Durée de l'épreuve : 1 heure 30 mn.
Aucun document n'est autorisé. La calculette n'est pas autorisée.
Les différents exercices (encadrés) sont indépendants.
Le barème, donné à titre indicatif, est sur 60 ; la note finale sera donnée sur 20.
Indiquer les réponses exclusivement sur ce document. Ne rien écrire dans la marge gauche.

NB : Pour l'ensemble des dossiers traités ici, on prendra les valeurs repères suivantes (valeur supérieure d'un effet faible et valeur inférieure d'un effet important) pour les différents indices :
0.20 et 0.40 pour une corrélation
0.04 et 0.16 pour les indices R^2, V^2, Eta^2
0.33 (1/3) et 0.67 (2/3) pour les écarts calibrés.

Dossier Argent de poche (12 pts)

Lors d'une enquête (Chombart de Lauwe & al., 1963) on a posé la question suivante à 333 personnes :
"*Qui, dans la famille, doit donner l'argent de poche aux enfants ?*".
La *Réponse* pouvait être : "La femme" (F), "l'homme" (H) ou "les deux indifféremment" (HF).
Les personnes interrogées appartenaient à 3 *Catégories sociales* : ouvriers (Ouvr), catégories intermédiaires (Inter) ou milieu aisé (Aisé).
On trouve ci-dessous la distribution des effectifs observés selon ces deux variables, *Catégorie sociale* et *Réponse*.

		Réponse			
		F	H	HF	
Catégorie sociale	Ouvr	83	13	17	113
	Inter	43	14	51	
	Aisé	25	56	31	
		151			333

1. Donner trois expressions synonymes (françaises ou anglaises) pour désigner un tel tableau :
Tableau de contingence – Tri croisé – Tableau de correspondances – Tableau de dépendance – Table Banner – Distribution d'effectifs bivariée - Cross tabulated data – Contingency table.

2. Le tableau suivant donne ce qui est habituellement appelé les "effectifs théoriques".

	F	H	HF
Ouvr	51.2	28.2	33.6
Inter	49.0	26.9	32.1
Aisé	50.8	27.9	33.3

a. A quoi correspondent ces effectifs théoriques ?
Ils correspondent aux effectifs que l'on observerait s'il y avait stricte indépendance entre les deux variables (les distributions marginales étant fixées).

b. Indiquer comment calculer l'effectif théorique 51.2 en haut à gauche de ce tableau :
L'effectif théorique d'un case s'obtient en calculant le produit des totaux marginaux correspondants, divisé par l'effectif total, soit (nj × nk) / n
Par conséquent 51.2 = (113 × 151) / 333

3. Le tableau suivant donne les valeurs des Taux de liaison (ou Écarts relatifs).

	F	H	HF
Ouvr	+62 %	-54 %	-49 %
Inter	-12 %	-48 %	+59 %
Aisé	-51 %	+101 %	-7 %

a. À partir des seules attractions, commenter les différences entre les trois catégories sociales :
Par rapport à l'ensemble des personnes interrogées,
- les ouvriers pensent plus souvent que c'est la femme qui doit donner l'argent de poche.
- les personnes des catégories intermédiaires sont proportionnellement plus nombreux à penser que cela revient aux deux (hommes et femmes) indifféremment.
- les personnes des milieux aisés pensent plus fréquemment que cela revient à l'homme.

b. Que signifie la valeur +101 % ?
Cela signifie qu'il y a beaucoup (cf. signe + soit une attraction) de personnes de milieux aisés qui pensent que c'est à l'homme de donner l'argent de poche.
Précisément, on observe 101 % de personnes en plus par rapport à l'effectif attendu (s'il avait indépendance). On observe un effectif de 27.9 + (101 % × 27.9) = 56.
Cela correspond à un doublement par rapport à l'effectif attendu.

c. Que signifie la valeur -49 % ?
Cela signifie qu'il y a peu (cf. signe - soit une répulsion) d'ouvriers qui pensent que cela revient indifféremment aux deux (H et F) de donner l'argent de poche.
Précisément, on observe 49 % de personnes en moins par rapport à l'effectif attendu (s'il avait indépendance). On observe un effectif de 33.6 - (49 % × 33.6) = 17.
Cela correspond à une réduction de moitié par rapport à l'effectif attendu.

d. On trouve également *Phi²* = 0.2870. Que peut-on dire sur l'importance de la liaison globale entre les deux variables dans cet échantillon ?

