CORRIGÉ ASDP4 (UE ADN)

Durée de l'épreuve : 1h30
Épreuve sans document. La calculette est autorisée (sans sa documentation).
Les exercices (encadrés) sont indépendants. Le barème (sur 40) est donné à titre indicatif. La note finale sera donnée sur 20.
Indiquer les réponses exclusivement sur ce document et aux endroits réservés à cet effet (ne pas écrire dans la marge).

Régression multiple : Dossier TMT (11 points)

Le test des points à relier (Trail Making Test : *TMT*) est un test servant à observer la capacité attentionnelle du sujet dans un test simple mais qui demande l'inhibition d'une routine de conduite.

> On demande au sujet de relier par un trait 13 chiffres et 13 lettres (présentés éparpillés sur une feuille) dans l'ordre de la série des chiffres et dans l'ordre alphabétique ; ce qui nécessite d'intercaler une lettre avec un chiffre (ex : 1-A-2-B, etc.). Si le sujet commet une erreur, l'expérimentateur le signale immédiatement et le sujet la corrige.

On note le temps de réalisation de l'épreuve (en secondes) en tenant compte du temps de correction. Dix sujets ont participé à cette épreuve. Ils ont également été sollicités pour la passation du test de la Tour de Londres : *TOURL* (noté sur 12) et pour une épreuve de résolution de problèmes : *PROB* (notée sur 20) pour laquelle on dispose également du temps : *TPPROB* (en secondes) mis pour résoudre les problèmes.

***Question initiale** : Quel est le meilleur prédicteur du temps réalisé à l'épreuve TMT ?*

Le logiciel *Statistica* donne les sorties suivantes :

Tableau 1 : Matrice des corrélations

REGRESS. MULTIPLE	TMT	TOURL	PROB	TPPROB
TMT	1.00	−.76	−.33	.22
TOURL	−.76	1.00	.83	−.60
PROB	−.33	.83	1.00	−.65
TPPROB	.22	−.60	−.65	1.00

1) Commenter les corrélations entre les variables prédictrices et la variable à prédire :

Les variables Tour de Londres (TRL) et Résolution de problèmes (PROB) sont corrélées négativement avec la performance au test TMT (respectivement r = -.76 et r = -.33) ; alors que la variable Temps de résolution (TPPROB) est corrélée positivement avec le test TMT (r = .22) ***1***

Tableau 2 : Synthèse de la régression multiple

REGRESS. MULTIPLE — R= .93386925 R²= .87211178 R² Ajusté= .80816767 F(3,6)=13.639 p<.00435 Err-Type de l'Estim.: 7.0321

N=10	BETA	Err-Type de BETA	B	Err-Type de B	t(6)	niveau p
OrdOrig.			131.4602	19.28111	6.81809	.000488
TOURL	−1.56094	.261228	−10.3864	1.73819	−5.97541	.000986
PROB	.85057	.274780	3.3895	1.09497	3.09547	.021239
TPPROB	−.16159	.193141	−.0526	.06281	−.83665	.434840

2) Indiquer ce que représentent les coefficients *R* et R^2 et les interpréter :

Le coefficient R représente le coefficient de corrélation multiple, et le coefficient R^2 le coefficient de détermination ***1***

R = .93 : la corrélation multiple est élevée (R > .40). R^2 = .87 indique que 87% de la variance de TMT est prise en compte par l'ensemble des prédictrices (ce qui est considéré comme important $R^2 > .16$) ***2***

3) Indiquer ci-après l'équation de la régression multiple :

$$\widetilde{TMT} = -10.38_{TOURL} + 3.39_{PROB} - .05_{TPPROB} + 131.46$$ ***1***

4) Le sujet 1 a obtenu 11.00 au test "Tour de Londres" (*TOURL*), 17.00 pour la résolution de problèmes (*PROB*) et a exécuté les problèmes en 105.00 secondes (*TPPROB*).

a) Quelle est la valeur prévue par la régression multiple pour le test *TMT* ?

$\widetilde{TMT} = (-10.38 \times 11) + (3.39 \times 17) - (.05 \times 105) + 131.46 = 69.31$ *(ou 69.66 si calcul sur arrondis)*

1 (0.5 pour calcul + 0.5 pour résultat)

b) Quelle est la valeur du résidu (sachant que le sujet 1 a obtenu 69.00 à l'épreuve *TMT*) ? Indiquer le calcul qui a permis de trouver ce résultat.

$y - \tilde{y} = 69.31 - 69.00 = 0.31$ *(ou 0.66 si calcul sur arrondis)* ***1 (0.5 calcul + 0.5 résultat)***

5) Que représentent les coefficients "*Bêta*" ?

