Université René Descartes Paris 5 - Institut de Psychologie - Licence – 2001/2002
U.E. RTD - Responsable: Denis Corroyer - Première session d'examen: 7/06/2002 -

Durée de l'épreuve : 3 heures.
Épreuve sans document. La calculette est autorisée (sans sa documentation).
Les dossiers BUDGET2 et EYSENCK sont présentés sur les trois dernières pages. Ces trois pages peuvent être détachées et conservées.
Les exercices (encadrés) sont indépendants. Le barème (sur 32) 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)

Les logiciels Addad, DS3Win, EyeLID-2, PAC et Statistica ont été utilisés pour analyser ces données.

Description et Inférence (1.5 points)

1/ Quelles sont les différences principales entre l'étape descriptive et l'étape inférentielle, lors de l'analyse statistique des données ?

L'étape descriptive consiste en l'analyse des données observées. Les indices statistiques (statistiques descriptives) ou les représentations graphiques, utilisés lors de cette étape, visent à décrire ces données. L'étape inférentielle consiste, lorsque ces données constituent un échantillon d'une population plus vaste (population parente) non observée, à se demander si certains résultats observés sur l'échantillon peuvent être généralisés à cette population parente. On essaie alors de préciser les paramètres (moyenne, corrélation…) de cette population. Pour cela on utilise des statistiques spécifiques (statistiques inférentielles, telles que Khi² et t de Student).

2/ Certains tests (Khi² par exemple) peuvent s'interpréter dans le cadre inférentiel ensembliste ou dans le cadre inférentiel fréquentiste. Citer les principales différences (préalables et interprétation) entre ces deux cadres de justification et d'interprétation (CJI) de l'inférence :

Le cadre inférentiel fréquentiste relève de ce que l'on appelle l'inférence généralisante. On l'utilise lorsque l'on cherche alors à généraliser, à partir des résultats observés sur un échantillon, à la population dont est issu cet échantillon. Ce cadre inférentiel suppose, strictement, que l'échantillon ait été obtenu par un tirage au hasard de l'échantillon.

Le cadre inférentiel ensembliste ne constitue pas, à strictement parler, un cadre inférentiel, dans la mesure où il ne s'agit plus de généraliser à une population plus vaste. Il s'agit dans ce cas de situer un groupe d'observations particulier parmi un ensemble de possibles. L'expression "inférence" fait référence, dans ce cas, aux procédures utilisées (tests de signification) qui sont de type inférentiel. La visée est, quant à elle, une visée plutôt descriptive (caractériser le groupe d'observations).

3/ Soit un tableau de contingence sur lequel on calcule les statistiques Phi² et Khi² ainsi que le seuil associé au test du Khi². Si l'on avait un effectif deux fois plus grand dans toutes les cases du tableau, que deviendraient...

a/ le Phi² ? (diminué/inchangé/augmenté ?)
Inchangé, comme toute statistique strictement descriptive.

b/ le Khi² ? (diminué/inchangé/augmenté ?)
Augmenté (très précisément doublé).

c/ le seuil observé du test du Khi²? (diminué/inchangé/augmenté ?)
Diminué (le seuil observé et le Khi² varient en sens inverse)

Typicalité (2 points)

Un groupe de 50 élèves a un QI moyen de 112, avec un écart-type de 16 et une distribution de forme proche d'une distribution normale. Un petit groupe de 3 élèves est très fier de son QI moyen, supérieur à celui du groupe, soit 120. On se demande si ce trio peut se targuer d'être supérieur au groupe, du point de vue de son QI moyen. Pour cela on va procéder à un test de typicalité, consistant à situer ce groupe dans la distribution d'échantillonnage de la moyenne des QI.

1/ Enoncer le théorème des 3 moyennes :
La moyenne de la distribution d'échantillonnage de la moyenne est égale à la moyenne parente

2/ En déduire la moyenne de la distribution d'échantillonnage de la moyenne :
Moy (M) = Mu = 112

3/ Calculer la variance puis l'écart-type de la distribution d'échantillonnage de la moyenne, en utilisant la formule suivante :

$$Var(M)=\frac{\sigma_2}{n}\frac{(N-n)}{(N-1)}=\frac{256}{3}\cdot\frac{(50-3)}{(50-1)}=85.33\cdot\frac{47}{49}=81.85$$

d'où Ety(M) = 9.05

4/ Procéder à un test de typicalité (test Z) en calculant la valeur de la statistique Z.
$z = (120 - 112) / \sqrt{81.85} = 0.88$

5/ On trouve p = .19. Que conclure par rapport à l'atypicalité autoproclamée de ce trio?
Le QI moyen du trio est supérieur au QI moyen du groupe plus vaste auquel il appartient. Néanmoins, le seuil observé (p = .19) n'étant manifestement pas petit, le trio ne peut pas être déclaré atypique, par rapport à ce groupe de 50 élèves.

Données expérimentales (12 points)

Ces analyses portent sur le dossier Eysenck, présenté sur la dernière page.

A. Structures (2 points)

1/ Indiquer le nom du(des) T-facteurs :
le facteur T
2/ Indiquer le nom du(des) G-facteurs :
le facteur A
3/ Exprimer à l'aide des symboles (∗ ou <>) la relation entre les couples de facteurs suivants :
S et T :
$S_{20}*T_2$
A et S :
$S_{10}<A_2>$
T et A :
$T_2 * A_2$

4/ Indiquer le plan le plus riche :
$S_{10}<A_2>*T_2$

5/ Indiquer un autre plan du protocole :
$S_{20}*T_2$

B. Calculs (3 points)

Donner tous les résultats de cette partie arrondis à deux décimales.

1/ Pour chacun des deux groupes (a1 et a2) du PDP1, calculer, et reporter dans le tableau suivant, la moyenne, l'écart-type, l'écart-type corrigé, la variance et la variance corrigée :

	moy	ety	s	var	s²
a1	19.00	3.49	3.68	12.20	13.56
a2	25.80	3.16	3.33	9.96	11.07

2/ Calculer la variance totale de ce PDP :
$v_{totale} = 22.64$

3/ Calculer la variance inter :
$v_{inter} = [(19.0-22.4)^2 + (25.8-22.4)^2] / 2 = 11.56$

4/ Calculer la variance intra de deux manières différentes (détailler les calculs) :
$v_{intra} = (12.20 + 9.96) / 2 = 11.08$

$v_{intra} = Vtotale – Vinter = 22.64 – 11.56 = 11.08$

5/ Calculer le rapport de corrélation Eta² :
$Eta^2 = Vinter / Vtotale = 11.56 / 22.64 = 0.51 (51\ \%)$

6/ Calculer s^2_{inter} sachant que :
$$s^2_{\text{inter}} = \frac{G}{G-1} \times v_{\text{inter}} = \frac{2}{1} \times 11.56 = 23.12$$
7/ Calculer s^2_{intra} sachant que :
$$s^2_{\text{intra}} = \frac{n}{n-G} v_{\text{intra}} = \frac{20}{20-2} \times 11.08 = 12.31$$
8/ Calculer l'effet calibré ECG :
$ECG = sinter / sintra = \sqrt{23.12} / \sqrt{12.31} = 4.81 / 3.51 = 1.37$

C. Effet de l'âge avec la tâche t1 (3 points)

On fait l'hypothèse que l'effet de l'âge est négligeable avec la tâche t1.

