Durée de l'épreuve : 3 heures.
Épreuve sans document. La calculette est autorisée (sans sa documentation).
Les exercices (encadrés) sont indépendants. Le barème, sur 100, est donné à titre indicatif.
Indiquer les réponses exclusivement sur ce document et aux endroits réservés à cet effet (ne pas écrire dans la marge)

COURS: Description et Inférence (4 points)

1. Expliquer en quelques lignes, en quoi consiste la distinction entre l'étape descriptive et l'étape inférentielle lors de l'analyse statistique des données.

L'étape descriptive consiste à analyser et décrire les données recueillies sur un échantillon. La conclusion porte uniquement sur ces données observées. L'étape inférentielle consiste à évaluer s'il est possible de généraliser au-delà de l'échantillon,, à une population plus vaste (population parente), les conclusions de l'étape descriptive.

2. Lors de l'analyse de ses données expérimentales, un chercheur s'attend à ce qu'un facteur B à deux modalités n'ait pas d'effet. Il teste l'hypothèse nulle selon laquelle l'effet parent (effet vrai) est égal à 0 (δ =0). Pour cela il calcule un T de Student et trouve un résultat non significatif.
Quelles précautions doit-il prendre lors de l'interprétation de ce résultat non significatif?

L'hypothèse nulle d'absence d'effet est compatible avec les données, mais cela ne veut pas dire que le facteur B n'a pas d'effet. D'autres hypothèses peuvent être également compatibles avec ces données $\Rightarrow$ constat d'ignorance.

Dossier INTERNET (13 points)

Dans une enquête sur le réseau Internet auprès de 1006 personnes, une des questions posées était la suivante: "Personnellement quelle est votre attitude à l'égard de cette nouvelle application de la micro-informatique? Vous êtes... Passionné (PASS), Intéressé (INTE), Indifférent (INDI), Dépassé (DEPA), Agacé (AGA)" ou ne se prononce pas (NSPA). Les personnes interrogées ont été regroupées en 5 catégories d'âge: 18-24 ans, 25-34 ans, 34-49ans, 50-64 ans, plus de 65 ans.
Les effectifs des différentes réponses en relation avec la catégorie d'âge sont reportés dans le tableau ci-dessous:

	PASS	INTE	INDI	DEPA	AGA	NSPA
18-24	15	72	34	10	7	0
25-34	5	84	68	10	7	0
34-49	9	112	112	23	26	6
50-64	9	96	144	21	24	6
PL-65	2	28	52	16	6	2

Source: Sondage Institut CSA - "Aujourd'hui" - "Le Parisien" (publié dans "SVM" Juillet-Août 1996), réalisé du 22 mai au 23 mai 1996 auprès d'un échantillon national représentatif de 1006 personnes âgées de 18 ans et plus, d'après la méthode des quotas (sexe, âge, catégorie socioprofessionnelle du chef de ménage), après stratification par région et taille d'agglomération.

A. Analyse des marges (2 points)

Une règle pour mener l'analyse d'un tel tableau veut que l'on commence par en analyser les marges. En particulier, tous âges confondus, on a le résultat suivant.

PASS	INTE	INDI	DEPA	AGA	NSP
4%	39%	41%	8%	7%	1%

a/ Commenter ces valeurs

L'essentiel des réponses se partage entre Intéressé (39%) et Indifférent (41%). Peu de gens donnent une réponse extrême: seulement 4% répondent Passionné 4%), 8% Dépassé et 7% Agacé (7%).

b/ Calculer le *pourcentage* de réponses "Passionné", *arrondi à 3 décimales* (xx.xxx %). Indiquer le calcul effectué :

*40/1006 * 100 = 3.98%*

B. Analyse de la liaison Âge × Réponse (8 points)

Question initiale : On se demande s'il existe une liaison entre le type de réponse et l'âge de la personne interrogée.

Pour cela on calcule tout d'abord le tableau des profils:

	PASS	INTE	INDI	DEPA	AGA	NSPA	TOTAL
18-24	11	52	25	7	5	0	100
25-34	3	48	39	6	4	0	100
34-49	3	39	39	8	9	2	100
50-64	3	32	48	7	8	2	100
PL-65	2	26	49	15	6	2	100
MOY	4	39	41	8	7	1	100

a/ à la vue de ce tableau un commentateur affirme "11% de ceux qui sont passionnés par Internet ont entre 18 et 24 ans". Commenter cette affirmation:

*Il y a inversion des proportions: 11% s'interprète ainsi: 11% des 18-24 ans sont passionnés par Internet. Ce sont en fait 37.5**%** des passionnés d'Internet qui ont entre 18 et 24 ans (15/40.*

b/ donner une première réponse à la question posée en indiquant les faits majeurs qui ressortent du tableau des profils :

On trouve le plus grand % de réponses Passionné et Intéressé chez les jeunes. Le % de réponses Indifférent augmente avec l'âge. La fréquences de Dépassé double chez les +65 ans.

c/ on calcule le tableau des effectifs théoriques. Compléter ce tableau en arrondissant à l'entier le plus près. Indiquer la procédure de calcul:

	PASS	INTE	INDI	DEPA	AGA	NSPA
18-24	5	54	56	11	10	2
25-34	7	68	71	14	12	2
34-49	11	112	117	23	20	4
50-64	12	117	122	24	21	4
PL-65	4	41	43	8	7	1

d/ construire le tableau des attractions-répulsions:

	PASS	INTE	INDI	DEPA	AGA	NSPA
18-24	+	+	−	−	−	−
25-34	−	+	−	−	−	−
34-49	−	=	−	=	+	+
50-64	−	−	+	−	+	+
PL-65	−	−	+	+	−	+

e/ répondre à nouveau à la question posée en commentant le tableau des attractions-répulsions.