Ce sont les coefficients de régression (b_k) centrés réduits ***1***

6) Interpréter la valeur "*Bêta*" obtenue pour la variable "Tour de Londres" (TOURL), donner le résultat du test *T* de Student qui lui est associé et formuler une conclusion inférentielle.:

Beta (TOURL) = -1.56. La performance obtenue au test TMT diminue de 1.56 écart-type quand la performance au test Tour de Londres augmente d'un écart-type. ***1***

Le test T de student est significatif au seuil p = .000986 pour la variable TOURL. Dans la population parente, la variable TOURL prend une part de variance non nulle de TMT, compte tenu de ce prennent déjà en compte les autres variables. ***2***

7) En se fondant sur les résultats contenus dans le tableau 2 (page précédente), répondre à la question initiale (justifier la réponse) :

La valeur "Bêta" la plus élevée concerne le test Tour de Londres. Cette variable a donc le poids prédicteur le plus fort de la performance au test TMT. ***1***

ACP Standard : Dossier Psychométrie (15 points)

Source : Rouanet, H, & Le Roux, B. (1993). L'analyse des données multidimensionnelles. Paris : Dunod.

Les données "Psychométrie" présentées ci-après, sont inspirées d'une recherche sur la pédagogie des mathématiques. Pour 20 sujets, on a les notes aux cinq épreuves individuelles suivantes : Combinatoire (COMB), Probabilités (PROB), Logique (LOG), notées de 0 à 10, QI verbal (QI) (dont les notes s'échelonnent ici de 85 à 125), Mathématiques (MATH), notée de 0 à 20. On aimerait étudier la structure de ces variables ainsi que les profils des 20 individus. Pour cela, on a effectué une ACP standard. L'analyse effectuée avec le logiciel *Addad* conduit aux résultats suivants :

Tableau 1 : Valeurs propres

NUM	VAL PROPRE	POURC.	CUMUL	VARIAT.	HISTOGRAMME DES VALEURS PROPRES
1	2.35638	47.128	47.128	*******	
2	1.12229	22.446	69.574	24.682	**********
3	.73187	14.637	84.211	7.808	********
4	.53531	10.706	94.917	3.931	*******
5	.25414	5.083	100.000	5.623	***

Tableau 2 : Aides à l'interprétation pour le "nuage des variables"

	J1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	COMB	796	1	200	-857	735	312	204	42	37	141	20	27
2	PROB	739	1	200	-666	443	188	412	170	151	-356	127	173
3	LOG	917	1	200	-582	338	144	-496	246	219	-577	332	454
4	QI	880	1	200	-828	686	291	183	33	30	400	160	219
5	MATH	878	1	200	-393	154	66	-795	632	563	304	92	126
			1000			1000			1000			1000	

1) A l'aide des tableaux 1 et 2 ci-dessus, on a décidé de retenir 3 axes factoriels pour cette analyse. Pourquoi ? Justifier.

Si on considère strictement le critère de Kaiser (VP >1), on ne retient que les 2 premiers axes. Cependant ces 2VF ne prennent en compte que 69.57% de la variance. Le 3ème axe apporte 14.64% de la variance. Si on retient ces 3 axes, on totalise un pourcentage 84%.