1/ Donner une écriture formelle de cet effet :
Effet partiel ou effet intra A/t1

2/ Parmi les PDP proposés, indiquer quel est le PDP pour cet effet (PDP1…PDP5) :
PDP2

3/ On trouve m^{a1}=7.00 et m^{a2}=6.50, et un effet calibré, noté ECG, égal à 0.22. Compte tenu de ces résultats, rédiger une première conclusion (descriptive) sur cet effet :
Parmi ces 20 sujets effectuant l'apprentissage involontaire d'une liste de mots, les sujets âgés rappellent en moyenne moins de mots que les sujets jeunes. Cependant la différence de performance est faible (|d| = 0.5 < 2 et ECG=0.22 < 0.33).

4/ En mettant en œuvre un test t de Student on trouve t = 0.68, ddl=18.
a/ A partir de l'extrait suivant de la table du T de Student, en déduire le résultat du test :

α α/2 ddl	.05 .025	.01 .005	.001 .0005
18	2.101	2.878	3.922

Le test n'est pas significatif au seuil .05 bilatéral (t_{obs} = 0.68 < $t_{[18].05}$ = 2.101)

b/ Que peut-on en conclure ?
Même si on observe une différence entre les deux groupes de 10 sujets, jeunes et âgés, on ne peut pas conclure à l'existence d'une différence entre les deux populations parentes respectives pour cette tâche particulière (t=0.68, ddl=18, NS)

5/ Le logiciel PAC donne, pour cet effet, le résultat bayésien suivant :

negligeable ? Pr(|effet vrai| < 2.00) = 0.9705

a/ Traduire ce résultat en langage naturel :
Il existe une probabilité de 97% que la différence vraie entre les sujets jeunes et les sujets âgés soit, en valeur absolue, inférieure à 2 mots.

b/ Que peut-on en conclure du point de vue de la question posée au départ ?
Pour l'apprentissage involontaire d'une liste de mots, la différence vraie entre les performances de rappel des sujets jeunes et des sujets âgés est négligeable (inférieure à 2 mots) avec une bonne garantie (g = .97).

D. Effet de l'âge avec la tâche t2 (2 points)

On fait l'hypothèse que les sujets jeunes réussissent mieux la tâche t2 que les sujets âgés.

1/ Parmi les PDP proposés, quel est le PDP pour cet effet (PDP1…PDP5) ?
PDP5

2/ Calculer les moyennes des deux groupes d'âge et l'effet observé d
m^{a1} = 12.00 m^{a2} = 19.30 d = 7.30

3/ Sachant également que ECG= 1.59, rédiger une conclusion descriptive :
En moyenne, les 10 sujets âgés réussissent mieux la tache d'apprentissage intentionnel que les 10 sujets les plus jeunes. La différence de performance (d = 7.30) apparaît importante (d > 5 mots et ECG=1.59 > 2/3)

4/ Prolonger cette conclusion par une conclusion inférentielle sachant que le logiciel PAC donne le résultat bayésien suivant :

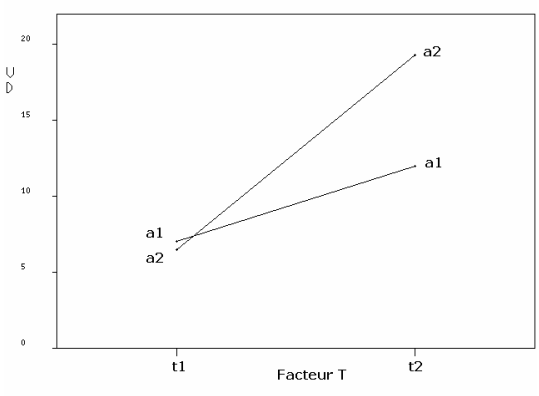
notable ? Pr(effet vrai > 5.00000000) = 0.9345

Pour l'apprentissage intentionnel d'une liste de mots, il existe une probabilité de 93% que la différence vraie de performance, entre les sujets jeunes et les sujets âgés, soit importante (au moins de 5 mots).

E. Comparaison des effets de l'âge avec t1 et avec t2 (1 point)

1/ Donner une expression formelle de la question posée :
Effet d'interaction A.T

2/ Répondre à la question posée au vu du graphe ci-après :
Chez ces deux groupes de 20 sujets (dont 10 sujets jeunes et 10 sujets âgés) l'effet de l'âge sur les performances de mémorisation d'une liste de mots est différent selon la nature de la tâche. Alors que la différence est en faveur des jeunes, mais faible, après un apprentissage involontaire, la différence est en faveur des sujets âgés, et importante, après un apprentissage intentionnel.



F. Protocoles Dérivés Pertinents (PDP) (1 point)

Parmi les PDP proposés, quels sont les PDP (PDP1…PDP5) pour les effets suivants :
1/ Effet de A :
PDP4 (ou PDP1)
2/ Effet de T :
PDP3
3/ Effet de A.T :
PDP3

Sondage (2.5 points)

Dans un pays lointain une élection se prépare. Les électeurs ont le choix entre deux candidats A et B.
Les sondages ayant un certain crédit dans l'opinion publique, un institut de sondage publie le résultat suivant qui met en relation le journal le plus souvent lu par les électeurs et le candidat préféré. Ce sondage a concerné "un échantillon représentatif de 1300 personnes".

Tableau 1 : Effectifs observés

	A	B	
La Montagne	80	550	
Le Peuple	70	600	

Tableau 2 : Pourcentages en ligne

	A	B	Total
La Montagne	12.70%	87.30%	100%
Le Peuple	10.45%	89.55%	100%
Moyenne	11.54%	88.46%	100%

1/ Comment est appelé, habituellement, un tableau tel que le tableau 1, qui rapporte une distribution d'effectifs selon deux variables qualitatives ?
Le terme le plus courant pour désigner ce type de tableau est tableau de contingence

2/ Donner 3 autres synonymes (français ou anglais) :
Contingency table, Tri croisé, Tableau de correspondances, Tableau de dépendance, Cross tabulated data,

3/ Un commentateur commente les résultats du sondage et, au vu du tableau exprimé en pourcentages, annonce : "Dans ce sondage, 12.70% des gens favorables au candidat A lisent *La Montagne*". Qu'en pensez-vous ?
C'est faux. Ce sont en fait 53.33% (80/150) des gens favorables au candidat A qui lisent La Montagne. Ce pourcentage (12.70%) doit être lu ainsi : 12.70% des lecteurs de La Montagne sont favorables au candidat A.