On trouve beaucoup de réponses Passionné et Intéressé chez les jeunes (<34 ans) et beaucoup de réponses Indifférent, Dépassé, Agacé et Ne sait pas chez les plus âgés.

f/ on souhaite visualiser les distances entre les profils des différentes catégories d'âges. Quelle méthode peut-on employer pour cela ? Justifier votre réponse :

L'Analyse Factorielle des Correspondances (AFC).
On analysera le graphe factoriel des lignes.

C. Analyse inférentielle (3 points)

S'agissant d'un sondage, on souhaite en fait se prononcer sur l'existence de telles différences de réponses liées à l'âge *dans l'ensemble de la population française*.

a/ rappeler la principale condition préalable pour être en mesure de répondre à cette question.

Cet échantillon doit être un échantillon tiré au hasard dans (ou représentatif de) la population française.

b/ quelle procédure inférentielle peut-on mettre en oeuvre?

Test du χ^2

c/ calculer le nombre de degrés de liberté associé à cette procédure :

ddl = (5-1) (6-1) = 4 x 5 = 20

COURS: Régression et corrélation (4 points)

1. Que devient le coefficient de corrélation linéaire de Bravais-Pearson entre deux variables X et Y si l'on divise toutes les valeurs de X et de Y par 2 ? Justifier votre réponse :

Le coefficient de corrélation linéaire de Bravais-Pearson est invariant pour toute transformation linéaire de l'une ou des deux variables.

2. Lors de la régression d'une variable Y sur un ensemble de k variables X1, X2...Xk, on trouve un coefficient de détermination $R^2 = 0.88$

a. Entre quelles limites peut varier cet indice?

Le coefficient de détermination varie entre 0 et +1.

b. Interpréter la valeur observée de R^2 :

Les variables X1, X2... Xk rendent comptent de 88% de la variance de Y, ce qui peut être considéré comme une proportion importante (proche de 1). La régression linéaire de Y sur les K variables X1, X2, ... Xk est très bonne.

3. Quel est l'objectif visé en calculant une corrélation partielle ?

On cherche à connaître la corrélation linéaire entre X et Y, en maintenant Z constant (en éliminant l'influence de Z sur chacune de ces deux variables).

PAPILLONS: Classification Ascendante Hiérarchique (4 points)

Au cours du printemps 1985, des enfants ont attrapé des papillons. Pour occuper un jour de pluie, ils les ont mesurés. A partir des mesures recueillies, on cherche à déterminer le nombre d'espèces présentes dans cette petite population de 23 papillons.

Source: Robert C; (1989) Analyse Descriptive Multivariée: Application à l'Intelligence Artificielle. Paris: Flammarion. Coll. Statistique en Biologie et en Médecine.

En procédant à la CAH, on obtient l'arbre suivant (figure 1):

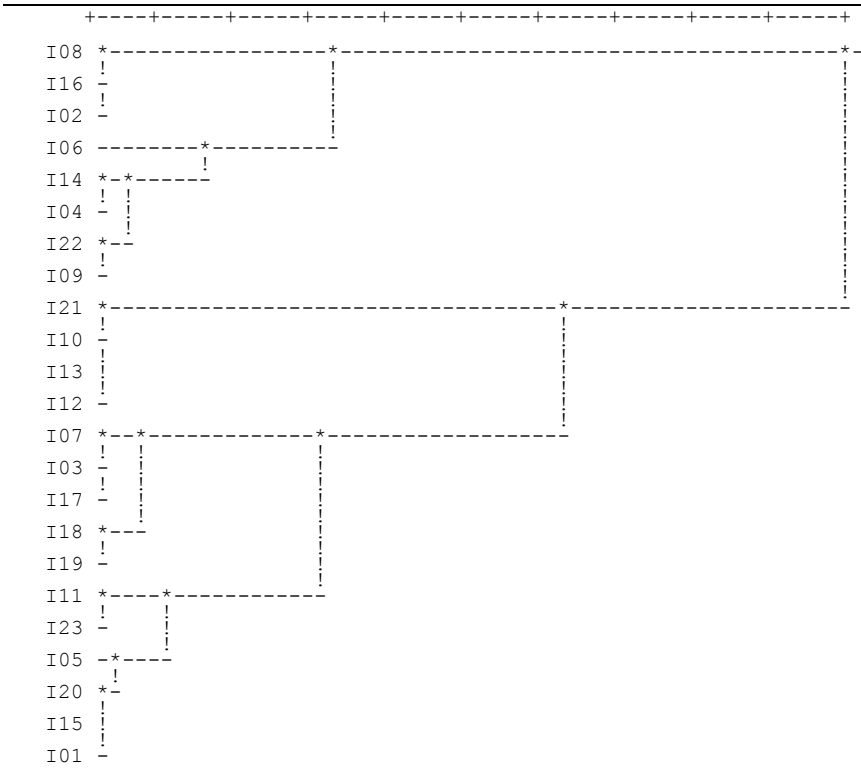

Figure 1: Dendrogramme de la Classification Ascendante Hiérarchique des papillons.

