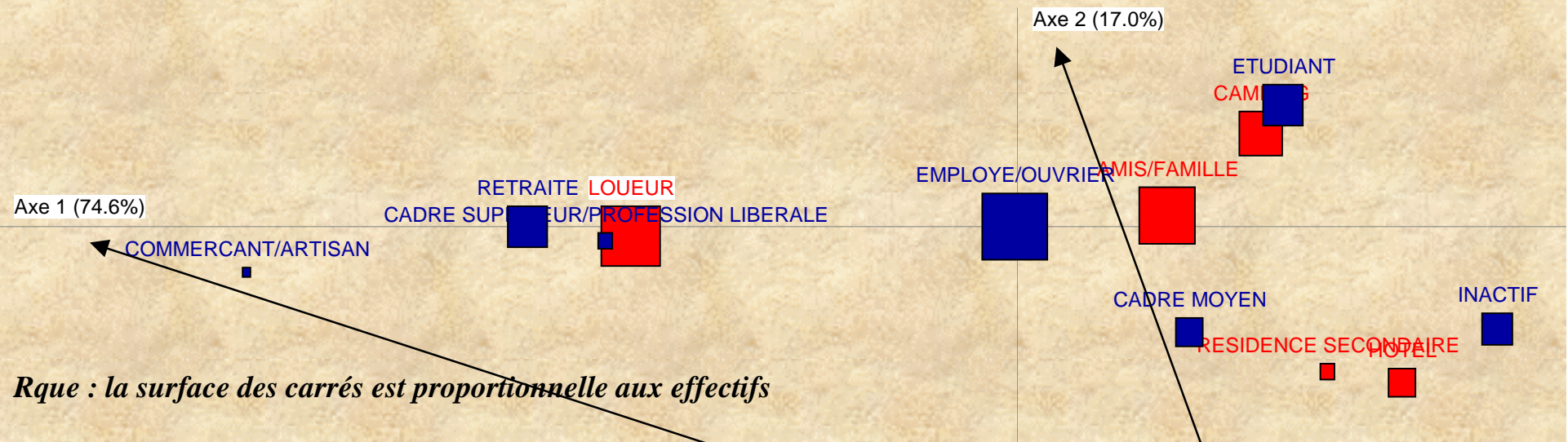
Le χ^2 est calculé sur le tableau des citations (effectifs marginaux égaux à la somme des effectifs lignes/colonnes).

Les valeurs du tableau sont les pourcentages en ligne établis sur 208 observations.

Pour mieux visualiser la relation entre ces deux variables, on représente les écarts à l'équi-répartition cad les dépendances entre des modalités des deux variables par la technique de l'AFC



3.32 / L'analyse bivariée : croiser deux variables nominales *L'Analyse Factorielle des Correspondances (AFC)*



Règle : la surface des carrés est proportionnelle aux effectifs

- La carte visualise les attractions et répulsions entre différentes modalités des deux variables
- Les axes affichent les % de variance expliquée : la carte restitue ici 91,6% de l'information initiale



3.32 / L'analyse bivariée : croiser deux variables nominales *L'Analyse Factorielle des Correspondances (AFC)*

Pour lire une carte, il faut s'appuyer sur les trois règles suivantes :

- 1** La règle d'éloignement par rapport au centre de gravité
Le centre de gravité exprime le comportement moyen : plus on s'en écarte, plus on observe des dépendances par rapport à la répartition.
- 2** La règle de proximité entre les modalités
Deux modalités proches dénotent une attractivité, deux modalités éloignées dénotent une répulsion.
- 3** La règle de superficie
C'est la dimension des carrés (ou des points) : plus le carré est important, plus la modalité a été citée.