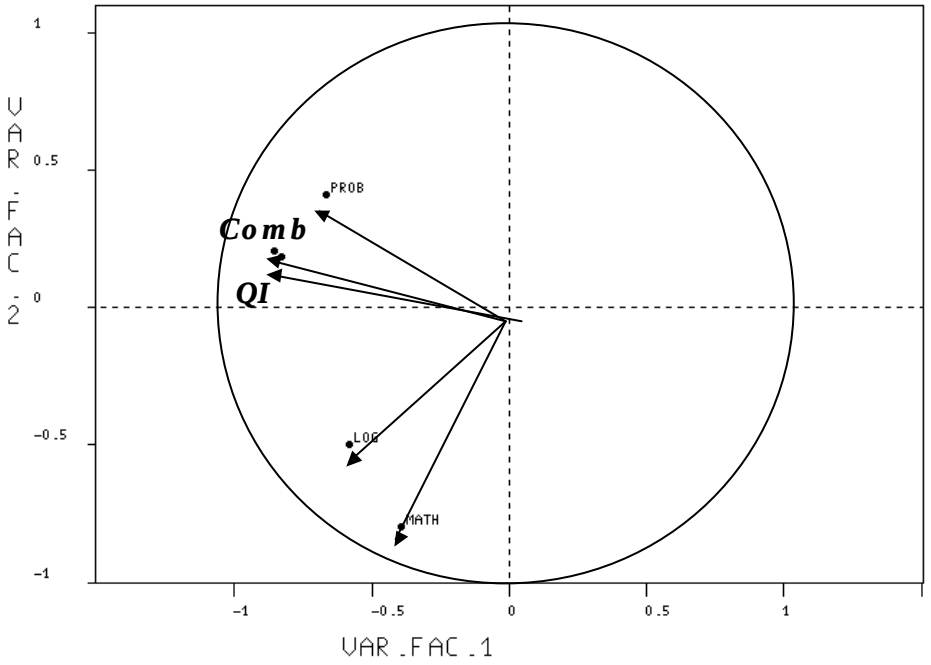
1.5

2) Quelle est la variable la moins bien représentée par l'ensemble de ces 3 axes ? (donner le nom et la valeur de l'indice correspondant) :

Il s'agit de la variable PROB, dont la QLT est égale à 739. ***1***

Ci-après le graphique factoriel pour le plan 1-2 (*logiciel EyeLID2*) :

Figure 1 : "Nuage des variables" ; Plan 1-2



3) A l'aide du tableau 2, reporter les deux étiquettes manquantes sur le graphique ci-dessus.

1 + 1

4) On s'intéresse maintenant à la représentation géométrique des corrélations entre les variables pour ***le plan 1-2*** (indépendamment des axes factoriels). Compléter le graphique ci-dessus de manière à en permettre l'interprétation.

Vecteurs ***1***
Cercle de corrélation ***1***

5) "Les variables COMB et QI présentent une corrélation forte et positive". Justifier cette affirmation :

Les vecteurs-variables COMB et QI, proches du cercle de corrélation, forment un angle proche de 0° (un cosinus carré proche de 0° indique une corrélation forte et positive) ***1***

5) On désire maintenant interpréter les 2 premiers axes factoriels.

a) Indiquer les variables qui contribuent au premier axe factoriel :

On retient les CTR > 1000/5 = 200 ***1***

-	+
COMB (312) QI (251)	

1

b) Interpréter ce premier axe factoriel :

Ce premier axe est représentatif des performances obtenues aux tests de Combinatoire et de QI.

1

c) Indiquer les variables qui contribuent au second axe factoriel :

-	+
LOG (219) MATH (563)	

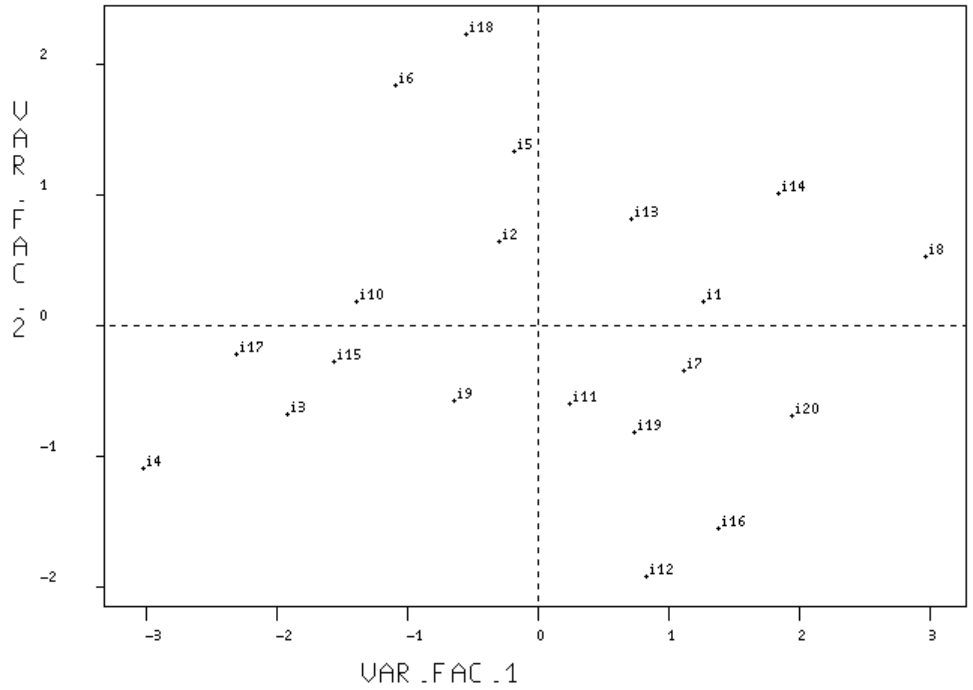
1

d) Interpréter ce second axe factoriel :

Cet axe est représentatif des performances plutôt mathématiques (tests de Logique et de Maths)
1