En coupant l'arbre, on définit 3 classes de papillons. Indiquer, pour chacune des ces 3 classes, les numéros des papillons qui la compose.

Classe 1: {I08, I16, I02, I06, I14, I04, I22, I09}
Classe 2: {I21, I10, I13, I12}
Classe 3: {I07, I03, I17, I18, I19, I11, I23, I05, I20, I15, I01}

HORLOGE: Structures (10 pts)

Une expérience porte sur l'exploration mentale d'un environnement imaginé (Amorim, Stuchhi, 1994). Le sujet voit la lettre "F" sur un écran (tridimensionnelle, projetée en 2 dimensions) et doit imaginer que la lettre se trouve au centre d'une horloge. Dans la condition c1 "sujet-centrée", on indique au sujet l'heure à laquelle il se trouve sur l'horloge, et la tâche consiste à donner l'heure à laquelle pointe le haut de la lettre; dans la condition c2 "objet-centrée", on indique l'heure à laquelle pointe la lettre et le sujet doit donner l'heure à laquelle il se trouve. Les 24 sujets (Facteur S) de l'expérience passent chacun les deux conditions, avec 12 essais par condition. Les 12 essais (facteur E) correspondent au croisement de: 6 angles (Facteur A) que peut faire la lettre avec le sujet (15°, 45°, 75°, 105°, 135°, 165°), et 2 côtés possibles, gauche ou droit (Facteur L, "Latéralité"). Enfin les sujets sont répartis en 4 groupes (facteur G), de 6 sujets chacun, obtenus par le croisement de: 2 dimensions suggérées de l'horloge (d1: 3m, d2: 30m); et 2 ordres de passation (o1, o2) des 2 conditions. On mesure l'erreur en degrés, une erreur positive indiquant une surestimation de l'angle.

Utiliser la notation indicée pour l'écriture des facteurs composés

1. Indiquer les relations (en utilisant les symboles <> et *) entre les facteurs :

   *S et C : S24*C2*

   *S et E : S24 * E12*

   *S et L : S24 * L2*

   *S et A : S24 * A6*

2. Indiquer la relation entre le facteur E12 et le facteur composé A6∗L2:

Relation de confusion :
E12 ~ A6 ∗ L2 ou E1<A6*L2>

3. Indiquer la relation entre le facteur G4 et le facteur composé D2∗O2:	Relation de confusion : G4 ~ D2 ∗ O2 ou G1<D2*O2>
4. Indiquer la relation entre les 3 facteurs S, D et O:	S6<D2 ∗ O2>
5. Indiquer la relation entre les 4 facteurs S, C, A et L:	S24 ∗ C2 ∗ A6 ∗ L2
6. Indiquer la formule du plan minimal:	S24 ∗ C2 ∗ E12 ou S24 ∗ C2 ∗ A6 ∗ L2
7. Indiquer la formule du plan le plus riche:	S6<D2 ∗ O2> ∗ C2 ∗ A6 ∗ L2 ou S6<G1<D2*O2>>*C2*E1<A6*L2>
8. Quel est le nombre total d'observations?	576
9. Quelle est la variable dépendante?	L'erreur en degrés entre l'heure réelle et l'heure estimée

PUBLICITE: Analyse Factorielle des Correspondances (17 pts)

Un corpus de 200 publicités télévisées visant la consommation de produits plus spécialement destinés aux enfants a été analysé. Ces publicités concernaient 11 catégories de produits: laitages nature (LNAT), laitages fruités (LFRU), des fromages frais (FROF), des fromages-pâtes (FROP), des eaux minérales (EAUX), des boissons sucrées (BOIS), des produits pour petits déjeuners (PDEJ), des jus de fruits (JUSF), confiseries (CONF), barres chocolatées (BARR), chocolats (CHOC).
Ces publicités s'appuient sur la mise en scène de différentes valeurs: Gourmandise/Plaisir (GOUR), Nature/Ecologie (NATU), Forme/santé (SANT), Aventure/Evasion (AVEN), Tendresse/Affection (TEND), Prestige/Luxe (PRES), Séduction/Erotisme (SEDU), Convivialité/Partage (CONV), Tradition/Gastronomie (TRAD), Innovation/Modernisme (INOV), Folie/Délire (FOLI).
Une publicité pour un produit peut s'appuyer sur plusieurs valeurs.
Les données rassemblées dans le tableau de contingence suivant (Tableau 1) indiquent le nombre de fois où une valeur a été associée à un produit.

Source: M. Watiez (1992) - Approche psychosociologique du processus de socialisation alimentaire chez l'enfant français. Etude du rôle de la publicité télévisée dans la formation des représentations sur l'alimentation. Thèse de doctorat. Université Paris V.