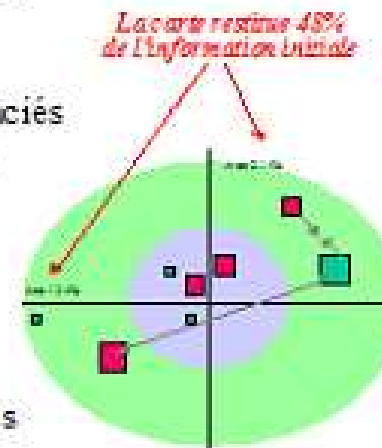
La qualité de la représentation dépend du % de variance expliquée par chacun des axes

Les zones de la carte :

- au centre les éléments peu ou mal différenciés
- à la périphérie les éléments remarquables

Les distances dans le plan de la carte :

- proximité = attraction ou forte valeur
- éloignement = répulsion ou faible valeur
- au centre peu de signification des distances



3.32 / L'analyse bivariée : croiser deux variables nominales

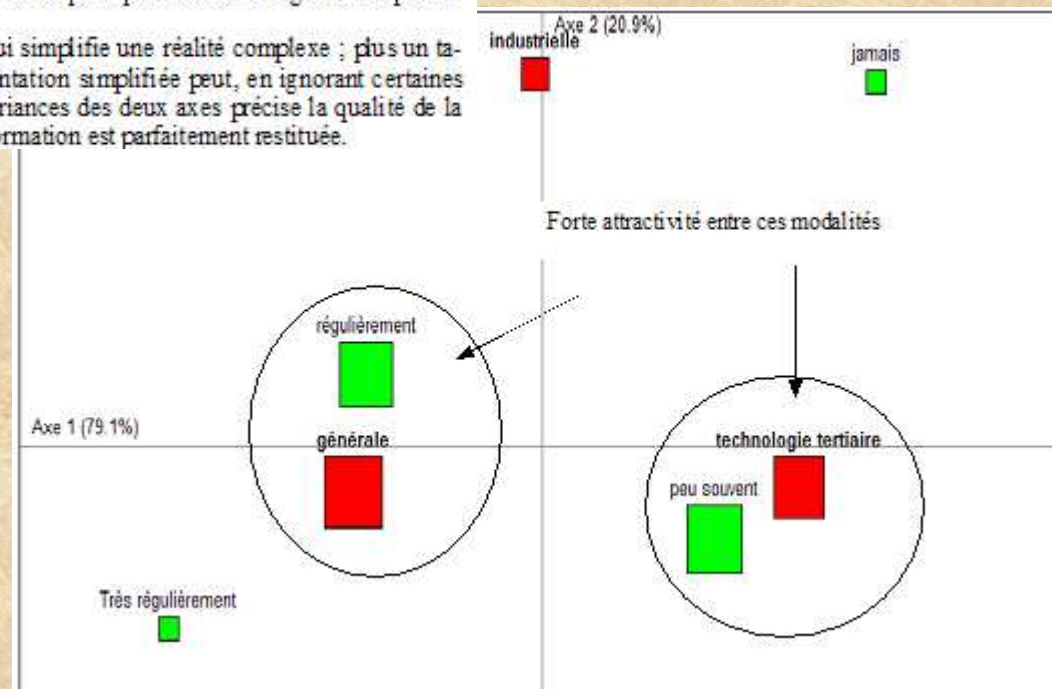
L'Analyse Factorielle des Correspondances (AFC) (autre exemple)

Fréquence MLI	Très régulièrement	régulièrement	peu souvent	jamais	TOTAL
Section					
générale	100%	68.4%	35.0%	0.0%	51.1%
technologie tertiaire	0.0%	15.8%	60.0%	66.7%	37.8%
industrielle	0.0%	15.8%	5.0%	33.3%	11.1%
TOTAL	100%	100%	100%	100%	100%

La dépendance est significative. $\chi^2 = 14.62$, ddl = 6, 1-p = 97.66%.

Si on se reporte au tableau (section*fréquence MLI), il faudrait 4 diagrammes empilés pour situer le degré de fréquentation de la MLI en fonction de la section des lycéens

A ces 4 diagrammes, l'AFC permet de substituer une carte perceptuelle qui simplifie une réalité complexe ; plus un tableau est grand, plus la représentation par l'AFC est utile ; Cette représentation simplifiée peut, en ignorant certaines données, l'éloigner de l'information initiale du tableau. La somme des variances des deux axes précise la qualité de la restitution de l'information : sur la représentation graphique suivante, l'information est parfaitement restituée.