4/ S'il n'y avait strictement aucune différence d'opinion entre les lecteurs de ces deux journaux, combien aurait-on observé de gens, à la fois lecteurs du journal *La Montagne* et préférant le candidat A ? Donner le nom usuel de la statistique calculée, sa formule de calcul, le détail de ces calculs et le résultat arrondi à 2 décimales :

Il s'agit de l'effectif théorique (ou effectif attendu ou expected frequency)

$T = (n_j \times n_k) / n = (630 \times 150) / 1300 = 72.69$

5/ A partir du tableau 1, calculer avec une précision de 4 décimales la fréquence des lecteurs du journal *Le Peuple* parmi ceux qui préfèrent le candidat B

$f = 600 / 1150 = 0.5217 (52.17\%)$

6/ Ce pays a pour devise "pourquoi faire simple quand on peut faire compliqué". Aussi, à l'entrée de ce pays, par ailleurs charmant, on exige de savoir résoudre le problème suivant :
Utiliser la formule suivante du théorème de Bayes, ainsi que les résultats du tableau 2 pour retrouver la valeur calculée ci-dessus (fréquence des lecteurs de *Le Peuple* parmi ceux qui préfèrent le candidat B).

$$P(j/k)=\frac{P(j)\cdot P(k/j)}{\sum_J P(j)\cdot P(k/j)}=$$

Appelons M le journal La Montagne *et P le journal* Le Peuple. *On a alors :*

$$P(P/B)=\frac{P(P)\cdot P(B/P)}{P(P)\cdot P(B/P)+P(M)\cdot P(A/M)}=\frac{\frac{670}{1300}\times\frac{600}{670}}{\frac{670}{1300}\times\frac{600}{670}+\frac{630}{1300}\times\frac{550}{630}}=\frac{\frac{600}{1300}}{\frac{600}{1300}+\frac{550}{1300}}=\frac{600}{600+550}=\frac{600}{1150}=0.5217$$

Analyses multivariées (10 points)

Ces analyses portent sur le dossier BUDGET2 présenté sur les dernières pages de ce document.

A. Analyse factorielle (1 point)

On met en œuvre une ACP normée sur ces données.

1/ Donner l'intitulé en clair de l'acronyme ACP ?

Analyse en Composantes Principales

2/ Que signifie "normée" dans l'expression "ACP normée" ?

Cela signifie que l'ACP a été appliquée au tableau des variables centrées–réduites.

3/ Pourquoi aurait-on pu également, sur ce dossier, procéder à une ACP simple ?

Parce que les variables sont toutes sur la même échelle (des temps d'occupation en minutes)

B. Valeurs propres (1.5 points)

```
-------------------------------------------------------------------------------------------------------------
 !NUM ! VAL PROPRE ! POURC.! CUMUL !VARIAT.!*! HISTOGRAMME DES VALEURS PROPRES
-------------------------------------------------------------------------------------------------------------
 !  1 !    3.87072 ! 24.192! 24.192!*******!*!***************!***************!***************!***************
 !  2 !    3.69041 ! 23.065! 47.257!  1.127!*!***************!***************!***************!************
 !  3 !    2.00030 ! 12.502! 59.759! 10.563!*!***************!***************!*
 !  4 !    1.52453 !  9.528! 69.287!  2.974!*!***************!*********
 !  5 !    1.09027 !  6.814! 76.101!  2.714!*!***************!**
 !  6 !     .83352 !  5.209! 81.311!  1.605!*!*************
 !  7 !     .77278 !  4.830! 86.141!   .380!*!*************
 !  8 !     .59963 !  3.748! 89.888!  1.082!*!*********
 !  9 !     .43013 !  2.688! 92.577!  1.059!*!*******
 ! 10 !     .37628 !  2.352! 94.928!   .337!*!******
 ! 11 !     .24748 !  1.547! 96.475!   .805!*!****
 ! 12 !     .22533 !  1.408! 97.883!   .138!*!***
 ! 13 !     .15763 !   .985! 98.869!   .423!*!**
 ! 14 !     .11706 !   .732! 99.600!   .254!*!**
 ! 15 !     .04105 !   .257! 99.857!   .475!*!*
 ! 16 !     .02291 !   .143!100.000!   .113!*!
```

1/ A quelle valeur est égale la somme de la colonne "val. propre" ?

La somme est égale à 16

2/ Expliquer à quoi correspond cette valeur (on attend deux réponses) :

Cette valeur (16) correspond :

- à la variance totale du nuage

- au nombre de variables du tableau car chaque variable étant centrée-réduite, sa variance est égale à 1.

3/ Expliquer en quoi consiste, de manière générale (quel que soit le type d'analyse factorielle), ce qui est appelé habituellement le critère de Kaiser :

Il consiste à retenir, pour résumer les données, uniquement les premières variables factorielles pour lesquelles le pourcentage de variance (par rapport à la variance totale du nuage) est supérieur au pourcentage moyen (100% divisé par le nombre de variables).

4/ Indiquer la règle pratique dans le cas d'une ACP, à laquelle conduit le critère de Kaiser :

Dans le cas d'une ACP, le critère de Kaiser revient à retenir les variables factorielles dont la valeur propre est supérieure à 1.