Tableau 1: Données PUBLICITE

	GOUR	NATU	SANT	AVEN	TEND	PRES	SEDU	CONV	TRAD	INOV	FOLI	Total
LNAT	1	4	9	2	7	2	1	2	2	2	0	32
LFRU	7	5	6	0	5	1	0	2	1	0	0	27
FROF	12	10	1	0	7	5	8	4	5	2	0	54
FROP	12	8	0	4	3	6	4	6	9	6	0	58
EAUX	1	9	13	4	2	4	1	0	0	0	2	36
BOIS	3	3	8	10	2	0	6	2	0	1	1	36
PDEJ	11	5	10	7	0	0	0	3	0	1	2	39
JUSF	0	4	3	1	0	3	2	2	0	0	0	15
CONF	5	4	1	5	1	0	4	4	0	2	3	29
BARR	19	8	6	15	4	1	2	0	1	1	3	60
CHOC	11	1	4	4	5	12	5	7	1	0	4	54
Total	82	61	61	52	36	34	33	32	19	15	15	440

On cherche à caractériser les profils des différents produits (fréquences des valeurs associées).

Tableau 2: Profils des différents produits (en %)

	GOUR	NATU	SANT	AVEN	TEND	PRES	SEDU	CONV	TRAD	INOV	FOLI	Total
LNAT	3	13	28	6	22	6	3	6	6	6	0	100
LFRU	26	19	22	0	19	4	0	7	4	0	0	100
FROF	22	19	2	0	13	9	15	7	9	4	0	100
FROP	21	14	0	7	5	10	7	10	16	10	0	100
EAUX	3	25	36	11	6	11	3	0	0	0	6	100
BOIS	8	8	22	28	6	0	17	6	0	3	3	100
PDEJ	28	13	26	18	0	0	0	8	0	3	5	100
JUSF	0	27	20	7	0	20	13	13	0	0	0	100
CONF	17	14	3	17	3	0	14	14	0	7	10	100
BARR	32	13	10	25	7	2	3	0	2	2	5	100
CHOC	20	2	7	7	9	22	9	13	2	0	7	100
Moy	19	14	14	12	8	8	8	7	4	3	3	100

1. Analyse des profils (cf. Tableau 2).

a/ Comment a été obtenue la première valeur, 3%, du profil de LNAT (case en haut à gauche) ?

3% = 0.03 = 1 / 32.

b/ Caractériser le profil de Chocolat (CHOC) par rapport au profil moyen.

Par rapport au profil moyen (par rapport à l'ensemble des autres produits) Chocolat est souvent associé à Prestige, Convivialité et Folie. Il est par contre particulièrement peu associé à Nature, Santé, Aventure et à Innovation.

c/ Quel indice peut-on calculer pour mesurer la distance entre ces deux profils (indiquer le nom de cet indice) ?

La distance du ϕ^2 entre les deux profils.

2. On a procédé à une analyse factorielle des correspondances (AFC). Un des résultats est le tableau des valeurs propres (cf. tableau 3).

a/ Qu'est-ce qu'une valeur propre?

C'est la variance d'une variable factorielle

Tableau 3: Valeurs propres de l'AFC

```
 ------------------------------------------------------------------------------
 !NUM ! VAL PROPRE ! POURC.! CUMUL !VARIAT.!*! HISTOGRAMME DES VALEURS PROPRES
 ------------------------------------------------------------------------------
 !  1 !     .19732 ! 34.890! 34.890!*******!*!***************!***************!
 !  2 !     .12182 ! 21.539! 56.429! 13.350!*!***************!****
 !  3 !     .08850 ! 15.649! 72.078!  5.890!*!*************
 !  4 !     .05778 ! 10.216! 82.293!  5.433!*!*********
 !  5 !     .03579 !  6.329! 88.622!  3.887!*!*****
 !  6 !     .02834 !  5.011! 93.633!  1.318!*!****
 !  7 !     .02107 !  3.725! 97.358!  1.286!*!***
 !  8 !     .01156 !  2.044! 99.402!  1.682!*!**
 !  9 !     .00319 !   .565! 99.967!  1.479!*!
 ! 10 !     .00019 !   .033!100.000!   .532!*!
```

b/ Compte tenu des informations fournies dans le tableau 3, vous paraît-il possible de s'intéresser uniquement aux deux premiers axes factoriels. Justifier votre réponse.

Les deux premiers axes totalisent (cumulent, rendent compte) de 56.4% de la variance du nuage (cf. colonne CUMUL). Ce pourcentage est relativement faible. Il paraît donc souhaitable de retenir un plus grand nombre d'axes pour rendre compte d'un plus fort pourcentage de la variance totale.

Si l'on prend l'autre critère qui consiste à retenir les axes pour lesquels la valeur propre représente un pourcentage de la variance du nuage supérieur à 1/ L (L = nombre d'axes), soit ici 1/10 = 10%, on devrait donc retenir les 4 premiers axes (cf. colonne POURC.) et non pas seulement 2.

c/ La somme des valeurs propres est égale à 0.56556. A quoi correspond cette valeur (on attend deux réponses)?