3.32 / L'analyse bivariée : croiser une variable nominale

avec une variable numérique : Le tableau de moyennes croisées

• Un tableau de moyennes croisées permet d'évaluer une variable fermée en fonction de variables ouvertes numériques ou fermées échelles.

• Les modalités de la variable à évaluer apparaissent en ligne. Chaque critère d'évaluation occupe une colonne; si la case est cochée, le test sur le t de Student est appliqué pour comparer la moyenne de la case avec la moyenne de l'ensemble des observations étudiées.

Les cases qui sont significativement différentes de la moyenne sont encadrées (en bleu si la moyenne de la case est supérieure, en rouge si elle est inférieure).

Marque voiture croisée avec ...

	Quel est votre âge ?		Quel est le nombre de personnes composant le ménag...		Quelle distance effectuez-vous mensuellement ?		Combien dépensez-vous par mois ?	
	Moyenne	Ecart-type	Moyenne	Ecart-type	Moyenne	Ecart-type	Moyenne	Ecart-type
Renault	18.67	0.49	3.83	1.03	1 403.83	885.64	903.42	602.35
Peugeot	19.57	0.79	4.50	1.22	2 771.17	817.72	1 841.33	596.86
Citroën	18.67	0.98	3.82	1.17	1 249.27	694.58	796.00	432.03
Talbot	18.25	0.50	4.00		1 604.00		989.00	
Ford	20.00	1.76	3.40	0.52	2 403.40	842.85	1 483.00	520.72
Fiat	18.00	0.00	3.67	0.82	1 287.00	1 183.10	802.67	753.63
Volkswagen	19.68	1.86	3.69	1.65	3 065.23	255.16	1 882.92	198.04
Opel	20.00	2.14	2.50	0.53	2 802.50	961.61	1 729.50	653.72
Japonaise	19.33	1.95	3.00	0.00	3 203.00	0.00	1 977.00	0.00
Autre	19.19	1.42	3.25	1.18	2 165.88	1 070.23	1 343.38	678.90
Total	19.19	1.51	3.52	1.14	2 202.31	1 057.89	1 377.83	672.03

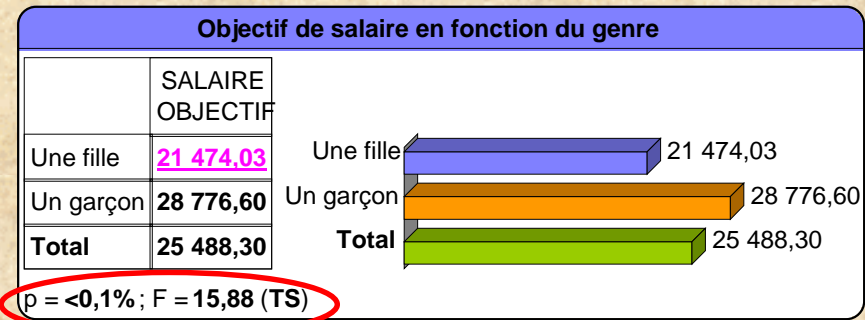
Les modalités en bleu (rose) sont sur-évaluées (sous-évaluées)



3.32 / L'analyse bivariée : croiser une variable nominale avec une variable numérique : Le test de Fisher (exemple de Sphinx développement)

- A partir d'une analyse des variances, le test F de Fisher nous indique si la relation entre les deux variables est significative.

les cases colorées nous montrent les catégories dont les moyennes sont statistiquement différentes de la moyenne générale (par le test de Student)



Le test de Fisher est significatif si la probabilité de rejet (p) est $< 5\%$ (ici il est très significatif avec $p < 0,1\%$)