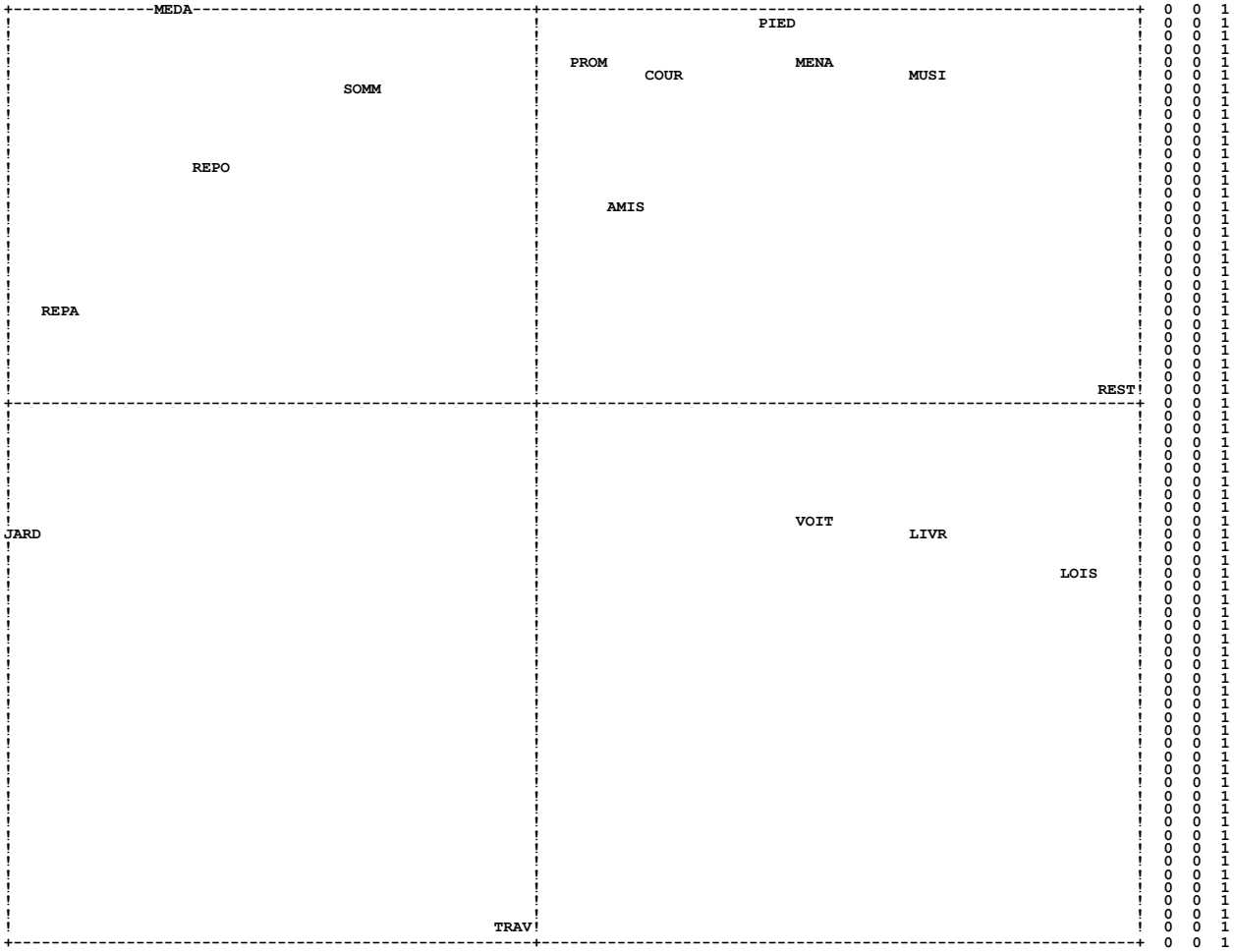
5/ Indiquer en justifiant, combien de variables factorielles seraient suffisantes pour résumer les données :

L'application du critère de Kaiser ($\lambda > 1$) conduit à retenir 5 variables factorielles. Elles totalisent 76% de la variance (ce que l'on pourrait éventuellement considérer comme insuffisant).

(l'examen du diagramme en étoiles montre un coude après la 7ème variable. En retenant 7 variables (sur 16) on aurait alors 86% de la variance totale).

C. Graphe factoriel des variables (3 points)

Sur le graphe ci-dessous, l'axe horizontal (orienté vers la droite) est l'axe factoriel n°1, l'axe vertical (orienté vers le haut) est l'axe factoriel n°2.



1/ Au vu de ce graphe donner une estimation des corrélations entre les variables :

- MENA et MUSI :

  Proche de +1

- TRAV et REST :

  Proche de 0

- REPA et REST :

  Proche de -1

2/ A l'examen de ce graphe factoriel, indiquer :
- une variable fortement corrélée, et positivement, avec l'axe 1 :
  REST
- une variable fortement corrélée, et négativement, avec l'axe 1 :
  REPA (ou JARD)
- une variable fortement corrélée, et positivement, avec l'axe 2 :
  PROM (ou COUR, SOMM)
- une variable fortement corrélée, et négativement, avec l'axe 2 :
  TRAV

3/ Toujours au vu du graphe, de quel ordre est la corrélation entre TRAV et l'axe 1 ?
Proche de 0 (ou faiblement négatiive)

4/ A partir du tableau des aides à l'interprétation fourni par le logiciel Addad (cf. dernières pages de ce document), construire les tableaux des principales contributions, pour l'axe 1 et pour l'axe 2 :
CTR > 1000/16 = 62

Axe 1	
-	+
JARD (148) REPA (116) MEDA (66)	REST (182) LOIS (141) LIVR (76) MUSI (72)

Axe 2	
-	+
TRAV (211)	MEDA (123) PIED (113) PROM (90) MENA (88) MUSI (83) COUR (81) SOMM (73)

5/ Donner une interprétation de la première variable factorielle :
La première variable factorielle oppose deux types de temps libre :
- à gauche des temps libres plutôt centrés sur la maison (Jardinage et Repas) associés à la fréquentation de Médias.
- à droite des temps libres plutôt tournés vers l'extérieur (Restaurant et Loisirs extérieurs) et un autre type de loisirs (Livres et Musique)

6/ Donner une interprétation de la deuxième variable factorielle :
La deuxième variable factorielle oppose le travail à un ensemble d'autres activités non professionnelles (médias, promenades, musique, courses, sommeil…)

D. Graphe factoriel des individus (3 points)

1/ Que suggère la position des individus suivants sur le graphe factoriel (cf. page suivante), concernant leurs caractéristiques ?
- i02, en haut à gauche du graphe :
  Cet individu i02 est éloigné du centre du gravité du nuage et éloigné des autres individus du groupe. Cela suggère qu'il a un profil (un emploi du temps) atypique.
- i15, près du centre du graphe :
  Cet individu est proche du centre de gravité du nuage. Cela suggère qu'il a un profil (un emploi du temps) proche du profil moyen du groupe.

2/ D'après le tableau des aides à l'interprétation fourni par le logiciel Addad (cf. dernières pages de ce document), la colonne INR (notation Addad) correspond à la contribution relative des individus à la variance totale du nuage.
a/ Indiquer les formules, de cette contribution relative et de la contribution absolue associée.
$Cta_i = f_i \times d^2(i,G)$
$Ctr_i = Cta_i / Var$
b/ Que peut-on dire de l'individu i02, sachant qu'il présente une valeur élevée pour INR ?
Tous les individus ayant le même poids (36/1000), cela signifie qu'il est éloigné du centre de gravité du nuage, et donc qu'il a un profil atypique.
3/ Que représentent les colonnes notées 1#F et 2#F par Addad ?
Ce sont les deux premières variables factorielles (ensemble des coordonnées factorielles des individus sur les axes factoriels 1 et 2).
4/ L'individu i10 a été effacé sur le graphe factoriel des individus. Compte tenu des informations du tableau des aides à l'interprétation (cf. dernières pages), reporter cet individu à la main sur le graphe

5/ Dans le tableau des aides à l'interprétation, on dispose pour chaque axe de valeurs présentées dans une colonne intitulée COR par Addad.
a/ Donner une autre notation pour cette statistique :
$Ctr\ (VF/i)$ ou Ctr^i_{VF}
b/ Commenter les faibles valeurs de COR pour i15, i16 et i17 sur l'axe 1 :
Ces faibles valeurs de COR indiquent que ces individus sont mal représentés sur cet axe factoriel.