Cela correspond à la variance totale du nuage, c'est-à-dire au .

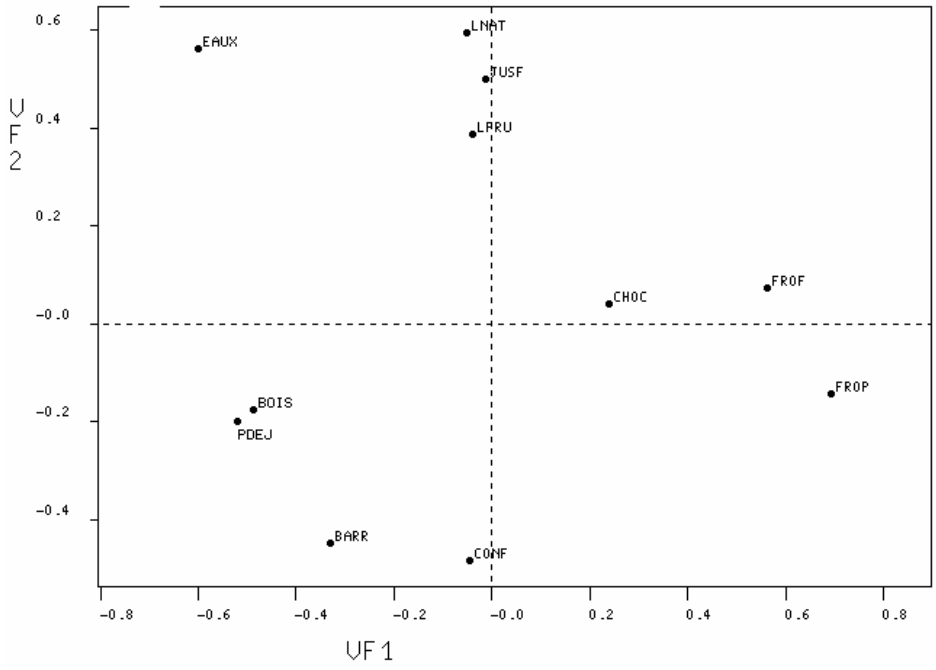

Figure 1: Nuage des produits dans l'espace des axes factoriels 1 et 2.

3. A quoi correspond l'intersection des axes factoriels?

L'intersection des axes factoriels correspond au centre de gravité du nuage, c'est-à-dire au profil moyen.

4. On constate que EAUX et FROP sont éloignés l'un de l'autre sur le graphique.

a. Qu'est-ce que cela suggère?

Cela suggère qu'ils ont des profils différents (relativement aux autres): ce ne sont pas les mêmes valeurs qui leur sont associées.

b. Vérifier en commentant les profils de ces deux produits (cf. Tableau 2, page précédente).

Les écarts de fréquences entre les éléments des deux profils sont souvent très importants. En particulier, Gourmandise, Convivialité, Tradition, Innovation sont beaucoup plus souvent associés aux Fromages qu'aux Eaux minérales, alors que Nature, Santé sont plus fréquemment associés aux Eaux minérales qu'aux Fromages.

Tableau 4: AFC des données PUBLICITE. Extrait des sorties du logiciel Addad (les valeurs sont exprimées en millièmes).

	I1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	LNAT	722	73	81	-51	4	1	595	564	212	311	154	80
2	LFRU	495	61	55	-39	3	0	387	295	76	316	197	69
3	FROF	755	123	95	563	724	197	73	12	5	91	19	11
4	FROP	847	132	154	693	726	321	-144	31	22	242	89	87
5	EAUX	850	82	115	-600	451	149	563	396	213	-43	2	2
6	BOIS	456	82	86	-485	398	98	-178	54	21	-53	5	3
7	PDEJ	690	89	74	-521	577	122	-199	84	29	118	29	14
6	JUSF	544	34	51	-14	0	0	500	298	70	-454	246	79
9	CONF	513	66	63	-45	4	1	-483	432	126	-204	77	31
10	BARR	781	136	103	-331	256	76	-448	468	224	155	56	37
11	CHOC	847	123	124	239	100	35	40	3	2	-651	744	587
			1000			1000			1000			1000	

5. La colonne POID nous renseigne sur le poids de chaque point dans l'analyse. La colonne INR indique l'inertie ou contribution relative (en millièmes) du point à la variance du nuage.
Expliquer comment retrouver par le calcul la valeur 73 (Poids de LNAT, en millièmes).

$$poids = n_j / n = 32 / 440 = 0.073$$

6. Contributions des points aux axes.

a. Rappeler le critère habituel pour retenir les principales contributions.

On retient les contributions supérieures à la contribution moyenne, c'est-à-dire > 1/J ou 1/K selon le cas, soit ici > 1/11 = 91/1000.

b. Indiquer dans le tableau suivant, les points qui contribuent le plus à l'axe 1.

-	+
EAUX (149)	*FROP (321)*
PDEJ (122)	*FROF (197)*
BOIS (98)	

7. Retour aux données: Expliciter, en se référant au tableau des profils (Tableau 2), ce qui distingue principalement ces deux groupes de produits s'opposant sur l'axe 1 :

Les valeurs Santé, Aventure, Folie sont fréquemment associées à EAUX, PDEJ et BOIS et peu fréquemment associées aux Fromages.