Dans cet exemple, l'objectif de salaire varie en fonction du genre, d'une manière très significative :

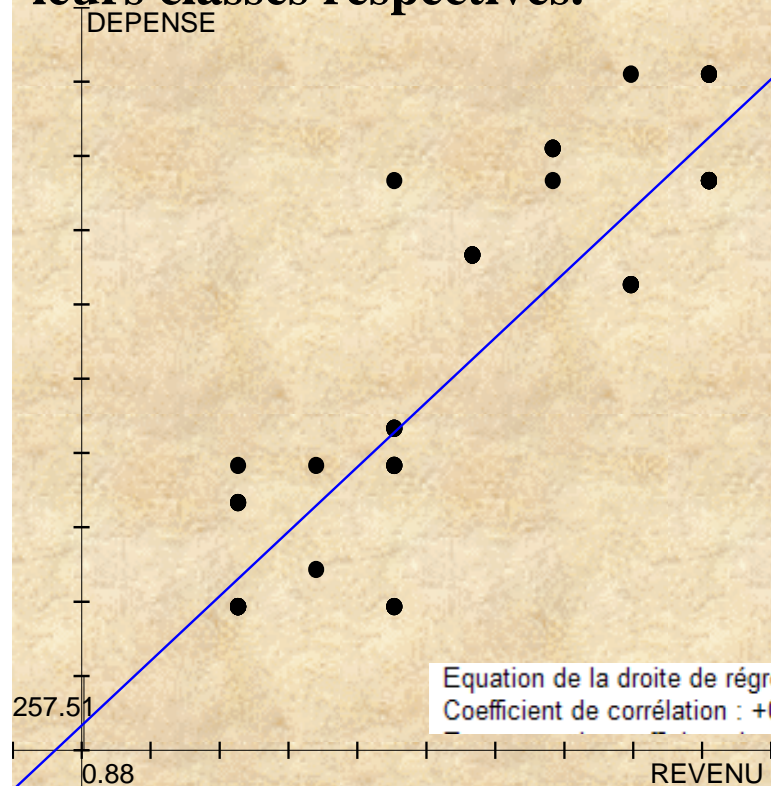
- les filles ont un objectif de salaire nettement inférieur à la moyenne



3.32 / L'analyse bivariée : croiser deux variables numériques :

La corrélation

- L'analyse permet de confronter deux variables correspondant à des nombres (c'est à dire ouvertes numériques ou bien fermées échelles). La première variable est la variable à expliquer (la dépense consacrée à son automobile), la seconde est une variable explicative (le revenu)
- Cette analyse est plus pertinente que le tableau croisé des deux variables avec leurs classes respectives.



La droite de régression linéaire, de type $y=ax+b$, permet de décrire une tendance, c'est à dire l'évolution générale de la dépense consacrée à son automobile en fonction du revenu.

Cette tendance est illustrée par une droite de régression qui ajuste linéairement le nuage de point.

Le coefficient de corrélation, ici =0,9 cad très forte corrélation positive (ce coefficient est toujours compris entre -1 et +1) caractérise la qualité de l'ajustement (donc très bonne qualité de l'ajustement)