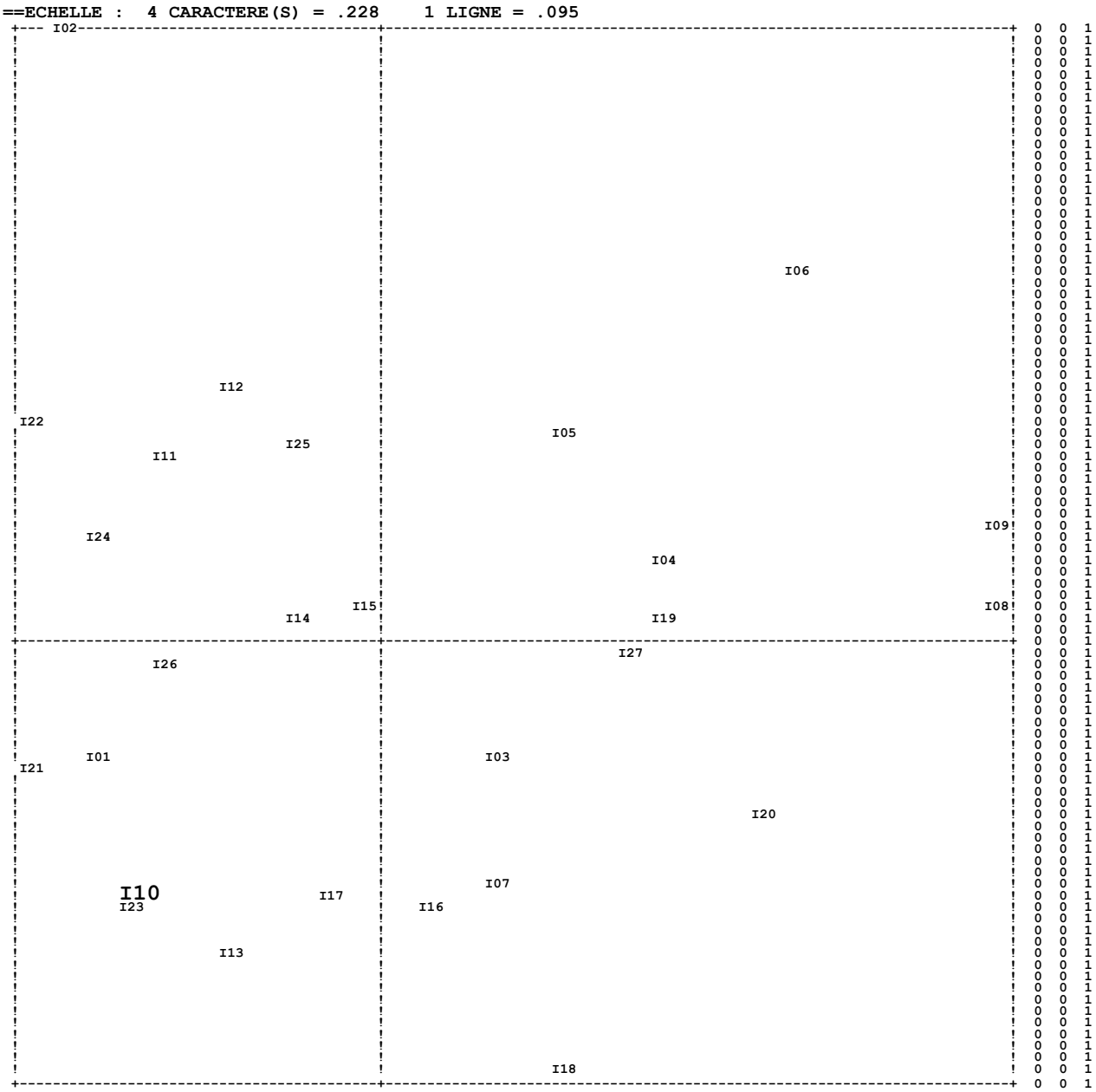
6/ Ce tableau comporte également, pour chaque axe, une colonne intitulée CTR par Addad.
a/ Donner une autre notation pour cette statistique :

Ctr (i/VF) ou Ctr_i^{VF}

b/ Donner la formule de calcul de la CTR, et de la contribution absolue associée, pour un individu sur l'axe 1 :

$Cta_i = f_i \,.\, (VF1_i)^2$

$Ctr_i = Cta_i / \lambda 1$



E. Mise en correspondance des deux nuages (1 point)

En mettant en correspondance les deux graphes (graphe des variables et graphe des individus), que suggère la position des individus suivants sur le graphe, concernant leurs caractéristiques ?

- i08 (à droite du graphe) :

Cela suggère qu'il passe, relativement au groupe :

- beaucoup de temps au restaurant, dans les loisirs extérieurs, à lire des livres et à écouter de la musique.

- peu de temps à jardiner, dans les repas à la maison, et à fréquenter les médias

- i18 (en bas du graphe) :

Cela suggère qu'il passe, relativement au groupe :

- beaucoup de temps à travailler,

- peu de temps dans beaucoup d'autres activités : fréquentation des médias, déplacements à pied, promenades, ménage, écouter de la musique, faire les courses, dormir.

F. Retour aux données (0.5 point)

Les individus i21 (à gauche du graphe), et i08 (à droite du graphe) sont éloignés sur le graphe factoriel des individus.

1/ Qu'est-ce que cela suggère ?

Cela suggère qu'ils ont des profils très différents

2/ Indiquer en quoi l'examen du tableau de données confirme (ou infirme) cette interprétation ?

Ils ont des occupations différentes : i08 se repose moins que i21 (31mn/51mn). Il passe moins de temps en repas à la maison (82mn/99mn) mais plus de temps au restaurant (26mn/6mn). i08 travaille beaucoup moins (270mn/356mn), passe plus de temps à faire le ménage (52mn/21mn), peu de temps en jardinage (36mn/82mn), plus de temps à écouter de la musique (6mn/0mn), à lire (8mn/1mn), à se déplacer en voiture (81mn/41mn). Seul le temps de fréquentation des médias apparaît strictement identique (108mn).

Remarque : Pour comparer plus efficacement leurs deux profils, relativement au reste du groupe, il faudrait disposer des temps centrés-réduits.