Étant donné que Phi²max est supérieur à 1, précisément à min(J,K)-1 soit min(3,3)-1 = 2,on calcule le V² de Cramér = Phi² / Phi²max = 0.2870 / 2 = 0.14
V² est compris entre 0.04 et 0.16.
On en conclut que, dans cet échantillon, la liaison entre la catégorie sociale et la réponse à cette question est d'importance intermédiaire (ni faible ni importante).

Dossier CES – Liaison entre deux variables numériques (9 pts)

cf. présentation des données CES sur la dernière page.

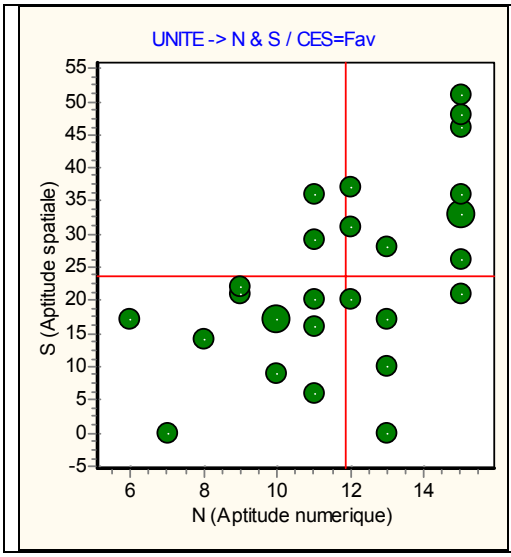
On s'intéresse à la liaison entre l'aptitude numérique (variable *N*) et l'aptitude spatiale (variable S) chez les seuls adolescents du collège favorisé (FAV).

1. Représentation graphique des données (4 pts)

Sur le graphique suivant on a représenté les "axes moyens".

a. A quoi correspond l'intersection de ces deux axes ?
Cela correspond au point moyen du nuage ou centre de gravité du nuage.

b. Ces axes moyens définissent 4 quadrants sur le graphe. Indiquer (à droite du graphe) en quoi la position des individus dans ces quatre quadrants suggère une covariance et une corrélation positive entre ces deux variables :



Ce qui suggère une covariance et une corrélation positive entre ces deux variables, c'est le fait que on trouve les individus principalement dans le quadrant en bas à gauche et dans celui en haut à droite.

Lorsque un individu est dans l'un de ces deux quadrants il contribue positivement à la covariance.
En bas à gauche, ses deux écarts à la moyenne (pour N et pour S) sont négatifs; le produit des deux écarts est donc positif.
En haut à droite, ses deux écarts à la moyenne sont positifs; le produit des écarts est donc également positif.

2. Description (2 pts)

L'équation de régression qui vise à prédire les scores en aptitude spatiale en fonction des scores en aptitude numérique est : $S = 2.94 \times N - 11.28$.

Expliquer en quoi cette équation indique également le signe de la covariance et de la corrélation :

Dans cette équation, de type Y = a X + b, le coefficient a (ici 2.94) est positif. Or ce coefficient est toujours de même signe que la covariance et la corrélation. Donc la corrélation est positive

3. Inférence (3 pts)

a. Décrire précisément en quoi consiste, pour les données analysées ici, la population parente :

La population parente est constituée de l'ensemble des élèves de 5ème de ce collège "favorisé" (de cette ville de région parisienne).

b. Le test de l'hypothèse nulle selon laquelle il n'existerait pas de la liaison entre ces deux variables dans la population parente (H0 : *RBPpar* = 0), indique $p = .0012$ (0.12 %)

Une erreur classique d'interprétation d'un test significatif (ou très significatif comme ici où le seuil p est très petit) consiste à conclure que la corrélation parente est importante.
Expliquer en quoi l'intervalle de confiance ci-dessous montre que on ne peut pas conclure, ici, à une corrélation importante dans la population :

IC (5 %) = [+.27 ; +.78]

Rappel : on considère qu'une corrélation est importante si elle est au moins égale (en valeur absolue) à .40.
On constate qu'un certain nombre de corrélations non importantes (.27, .30, .35, .39 par exemple), sont à l'intérieur de l'intervalle de confiance. Cela signifie que ces valeurs ne peuvent être exclues de l'ensemble des valeurs "possibles" de la corrélation parente. On ne peut donc pas conclure à une corrélation parente importante (> 40).