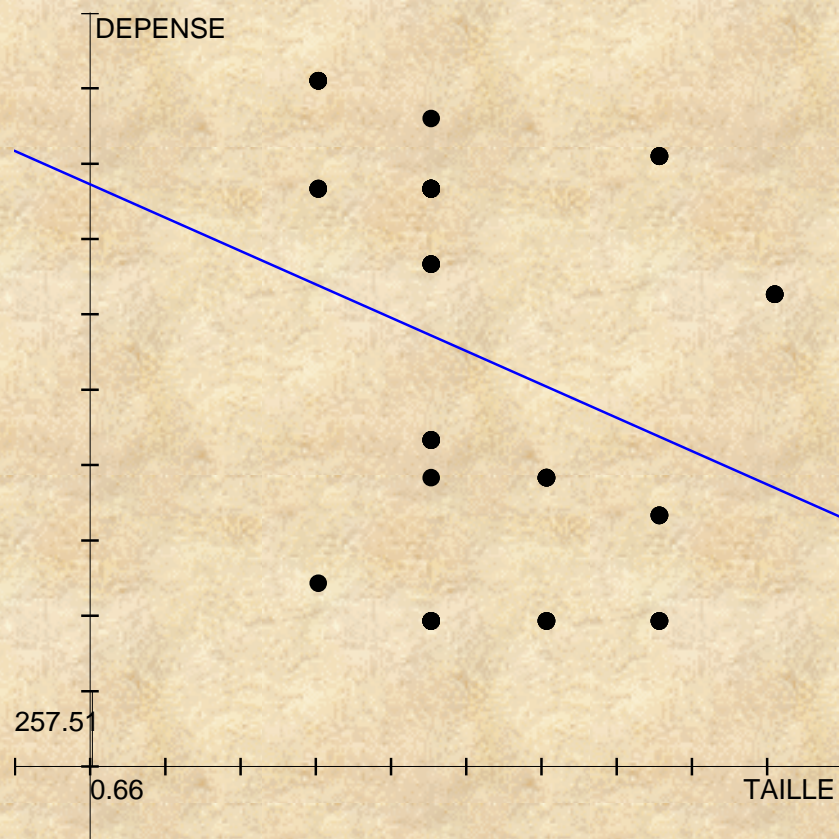
Equation de la droite de régression : $DEPENSE = 254.66 * REVENU + 86.67$
Coefficient de corrélation : +0.90 (REVENU explique 81% de la variance de DEPENSE)



3.32 / L'analyse bivariée : croiser deux variables numériques : La corrélation (autre exemple)

La dépense consacrée à son automobile en fonction du nombre de personnes composant la famille

Equation de la droite de régression : $DEPENSE = -172.60 * TAILLE + 1992.71$
Coefficient de corrélation : -0.29 (TAILLE explique 8% de la variance de DEPENSE)



L'ajustement est peu précis et ce degré d'imprécision est vérifié par un coefficient de corrélation sans signification ($r = -0,29$).



3.3 Les différents niveaux de l'analyse des données :



3.33 l'analyse multivariée



3.33 l'analyse multivariée

• Ses objectifs et modes d'expression

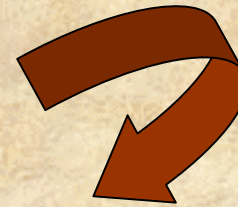
- Pour expliquer → *Régression multiple et Analyse de la variance*
- Pour synthétiser → *Analyse en Composantes principales et Analyse Factorielle des Correspondances multiples*
- Pour classifier → *Regrouper les individus par la classification Automatique et la typologie*



3.33 l'analyse multivariée : Pour expliquer

La régression multiple

- Pour déterminer sous forme d'équation linéaire la relation explicative amenant des variables (variables explicatives) à expliquer ou non un phénomène (variable expliquée)



Equation de régression multiple

Variables explicatives

$$V_0 = a_1 \times V_1 + a_2 \times V_2 + a_3 \times V_3 \dots + a_n \times V_n$$

La qualité de l'ajustement s'apprécie par rapport à la valeur du coefficient de corrélation



3.33 l'analyse multivariée : Pour expliquer

La régression multiple : exemple illustratif

Expliquer la dépense touristique totale V1 en fonction des dépenses d'hébergement V2, d'alimentation V3, de restauration V4 et de loisir V5

Choix des variables pour la régression multiple

Variable à expliquer : 7. Dépense totale

Variables explicatives : 7. Dépense totale, 8. Dépense hébergement, 9. Dépense alimentaire, 10. Dépense restaurant, 11. Dépense loisirs, 12. Soleil

Non-réponses : Ignorer, Valeur 0

4 éléments sélectionnés

Cette fonction permet d'obtenir l'équation de régression ainsi que la matrice des corrélations et divers profils des variables explicatives pour la variable à expliquer.