Régression multiple (4 points)

Ces analyses portent sur le dossier BUDGET2, présenté sur les dernières pages de ce document.

On se demande si le temps passé avec les amis, dépend de différentes variables : le temps passé au restaurant (REST), au travail (TRAV), à faire le ménage (MENA), à jardiner (JARD).

Pour cela on procède à une régression multiple, en prenant la variable amis comme "variable dépendante", et les 4 autres comme "variables indépendantes".

Les résultats de la régression multiple sont présentés dans les dernières pages de ce document sous forme d'une copie d'écran lors de l'exécution du logiciel Statistica.

1/ Donner au moins deux autres synonymes, dans le contexte de la régression, pour chacune de ces deux expressions :

- variable dépendante :

Variable expliquée, variable critère, variable à prédire, variable à estimer

- variable indépendante :

Variable explicative, variable prédictrice.

2/ A partir des résultats affichés par le logiciel Statistica, écrire les deux équations de régression :

- avec les variables brutes :

$AMIS = 0.39\ REST - 0.16\ TRAV - 0.21\ MENA + 0.05\ JARD + 69.34$

- avec les variables centrées-réduites :

$AMIS = 0.23\ REST - 0.54\ TRAV - 0.15\ MENA + 0.11\ JARD$

3/ Commenter la valeur -.15519 reportée dans la colonne B :

Cette valeur signifie que lorsque le temps de travail augmente de 1 minute, et les autres variables étant maintenues constantes, le temps passé avec les amis diminue, en moyenne, de 0.16 minutes.

4/ Indiquer le nom de l'indice descriptif qui évalue la qualité de la prédiction de la variable AMIS par les quatre autres variables prises dans leur ensemble :

Il s'agit du coefficient de détermination R^2

5/ Commenter la valeur de cet indice :

$R^2 = .25$ (25%). Cela signifie que ces 4 variables permettent de prédire 25% de la variance des temps passés avec les amis, ce qui est important ($R^2 = .25 > .16$).

6/ Quelle statistique nous permet d'évaluer si, dans la population parente, ces quatre variables, prises globalement, prédisent le temps passé avec les amis (AMIS) ?

Il s'agit du test F.

7/ Commenter le résultat obtenu :

Le test F n'est pas significatif ($p = .16$). Cela signifie que l'on ne peut pas rejeter l'hypothèse selon laquelle ces 4 variables ne permettent pas, dans la population parente, de prédire la variance de la variable AMIS.

8/ Quelle est la meilleure variable prédictrice du temps passé avec les amis? La réponse à cette question est-elle unique? Justifier et discuter :

La réponse dépend du critère considéré :

- Si on prend les coefficients β (régression multiple sur les variables réduites) c'est la variable TRAV qui apparaît la meilleure prédictrice (coefficient β le plus élevé), quand les autres variables sont maintenues constantes, du temps passé avec les amis (plus on passe de temps au Travail moins on en passe entre Amis).

- Si on prend les coefficients B (régression multiple sur les variables brutes, ce qui est possible ici car les variables sont sur la même échelle) on conclut que c'est le restaurant (REST) qui est, quand les autres variables sont maintenues constantes, le meilleur indicateur du temps passé avec les amis (plus on passe de temps au Restaurant plus on en passe entre Amis) car cette variable a le coefficient B le plus élevé.

DOSSIER BUDGET2

Source : d'après Lebart, L., Morineau, A., Piron, M. (1995) – Statistique exploratoire multidimensionnelle, Paris : Dunod, p.58.

Le CESP (Centre d'Etude des Supports de Publicité) a relevé, dans son enquête Budget-Temps Multimédia de 1991/1992 auprès de 17 665 personnes, des temps d'activité quotidienne. Nous nous intéressons ici à une sous-population de 27 "individus" (en fait 27 sous-groupes) composée uniquement d'hommes actifs.
Les activités répertoriées sont les suivantes :
Sommeil (SOMM), Repos (REPO), Repas chez soi (REPA), Repas au restaurant (REST), Travail rémunéré (TRAV), Ménage (MENA), Visite à amis (AMI), Jardinage et bricolage (JARD), Loisirs extérieurs (LOIS), Ecoute de musique (MUSI), Lecture de livres (LIVR), Courses et démarches (COUR), Promenades (PROM), Déplacements à pied (PIED), Déplacements en voiture (VOIT), et Fréquentation de médias (MEDA)

Le tableau ci-dessous rapporte les temps (en minutes) passés en moyenne, chaque jour, dans chacune de ces activités par les 27 individus.

	SOMM	REPO	REPA	REST	TRAV	MENA	AMIS	JARD	LOIS	MUSI	LIVR	COUR	PROM	PIED	VOIT	MEDA
I01	464	24	107	5	300	21	51	82	10	1	0	41	7	7	52	136
I02	516	58	103	10	209	42	30	33	2	5	1	34	8	25	29	226
I03	463	34	85	17	298	18	38	56	18	6	3	31	6	9	56	136
I04	456	43	74	22	239	26	51	60	18	4	5	52	10	11	73	142
I05	478	44	77	15	212	22	42	44	18	2	6	48	15	15	73	168
I06	465	42	85	24	226	37	42	16	11	9	9	44	14	20	59	145
I07	458	47	95	15	314	25	39	42	17	1	17	34	5	6	61	103
I08	457	31	82	26	270	52	38	36	26	6	8	43	10	12	81	108
I09	465	40	79	31	269	36	22	4	19	6	15	47	11	22	48	82
I10	449	42	86	8	312	15	16	113	15	0	2	32	8	8	60	154
I11	450	63	87	10	250	40	56	83	3	2	0	45	9	10	62	145
I12	455	47	96	9	251	30	13	57	8	3	7	52	15	16	49	195
I13	462	39	90	8	323	15	22	82	15	1	5	26	4	7	60	131
I14	454	45	97	19	269	23	40	93	3	3	12	42	12	11	62	129
I15	433	50	92	13	284	22	21	63	13	6	7	38	12	12	48	169
I16	438	33	102	11	338	28	6	65	14	1	20	35	7	14	53	130
I17	458	44	88	7	313	24	23	64	9	1	12	30	7	7	70	108
I18	455	47	79	32	381	24	7	40	13	0	10	23	1	9	59	100
I19	467	37	87	22	264	41	28	33	12	2	11	45	7	11	73	135
I20	433	36	76	17	355	34	13	32	13	3	13	37	8	22	57	96
I21	473	51	99	6	356	21	28	82	9	0	1	36	13	7	41	108
I22	462	60	104	9	240	35	14	83	1	2	7	46	6	17	53	184
I23	453	46	86	8	359	13	18	54	4	0	5	34	3	10	49	143
I24	485	53	86	0	222	25	23	92	8	0	4	53	7	10	75	166
I25	457	43	95	12	265	30	24	61	9	2	12	50	18	13	46	185
I26	444	54	91	7	302	32	16	98	5	2	4	39	14	11	62	127
I27	438	51	81	11	307	19	24	10	14	0	18	68	8	19	63	143