Ci-après (figure 2), le nuage des individus :

Figure 2 : Nuage des individus ; plan 1-2



6) Compte-tenu de l'analyse qui a été faite de la première et de la deuxième variable factorielle, indiquer ce que suggère l'ACP concernant les profils des individus i10 (à gauche), i8 (à droite) et i18 (en haut).

i10 a plutôt obtenu de bonnes performances aux tests de Combinatoire et de QI ***0.5***
i8 a plutôt obtenu de faibles performances aux tests de Combinatoire et de QI ***0.5***
i18 a plutôt obtenu de faibles performances aux tests de Logique et de Mathématiques ***0.5***

7) Pour analyser plus finement les profils de groupes d'individus, quelle analyse pourrait être mise en œuvre après cette ACP ? Indiquer le nom complet de cette analyse, ainsi que son abréviation usuelle.

Une Classification Ascendante Hiérarchique (CAH) ***0.5 + 0.5***

Analyse d'un tableau de contingence : Dossier Environnement (14 points)

Source : d'après Lebart, L. (1993). Analyse exploratoire de fichiers d'enquêtes ; apports de l'analyse des correspondances. In : *Développements récents dans l'analyse de grands ensembles de données*. Eurostat, Luxembourg.

Les données suivantes ont été recueillies pour étudier la relation entre la profession et la source d'information sur les problèmes d'environnement : 1283 personnes classées selon leur profession (7 modalités) ont été interrogées. Le tableau suivant donne pour chaque profession le nombre de personnes qui ont pour principale source d'information l'une des 6 sources suivantes : télévision, journaux, radio, livre, associations, et mairie. Les 7 catégories socio-professionnelles étudiées sont : agriculteur, cadre supérieur, cadre moyen, employé de bureau, ouvrier, retraité, chômeur. On se pose la question suivante : *existe-t-il une liaison entre les catégories socio-professionnelles et le choix de la source d'information ?*

Ci-après le tableau de données (tableau 1).

Tableau 1 : Catégories socio-professionnelles / Sources d'information

	Télé	*Journal*	*Radio*	*Livre*	*Association*	*Mairie*	*Total*
Agriculteur	26	18	9	5	4	6	68
Cadre supérieur	19	49	4	16	5	3	96
Cadre Moyen	44	87	4	39	14	3	191
Employé	83	87	13	24	5	1	213
Ouvrier	181	107	16	31	7	7	349
Retraité	167	95	29	15	7	7	320
Chômeur	27	9	4	2	2	2	46
Total	547	452	79	132	44	29	1283

1) Si l'on veut obtenir une représentation factorielle de ces données, quelle est la méthode appropriée ? Donner son nom complet et son abréviation usuelle.

Analyse (Factorielle) des Correspondances (AFC ou AC) ***0.5 + 0.5***

2) On procède à cette analyse avec le logiciel *Addad*. On obtient le tableau des valeurs propres suivant :

Tableau 2 : Valeurs propres

```
-------------------------------------------------------------------------------
 !NUM ! VAL PROPRE ! POURC.! CUMUL !VARIAT.!*! HISTOGRAMME DES VALEURS PROPRES
 -------------------------------------------------------------------------------
 !  1 !     .09151 ! 75.099! 75.099!*******!*!***************!***************!******
 !  2 !     .02193 ! 18.001! 93.100! 57.098!*!**************
 !  3 !     .00642 !  5.269! 98.369! 12.732!*!****
 !  4 !     .00175 !  1.437! 99.806!  3.832!*!*
 !  5 !     .00024 !   .194!100.000!  1.243!*!
```

3) On trouve $\phi^2 = 0.1218$. Retrouver cette valeur à l'aide du tableau 2 et indiquer la procédure utilisée :

La somme des valeurs propres est égale à f^2. ***1***

D'où $f^2 = .09151 + .02193 + ... + .00024 = 0.1218$) ***1***

4) Peut-on considérer que la liaison entre la catégorie socio-professionnelle et la source d'information est importante dans cet échantillon ? Calculer l'indice qui permet de répondre à cette question et conclure.