OK Annuler

On obtient un modèle du type
 $V1 = aV2 + bV3 + cV4 + \text{résidu.}$

La qualité de l'ajustement s'apprécie par rapport à la valeur du coefficient de corrélation. Plus la valeur absolue est élevée, plus faible est l'écart entre les valeurs calculées par l'équation et les valeurs observées en réalité (cet écart est appelé résidu) :

Equation de la régression :

Dépense totale = +0.796 * Dépense hébergement + 1.638 * Dépense alimentation + 0.734 * Dépense restaurant + 1.858 * Dépense loisirs + 174.856



3.33 l'analyse multivariée : Pour expliquer

La régression multiple : exemple illustratif

Equation de la régression :

Dépense totale = +0.796 * Dépense hébergement +1.638 * Dépense alimentation +0.734 * Dépense restaurant +1.858 * Dépense loisirs +174.856

- Les 4 variables expliquent 80.2% de la variance de Dépense totale et le coefficient de régression multiple = 0,9

Significativité des paramètres :

'Dépense hébergement' : coefficient = 0,80, écart-type = 0,23

'Dépense alimentation' : coefficient = 1,64, écart-type = 0,28

'Dépense restaurant' : coefficient = 0,73, écart-type = 0,45 (Peu influent)

'Dépense loisirs' : coefficient = 1,86, écart-type = 0,29

L'effet de chaque variable explicative dépend du coefficient de régression figurant dans l'équation. Plus celui-ci est élevé, plus la variable explicative considérée influence la variable expliquée.



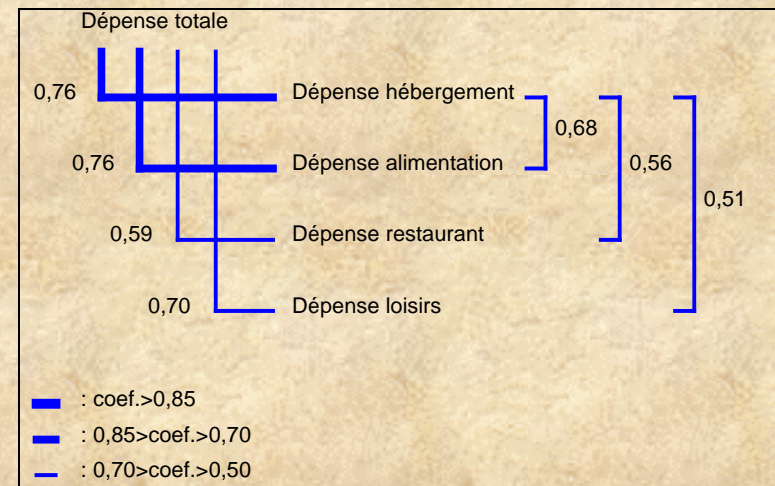
3.33 l'analyse multivariée : *Pour expliquer*

La régression multiple : exemple illustratif

- *Cependant, il faut également prendre en compte l'écart type de chacun de ces coefficients : plus il est élevé, moins l'influence de la variable considérée est marquée. Certains termes de l'équation sont peu influents, leur rapport coefficient / écart-type est inférieur à 2*

La matrice des coefficients de corrélation peut se présenter sous la forme d'un tableau ou d'un diagramme :

	Dépense totale	Dépense hébergement	Dépense alimentation	Dépense restaurant	Dépense loisirs
Dépense totale	1,00				
Dépense hébergement	0,76	1,00			
Dépense alimentation	0,76	0,68	1,00		
Dépense restaurant	0,59	0,56	0,48	1,00	
Dépense loisirs	0,70	0,51	0,41	0,45	1,00



3.33 L'analyse multivariée : *Pour expliquer*

L'analyse de la variance à deux facteurs (MANOVA)

- Pour comparer pour chaque modalité de deux variables nominales, la valeur d'une variable numérique

Exemple : analyser les valeurs de la dépense d'un séjour touristique selon le sexe et le mode d'hébergement choisi

	M	F	TOTAL
Hôtel	6254,84	4662,00	5543,75
Camping	3270,51	2194,12	2943,75
Location / gîte	5653,19	3882,89	4861,76
Famille / amis	2739,29	3480,00	3106,31
TOTAL	4280,64	3650,37	4004,38

Les valeurs du tableau sont les moyennes de la variable Dépense totale pour chaque couple de citations.