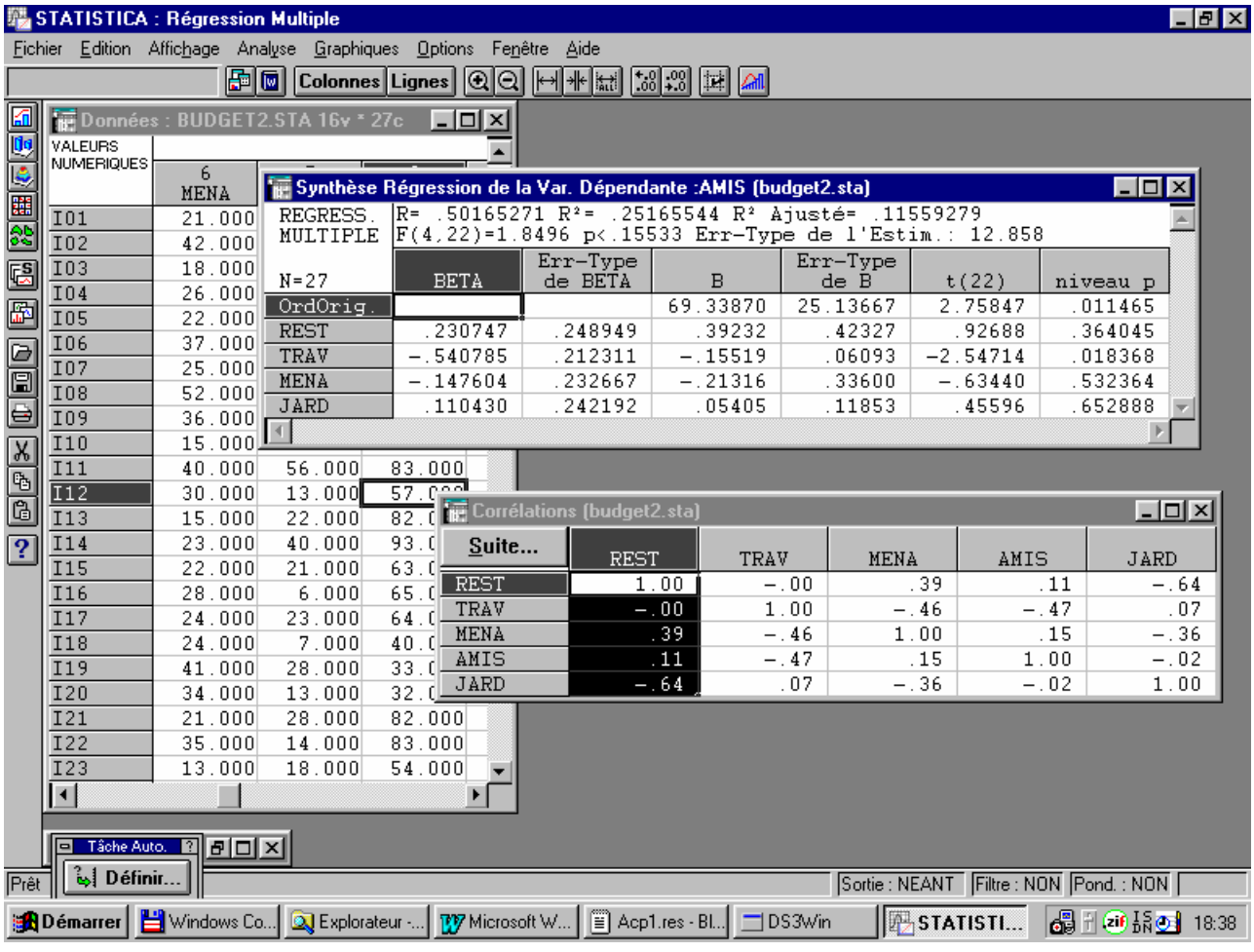
Aides à l'interprétation pour l'ACP du nuage des variables (logiciel Addad) :

! J1 !	QLT	POID	INR!	1#F	COR	CTR!	2#F	COR	CTR!	3#F	COR	CTR!	4#F	COR	CTR!	5#F	COR	CTR!
1!SOMM!	723	1	62!	-228	52	13!	521	271	73!	-155	24	12!	506	256	168!	-345	119	109!
2!REPO!	714	1	62!	-466	217	56!	387	150	41!	166	28	14!	-266	71	47!	-498	248	228!
3!REPA!	724	1	62!	-669	447	116!	147	22	6!	242	59	29!	212	45	29!	389	152	139!
4!REST!	785	1	62!	839	704	182!	15	0	0!	80	6	3!	256	65	43!	-98	10	9!
5!TRAV!	906	1	62!	-35	1	0!	-883	780	211!	339	115	57!	75	6	4!	63	4	4!
6!MENA!	516	1	62!	391	153	40!	570	325	88!	84	7	4!	145	21	14!	-98	10	9!
7!AMIS!	700	1	62!	133	18	5!	331	110	30!	-727	529	264!	193	37	24!	81	7	6!
8!JARD!	812	1	62!	-758	574	148!	-234	55	15!	-370	137	69!	-102	10	7!	188	35	32!
9!LOIS!	734	1	62!	738	545	141!	-288	83	22!	-275	76	38!	77	6	4!	155	24	22!
10!MUSI!	800	1	62!	528	279	72!	552	305	83!	-5	0	0!	371	138	90!	278	77	71!
11!LIVR!	732	1	62!	542	293	76!	-219	48	13!	492	242	121!	-381	145	95!	62	4	4!
12!COUR!	835	1	62!	196	38	10!	547	299	81!	-122	15	8!	-695	483	317!	23	1	0!
13!PROM!	768	1	62!	80	6	2!	576	331	90!	-94	9	4!	-312	97	64!	569	324	297!
14!PIED!	864	1	62!	345	119	31!	644	415	113!	567	321	161!	-44	2	1!	-81	7	6!
15!VOIT!	846	1	62!	408	167	43!	-211	44	12!	-656	430	215!	-368	135	89!	-264	70	64!
16!MEDA!	718	1	62!	-506	256	66!	672	452	123!	50	2	1!	-82	7	4!	10	0	0!
!	!	1000!			1000!			1000!			1000!			1000!			1000!	