La valeur de Tradition est souvent associée aux Fromages et peu souvent à EAUX, PDEJ et BOIS.

Protocoles Dérivés Pertinents (12 points)

Supposons que le tableau suivant rapporte les résultats d'une expérience pour laquelle la formule du plan de recueil des données était : S<A2>*B2*C2.
On s'intéresse à différents effets, notés ci-dessous de manière formelle (A, B/c1, A/c1, A.B, B/c1a1, B.C). Pour chacun de ces effets :

1. Calculer, et reporter dans la colonne correspondante du tableau de droite, le protocole dérivé pertinent (PDP) permettant de calculer l'effet calibré et le T de Student.

2. Indiquer dans la dernière ligne du tableau ("D/M") si l'effet moyen peut être obtenu, soit en calculant la moyenne générale du PDP (dans ce cas noter "M" dans la case du PDP correspondant), soit en calculant une différence de moyennes entre deux groupes du PDP (dans ce cas noter "D").

	b1c1	b2c1	b1c2	b2c2
s1a1	3	2	8	3
s2a1	2	4	4	2
s3a1	0	1	3	4
s4a2	0	2	1	1
s5a2	6	7	4	3
s6a2	2	2	3	1

	A	B/c1	A/c1	A.B	B/c1a1	B.C
	4	-1	2.5	3	-1	4
	3	2	3	0	2	4
	2	1	0.5	-1	1	0
	1	2	1	-1		2
	5	1	6.5	0		2
	2	0	2	1		2
D/M	D	M	D	D	M	M

3. Calculer les moyennes (m) et écarts-types corrigés (s) des deux premières colonnes du tableau de données, arrondis à trois décimales :

	b1c1	b2c1
m	2.167	3.000
s	2.229	2.191

TESTS: Analyse en Composantes Principales (10 points)

Une batterie de 11 tests a été soumise à 42 élèves de classes de Cinquième:
Dépendance-Indépendance à l'égard du champ (DIC), Facteur G (G), Facteur Spatial (S), Test numérique (N), Test verbal (V), Réflexivité-impulsivité (RIT), Réactivité à la difficulté de la tâche (MRD), Métaconnaissances sur les méthodes de travail (MEN), Rapidité au test Spatial (SI), Réflexion-Impulsivité- précision (RIEI), Précision au test Spatial (SEI)
Source: P. Rozencwajg (1994) - Stratégies de résolution de problèmes scolaires et différences individuelles. Thèse de doctorat. Université Paris V.

Afin de comprendre la structure des performances aux 11 tests, on procède à une analyse en composantes principales (ACP) normée.

1. A quoi est égale la somme des valeurs propres ? On attend deux réponses.

$\Sigma \lambda =$ *Variance du nuage = 11*

$\Sigma \lambda =$ *Nombre de variables = 11.*

On s'intéresse aux deux premiers axes factoriels qui rendent compte de 64% de la variance totale.
2. Sur la figure 1, tracer les 11 vecteurs représentant les variables.

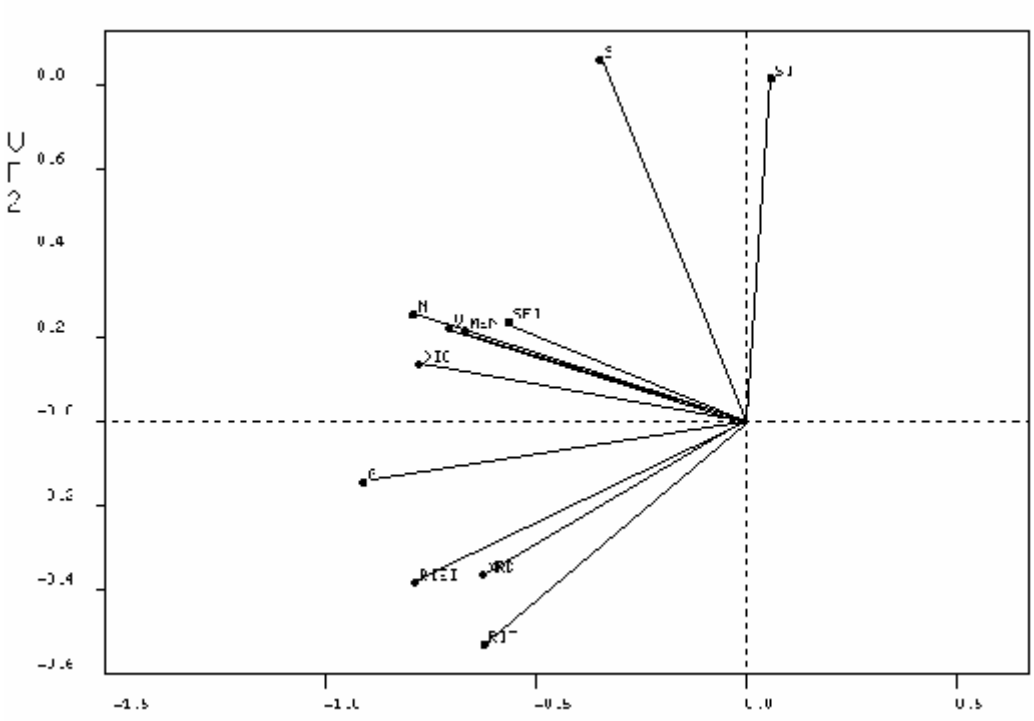

Figure 1: Nuage des variables dans le plan défini par les axes factoriels 1 (axe horizontal, orienté à droite) et 2 (axe vertical, orienté vers le haut).