Analyse de la variance à deux facteurs :

- L'effet principal de 'Mode d'hébergement' est très significatif (V_inter = 115902378,97, V_intra = 14637378,43, F = 7,92.)
- L'effet principal de 'Sexe' n'est pas significatif (V_inter = 22382105,57, V_intra = 14637378,43, F = 1,53.)
- L'interaction de 'Mode d'hébergement' et 'Sexe' est peu significative (V_inter = 35840396,78, V_intra = 14637378,43, F = 2,45.)

La dépense est significativement différente selon le sexe pour les touristes en camping

Quelque soit le sexe, la dépense totale est significativement différente selon les modes d'hébergement sauf pour location/gîte



3.33 l'analyse multivariée : Pour synthétiser

- Pour rendre lisible une analyse multi-dimensionnelle en trouvant des facteurs qui permettent de réduire le nombre de dimensions cad de variables à considérer

Les techniques :

Variables numériques : *Analyse en composantes principales*

Variables nominales : *Analyse factorielle multiple*

(table individus x modalités)

Analyse factorielle des correspondances

(tableaux croisés)



3.33 l'analyse multivariée : Pour synthétiser

Analyse en Composantes Principales

- L'ACP reprend les individus en lignes et analyse les correspondances avec des variables numériques (critères) placées en colonnes

Exemple : (tiré de Sphinx développement); pour les touristes (individus) ce que sont les critères illustratifs de vacances idéales : soleil, sport, contacts, activités, confort, nature, le monde, le repos, la famille, le tout compris

Rque : ces critères apparaissent sous forme d'échelles dans le questionnaire et sont donc traités comme variables numériques

Problème : nous sommes ici dans un espace d'analyse à 10 dimensions qu'il convient de retranscrire dans un espace synthétique à deux dimensions (à l'image d'une photographie qui retranscrit en 2 dimensions l'espace réel composé de 3 dimensions quitte à sacrifier certains détails).