Aides à l'interprétation pour l'ACP du nuage des individus (logiciel Addad)) :

! I1 !	QLT	POID	INR!	1#F	COR	CTR!	2#F	COR	CTR!	3#F	COR	CTR!	4#F	COR	CTR!	5#F	COR	CTR!
1! I01!	680	37	45!	-1955	195	37!	-892	41	8!	-1730	153	55!	1457	109	52!	1886	182	121!
2! I02!	978	37	111!	-2319	112	51!	5114	546	263!	2089	91	81!	3031	192	223!	-1337	37	61!
3! I03!	732	37	23!	761	57	6!	-924	84	9!	-1192	140	26!	2051	415	102!	595	35	12!
4! I04!	886	37	31!	1929	280	36!	738	41	5!	-2652	529	130!	-597	27	9!	-342	9	4!
5! I05!	662	37	33!	1123	88	12!	1766	217	31!	-2083	301	80!	-819	47	16!	-367	9	5!
6! I06!	909	37	45!	2704	373	70!	3038	471	93!	31	0	0!	921	43	21!	642	21	14!
7! I07!	413	37	25!	642	39	4!	-1941	352	38!	-143	2	0!	323	10	3!	-343	11	4!
8! I08!	819	37	62!	4207	656	169!	352	5	1!	-1862	129	64!	823	25	16!	353	5	4!
9! I09!	892	37	58!	4224	714	171!	1024	42	11!	1719	118	55!	666	18	11!	6	0	0!
10! I10!	699	37	29!	-1855	272	33!	-2104	350	44!	-908	65	15!	-215	4	1!	334	9	4!
11! I11!	538	37	40!	-1481	127	21!	1524	135	23!	-1814	191	61!	-481	13	6!	-1106	71	42!
12! I12!	878	37	25!	-1064	105	11!	2118	417	45!	1114	115	23!	-1177	129	34!	1099	112	41!
13! I13!	882	37	25!	-1249	143	15!	-2552	598	65!	-627	36	7!	1057	103	27!	-160	2	1!
14! I14!	228	37	18!	-640	54	4!	197	5	0!	-519	36	5!	-453	27	5!	896	106	27!
15! I15!	360	37	19!	-320	13	1!	356	16	1!	714	63	9!	-188	4	1!	1461	264	73!
16! I16!	763	37	37!	146	1	0!	-2113	281	45!	2324	340	100!	-470	14	5!	1417	126	68!
17! I17!	642	37	17!	-469	30	2!	-2035	559	42!	-321	14	2!	-163	4	1!	-521	37	9!
18! I18!	888	37	58!	1231	61	14!	-3527	499	125!	1442	83	38!	1260	64	39!	-2123	181	153!
19! I19!	516	37	18!	1737	383	29!	223	6	0!	-504	32	5!	72	1	0!	-860	94	25!
20! I20!	754	37	39!	2499	370	60!	-1415	119	20!	2085	258	81!	-371	8	3!	23	0	0!
21! I21!	584	37	36!	-2618	439	66!	-1050	71	11!	54	0	0!	763	37	14!	763	37	20!
22! I22!	802	37	38!	-2532	392	61!	1868	213	35!	1421	124	37!	-548	18	7!	-940	54	30!
23! I23!	775	37	30!	-1832	256	32!	-2168	359	47!	1044	83	20!	237	4	1!	-968	72	32!
24! I24!	815	37	40!	-2138	267	44!	917	49	8!	-1787	187	59!	-1448	123	51!	-1801	189	110!
25! I25!	798	37	25!	-645	39	4!	1681	266	28!	794	59	12!	-1186	132	34!	1789	301	109!
26! I26!	427	37	22!	-1625	277	25!	-158	3	0!	-16	0	0!	-1092	125	29!	465	23	7!
27! I27!	765	37	51!	1543	108	23!	-38	0	0!	1328	80	33!	-3452	542	289!	-860	34	25!
! !		1000!			1000!			1000!			1000!			1000!			1000!	

Régression multiple (copie d'écran du logiciel Statistica) :



Dossier EYSENCK

Le modèle de mémorisation de Craik et Lockart (1972) stipule que le niveau de rappel d'un matériel verbal est fonction du niveau de traitement lors de sa présentation initiale.

On rapporte ici les données (adaptées et partielles) d'une étude de Eysenck (1974) consacrée à la rétention de matériel verbal en fonction du niveau de traitement. Il s'agissait de comparer la mémorisation d'une liste de mots dans deux conditions :

- d'une part après un apprentissage involontaire. Lors de cette tâche t1, le sujet devait lire une liste de mots et calculer le nombre de lettres de chacun de ces mots. Il ne savait pas qu'il devrait, plus tard, se rappeler la liste de mots,
- d'autre part un apprentissage intentionnel. Lors de cette tâche t2, le sujet devait lire la liste et essayer de mémoriser les mots.

Les sujets passaient les deux tâches dans l'ordre t1 puis t2, avec deux listes distinctes, mais de difficultés équivalentes (d'après un test sur un autre ensemble de sujets).

Par ailleurs, afin d'étudier l'effet du vieillissement sur la mémorisation, deux groupes de sujets ont été comparés :

- un groupe (a1) de 10 sujets dont l'âge se situait en 18 et 30 ans,
- un groupe (a2) de 10 sujets dont l'âge se situait entre 55 et 65 ans.

Les sujets de cette expérience ont été tirés au hasard parmi un groupe de volontaires, étudiants et enseignants de Sciences Humaines.

Source : Howell, D.C. (1998) – Méthodes statistiques en Sciences Humaines, Paris : DeBoeck Université, p.454.

On considère qu'un effet inférieur ou égal à 2 est faible ou négligeable, et qu'un effet supérieur ou égal à 5 est important ou notable.

Les calculs seront faits dans le sens "modalité 2 – modalité 1"

Les données et plusieurs protocoles dérivés sont reportés ci-dessous :

	A	t1	t2		PDP1	PDP2	PDP3	PDP4	PDP5
s1	1	9	10		19	9	1	9.50	10
s2	1	8	19		27	8	11	13.50	19
s3	1	6	14		20	6	8	10.00	14
s4	1	8	5		13	8	-3	6.50	5
s5	1	10	10		20	10	0	10.00	10
s6	1	4	11		15	4	7	7.50	11
s7	1	6	14		20	6	8	10.00	14
s8	1	5	15		20	5	10	10.00	15
s9	1	7	11		18	7	4	9.00	11
s10	1	7	11		18	7	4	9.00	11
s11	2	8	21		29	8	13	14.50	21
s12	2	6	19		25	6	13	12.50	19
s13	2	4	17		21	4	13	10.50	17
s14	2	6	15		21	6	9	10.50	15
s15	2	7	22		29	7	15	14.50	22
s16	2	6	16		22	6	10	11.00	16
s17	2	5	22		27	5	17	13.50	22
s18	2	7	22		29	7	15	14.50	22
s19	2	9	18		27	9	9	13.50	18
s20	2	7	21		28	7	14	14.00	21