3. Au vu du graphique ainsi complété, que peut-on dire des corrélations:

- entre les variables N et V:
  corrélation (fortement) positive, proche de +1
- entre les variables S et RIEI:
  corrélation proche de 0
- entre les variables SI et RIT:
  corrélation négative

Tableau 1: Caractéristiques des variables dans l'espace factoriel. Extrait de sorties du logiciel Addad (les valeurs sont exprimées en millièmes).

	J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR
1	DIC	624	1	91	-778	606	125	133	18	8
2	G	844	1	91	-907	823	170	-144	21	9
3	S	861	1	91	-348	121	25	860	740	334
4	N	688	1	91	-790	624	129	253	64	29
5	V	549	1	91	-707	499	103	223	50	22
6	RIT	666	1	91	-623	388	80	-527	278	125
7	MRD	527	1	91	-627	393	81	-367	134	61
8	MEN	495	1	91	-670	449	93	215	46	21
9	SI	665	1	91	58	3	1	814	662	299
10	RIEI	766	1	91	-786	618	128	-385	148	67
11	SEI	376	1	91	-566	320	66	237	56	25
			1000			1000			1000	

4. L'examen du tableau 1 montre pour la variable SEI, un indice QLT égal à 376. On considère que cette valeur est faible. Que signifie-t-elle?

Le point SEI est mal représenté sur le plan factoriel défini par les deux axes retenus.

5. Quelle est la variable la plus fortement corrélée avec l'axe 1? Justifier votre réponse.

Facteur G (G). C'est le vecteur-variable qui forme l'angle le plus faible avec cet axe.

6. Quelles sont les deux variables les plus fortement corrélées avec l'axe 2? Justifier votre réponse.

Facteur Spatial (S) et rapidité au test spatial (SI). Ce sont les deux variables qui forment l'angle le plus faible avec l'axe 2 (vertical).

7. A partir de ce constat, comment peut-on interpréter cet axe 2?

Il s'agit d'un axe Spatial, "saturé en facteur Spatial" (contrairement à l'axe 1, axe de taille, corrélé avec tous les tests). D'un côté de l'axe (en haut), les performances élevées en spatial. De l'autre côté (en bas), les performances plus faibles en Spatial.

COURS: Structuration des données (4 points)

Dans une expérimentation, la structure du protocole est décrite à l'aide de 6 facteurs F, R, K, D, E, J.

1. Sachant que F4<R6> et R2<K3>, écrire la relation (en précisant les indices) liant les facteurs F et K:

F8<K3>

2. Compte tenu des conventions habituelles de la notation indicée, indiquer le nombre de modalités du facteur F:

24 modalités

3. Sachant les relations suivantes: D∗E et E∗J, que peut-on dire de la relation entre ces 3 facteurs?

Rien: en particulier, sans information complémentaire, on ne peut <u>pas</u> conclure que les 3 facteurs sont croisés dans leur ensemble.

4. Dans quel cas parle-t-on de "niveaux" pour les modalités d'un facteur?

Lorsque ces modalités sont ordonnées (exemple: groupes d'âges différents)

Dossier EUROS (10 points)

On s'intéresse aux comportements induits par l'utilisation des "Euros" chez les personnes âgées. On a donc testé deux fois, dans la même journée, 80 personnes âgées (prises au hasard dans différentes maisons de retraite de Paris). L'expérience de simulation est la suivante : on donne un budget à ces personnes puis on leur propose d'acquérir des articles alimentaires et des articles dits de confort. Le budget ainsi que les achats à effectuer sont exprimés en "Francs" lors du premier test, en "Euros" lors du second test.

Le tableau suivant regroupe le nombre d'achats simulés pour chaque test et pour chaque type d'article :

		2nd Test (Euros)	
		Alimentaire	*Confort*
1er Test (Francs)	*Alimentaire*	35	20
	Confort	17	8

Question initiale *: on se demande si le fait d'utiliser une monnaie inhabituelle peut avoir une influence sur le comportement d'achat des personnes âgées, pour ce qui concerne les articles de confort.*

On jugera un effet faible s'il est inférieur ou égal à 5 points de pourcentage, important s'il est supérieur ou égal à 10 points de pourcentage.

1. Calculer l'effet "Monnaie" sur la fréquence d'achat des articles de confort dans cet échantillon :
   d = 28/80 – 25/80 = 0.375 (3.75 points de pourcentage)

2. Élaborer une conclusion descriptive :
   Chez ces 80 personnes âgées, la fréquence du comportement d'achat de produits de confort est plus élevée avec des Euros qu'avec des Francs. Cependant, cet effet est faible (d=3.75 points de pourcentage < 5 points).