3.33 l'analyse multivariée : Pour synthétiser

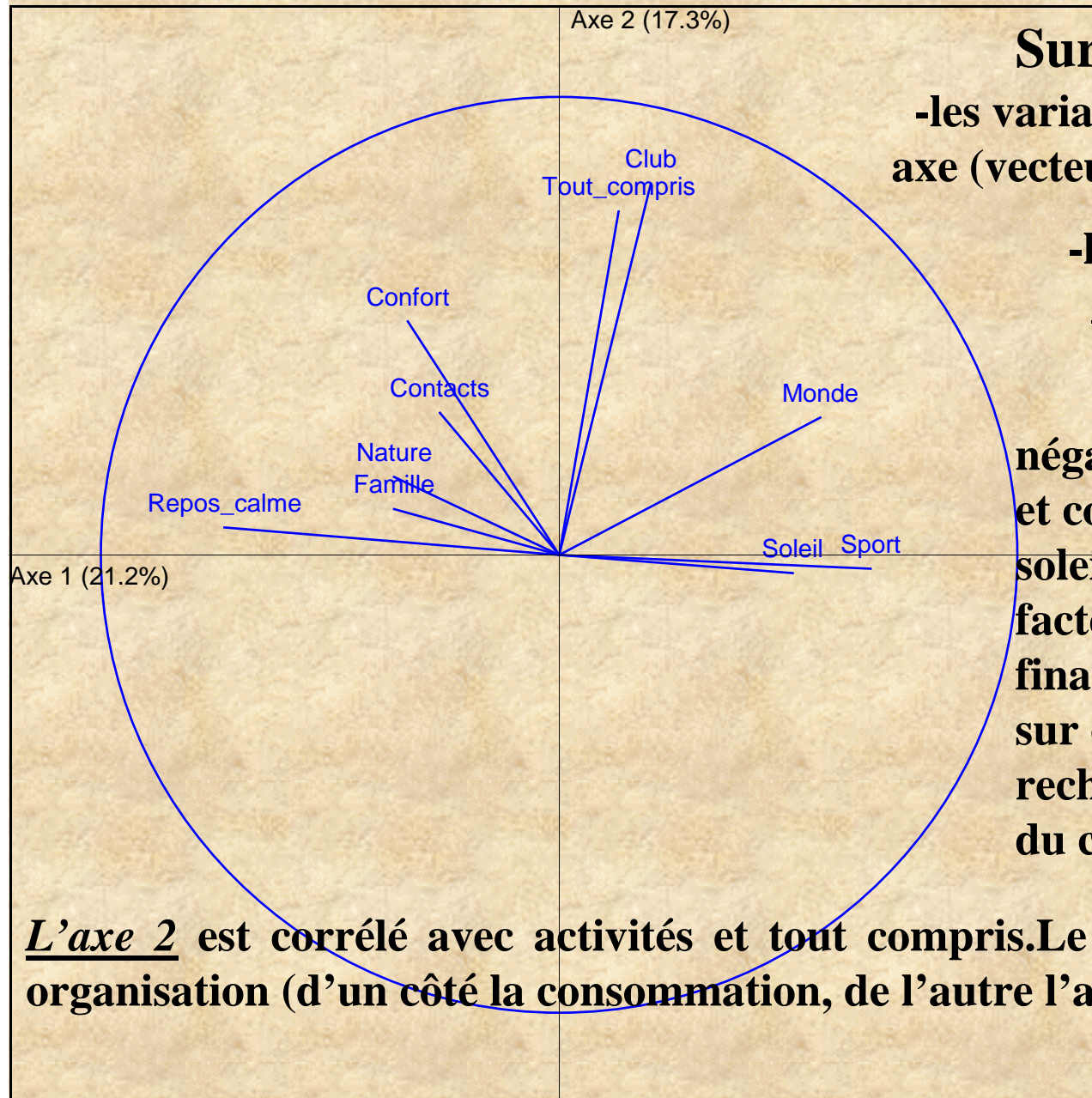
Analyse en Composantes Principales

- Les deux dimensions sont représentées par deux axes :
 - **l'axe principal 1** est la première composante principale cad qu'il représente l'indicateur pour laquelle la variance des individus est maximale afin d'intégrer un maximum d'observations.

Cet axe principal disposera les variables qui contribuent positivement à son sens .

- même démarche pour **le second axe** qui disposera de variables fortement corrélées avec lui mais qui sera indépendant du 1er axe





Sur la carte, on trouve :
-les variables représentées par un axe (vecteurs propres)

-le cercle de corrélation
-au niveau des facteurs :

L'axe 1 est corrélé négativement avec repos, calme et confort et positivement avec soleil, sport et monde. Le facteur 1 est donc relatif à la finalité des vacances (il oppose sur cette dimension la recherche de l'excitation à celle du calme et du repos)

L'axe 2 est corrélé avec activités et tout compris. Le facteur 2 est relatif à leur organisation (d'un côté la consommation, de l'autre l'autonomie).



3.33 l'analyse multivariée : Pour classifier et dresser des typologies

Analyse en Composantes Principales

•- au niveau des individus : dans le même plan factoriel, on peut représenter les individus :

les points situés à gauche ont un score élevé sur l'axe 1 et correspondent à des individus exprimant un degré d'accord élevé avec l'opinion selon laquelle les vacances idéales sont « repos-calme »; on trouvera donc à gauche ceux qui cherchent le calme, à droite ceux que le sport, le monde, le soleil attirent, en haut ceux qui cherchent les ambiances club, en bas les autonomes. On aboutit là à la construction d'une typologie.