On doit calculer le V^2 de Cramer. ***0.5***

$$V^2 = \frac{f^2}{f^2_{Max}} = \frac{0.1218}{5} = 0.024$$ ***1 (0.5 calcul + 0.5 résultat)***

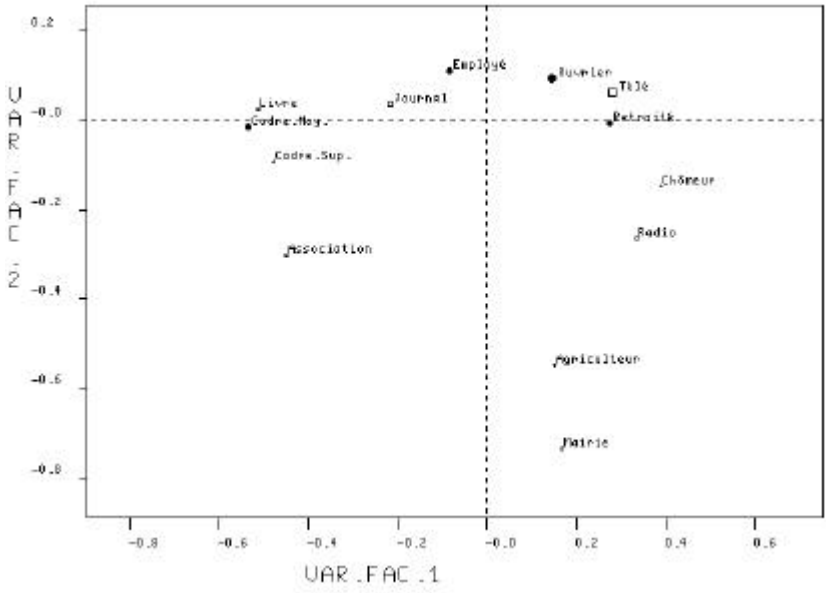
$V^2 = 0.024 < 0.04$: on conclut que la liaison entre la catégorie socio-professionnelle et la source d'information est faible. ***1***

5) On a décidé ici de retenir les 2 premiers axes factoriels. Pourquoi ? Justifier.

Un premier critère consiste à retenir les axes pour lesquels le pourcentage de valeur propre est supérieur à 100% divisé par nombre d'axes. Ici 100/5 = 20 (ou retenir la valeur propre supérieure à une VP moyenne : 0.1218/5 = 0.024). Selon ce critère, on ne retiendrait que le 1er axe. Ensuite, on s'assure que le pourcentage de variance pris en compte est suffisant (75% pour ce 1er axe). En prenant un second axe, on rajoute 18% de la variance, ce qui fait un cumul de 93% (ce qui est important). ***2***

La mise en correspondance des lignes et des colonnes du tableau donne la représentation géométrique suivante :

Figure 1 : Mise en correspondance des modalités lignes et des modalités colonnes



On donne également le tableau des taux de liaison :

Tableau 3 : Taux de liaison

	Télé	*Journal*	*Radio*	*Livre*	*Association*	*Mairie*
Agriculteur	-0.10	-0.25	1.15	-0.29	0.72	2.90
Cadre supérieur	-0.54	0.45	-0.32	0.62	0.52	0.38
Cadre Moyen	-0.46	0.29	-0.66	0.98	1.14	-0.31
Employé	-0.09	0.16	-0.01	0.10	-0.32	-0.79
Ouvrier	0.22	-0.13	-0.26	-0.14	-0.42	-0.11
Retraité	0.22	-0.16	0.47	-0.54	-0.36	-0.03
Chômeur	0.38	-0.44	0.41	-0.58	0.27	0.92

6) Illustrer à partir du point "Agriculteur" que ce graphe factoriel est bien une représentation géométrique des taux de liaison (cf. figure 1 et tableau 3 ci-dessus) :

Si l'on considère les point Agriculteur (en haut, à droite du graphique et 1ère ligne du tableau), on constate 1/ qu'il est proche de Mairie et de Radio avec lesquels il présente des taux de liaison élevés et positifs.2/ qu'il est éloigné des points Télé, Journal, Livre avec lesquels ils présente des taux de liaison négatifs. ***2***

7) Inférence. Indiquer ci-après :

a) La statistique de test appropriée, et une formule de cette statistique qui permette de calculer rapidement sa valeur (156.27) à partir des résultats précédents :

Il s'agit du test du Khi². ***1***

Khi² = n ´ Phi²

Khi² = 1283 ´ 0.1218 = 156.27 (valeur exacte : 156.33) ***1 (0.5 procédure + 0.5 résultat)***

b) Le nombre de degrés de liberté associé à ce test :

ddl = (J-1) (K-1) ; ddl = (6-1) (7-1) = 30 ***1 (0.5 procédure + 0.5 résultat)***

c) Ce test est significatif au seuil .001. Dans la mesure où l'on peut considérer que ces 1283 personnes ont été échantillonnées au hasard dans la population française, conclure sur l'existence d'une liaison entre les deux variables dans cette population parente :

Il existe une liaison entre la catégorie socio-professionnelle et la source principale d'information chez l'ensemble des français (test du Khi² significatif au seuil .001). ***1.5***