3. On désire généraliser les résultats à la population des personnes âgées vivant à Paris.
a/ Indiquer précisément le nom du test à utiliser. Justifier l'utilisation de ce test :
   Test du Khi² de McNémar.

b/ Développer le calcul :

$$\chi^2 = \frac{(n_{+-} - n_{-+})^2}{(n_{+-} + n_{-+})} = \frac{(17-20)^2}{17+20} = \frac{9}{37} = 0.24$$

c) Sachant que l'on trouve p = .62, indiquer le résultat du test :
   Le test est non significatif.

d) Elaborer une conclusion inférentielle qui réponde à la question initiale :
   Chez ces 80 personnes âgées, le fait d'utiliser une monnaie inhabituelle semble avoir une influence sur le comportement d'achat, pour ce qui concerne les articles de confort. Mais cette influence apparaît faible. De plus on ne peut pas conclure à l'existence d'une telle influence au-delà de cet échantillon, chez l'ensemble des personnes âgées (Test Khi² de McNémar = 0.24, p=.62, NS).

DOCIMOLOGIE: Test de typicalité (6 points)

Lors de la correction du baccalauréat, la moyenne générale (μ) des notes de l'épreuve de philosophie est égale à 9.45, l'écart-type (σ) est égal à 2.32. La distribution de ces notes est approximativement normale.
Chaque correcteur a un paquet de 64 copies à corriger.

1. Calculer la moyenne, moy(M) et l'écart-type, ety(M) de la distribution d'échantillonnage de la moyenne pour des échantillons de taille n=64.

moy(M) = μ *9.45*

ety(M) = $\sigma/\sqrt{n}$ *2.32 / 8 = 0.29*

2. Un des correcteurs a obtenu des notes relativement faibles, avec une moyenne (m) de 8.67. Il se demande si son paquet de copies peut être qualifié d'atypique par rapport à l'ensemble des copies. Pour cela il situe ses copies dans la distribution d'échantillonnage définie précédemment. Il trouve P(M < m) = .0036. Indiquer comment il a trouvé cette valeur (cf. ci-dessous un extrait d'une table de la loi normale).

$$P(M < m) = P(Z < \frac{m - moy(M)}{ety(M)} = P(Z < \frac{(8.67 - 9.45)}{0.29} = P(Z < -2.69) = 1 - P(Z < 2.69) = 1 - 9964 = .0036$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

3. Quelle va être sa conclusion sur l'atypicalité éventuelle de ses copies ? Justifier.

Ce paquet de copies peut être qualifié d'atypique au seuil .0036. En tirant un paquet de 64 copies au hasard, il y a 0.36% de chances d'obtenir un paquet avec une moyenne si basse.

4. Sachant en fait que les copies ont été mélangées et distribuées au hasard aux différents correcteurs, on peut de demander si le système de correction utilisé par ce correcteur n'est pas différent de celui des autres correcteurs. Pour cela on teste l'hypothèse nulle selon laquelle son système de correction est le même. Utiliser le résultat précédent pour répondre à cette nouvelle question :

On peut rejeter l'hypothèse nulle et conclure à un système de correction différent de celui des autres correcteurs (test Z significatif au seuil .0036 unilatéral).

Théorie: Densité dans un tableau de contingence (6 points)

Pour l'analyse des tableaux de contingence on définit un nouvel indice: la densité. On désigne par d^{jk} la densité d'une case (j,k) du tableau. Cette densité est ainsi définie:

$$d^{jk} = \frac{n_{jk}}{\widetilde{n}_{jk}} = \frac{f_{jk}}{\widetilde{f}_{jk}}$$

($\widetilde{n}_{jk}$ et $\widetilde{f}_{jk}$ désignent les effectifs théoriques et les fréquences théoriques)

1/ Dans quel cas la densité d^{jk} est-elle égale à 1?

d^{jk} *= 1 si* $n_{jk} = \widetilde{n}_{jk}$*, c'est-à-dire si l'effectif observé est égal à l'effectif théorique, c'est-à-dire dans le cas de l'indépendance.*

2/ Quelle est la valeur du taux de liaison t^{jk} correspondant à une densité d^{jk} égale à 1 ?

$d^{jk} = 1 \Leftrightarrow t^{jk} = 0$

3/ A quel intervalle de valeurs des taux de liaison t^{jk} correspond:

- une densité d^{jk} inférieure à 1?

$n_{jk} < \widetilde{n}_{jk}$ *(Observés* n_{jk} *< Théoriques)* $\Rightarrow$ *Taux de liaison négatif*

- une densité d^{jk} supérieure à 1?

$n_{jk} > \widetilde{n}_{jk}$ *(Observés > Théoriques)* $\Rightarrow$ *Taux de liaison positif*

4/ Existe-t-il des valeurs minimale et maximale pour cet indice d^{jk} ? Si oui lesquelles?

Min = 0

Max = + ∞ *(pas de limite maximale)*

Les logiciels Addad, DS3Win, EyeLID et Statistica ont été utilisés pour analyser ces données.