Dossier CES – Comparaison des collèges (8 points)

cf. présentation des données CES sur la dernière page.

On s'intéresse à la liaison entre les variables CES (collège) et N (aptitude numérique).

On considérera qu'une différence de scores d'aptitude numérique (N) est faible si elle est inférieure à 2 points et importante si elle est supérieure à 4 points.

Moyennes des groupes

CES	*Moy*
Def	7.3
Fav	11.9

On trouve :
d = 4.6
IC (5 %) = [2.8 ; 6.4]
Eta² = 40 %
t = 5.15, *ddl* = 40, *p* = 0.01 %
ECG = 1.12

1. Distinguer, parmi les statistiques *d, IC, Eta², t et ECG*, celles qui permettent de décrire l'échantillon et celles qui permettent de procéder à des inférences sur la population :

Pour décrire l'échantillon : *d, Eta², ECG*

Pour les inférences sur la population : *t, IC*

Rédiger une conclusion générale, concernant uniquement les différences d'aptitude numérique (*N*) entre les deux collèges, en distinguant nettement les aspects descriptifs, puis inférentiels :

Pour cet échantillon de 42 adolescents scolarisés en 5ème dans les deux collèges d'une ville de la région parisienne,
on constate que les scores en aptitude numérique sont meilleurs dans le collège "favorisé".
La différence est importante,
que l'on considère l'écart entre les moyennes (d = 4.6 > 4.0)
l'écart calibré (ECG = 1.12 > 0.67)
ou le rapport de corrélation (Eta² = 40 % > 16 %).

Il semble que,
pour l'ensemble des élèves de 5ème de ces deux collèges,
il existe bien une différence d'aptitude numérique, en faveur du collège "favorisé"
(t = 5.15, ddl = 40, p = 0.01 %).
Cependant on ne peut pas conclure que cette différence est importante
(IC (5 %) = [2.8 ; 6.4]).

Choix du test statistique (4 pts)

Le test statistique à mettre en œuvre pour tester l'hypothèse nulle classique d'une absence de liaison dans la population, dépend du type des variables analysées (nominale, numérique…) et du statut de ces variables (variable explicative, VI, VD…).
Attribuer chacun des 4 tests suivants à chacune des 4 situations décrites ci-dessous.

Khi² d'indépendance, t de Student, F de Fisher-Snedecor, Khi² de McNémar.

Liaison entre deux variables nominales dont l'une définit des groupes indépendants :

Khi² d'indépendance

Liaison entre une variable nominale à 3 modalités et une variable numérique :

F de Fisher-Snédécor

Deux mesures répétées sur une variable nominale binaire :

Khi² de McNémar

Liaison entre deux variables numériques :

T de Student

Décomposition de la variance (10 pts)

Soient les données suivantes où la variable *G* est une variable nominale définissant des groupes et où *X* désigne une variable numérique.

	G	X
s1	1	15
s2	1	16
s3	1	17
s4	2	18
s5	2	12
s6	3	15
s7	3	12
s8	3	10
s9	3	11

G	N	Moy	Var
g1	3	16	0.67
g2	2	15	9.00
g3	4	12	3.50

1. Indiquer comment calculer la moyenne générale des 9 notes, à partir des moyennes des 3 groupes. On trouve $m = 14$.

$$m = \sum\left(p_g \times m_g\right) = \left(\frac{3}{9}\times 16\right) + \left(\frac{2}{9}\times 15\right) + \left(\frac{4}{9}\times 12\right) = 14$$

2. La variance inter-groupes, notée *Vinter*, est égale à 3.33.

a. Indiquer brièvement, ce que mesure cette variance inter-groupes :

La variance inter mesure la dispersion des moyennes des groupes.

b. Rappeler sa formule puis développer avec les valeurs de cet exemple :

C'est la variance des moyennes des groupes :

$$Vinter = \sum\left(p_g \times \left(m_g - m\right)^2\right) = \left(\frac{3}{9}\times\left(16-14\right)^2\right) + \left(\frac{2}{9}\times\left(15-14\right)^2\right) + \left(\frac{4}{9}\times\left(12-14\right)^2\right) = 3.33$$

3. La variance intra-groupes, notée *Vintra*, est égale à 3.78.

a. Indiquer brièvement, ce que mesure cette variance intra-groupes :

La variance intra-groupes mesure la dispersion des scores à l'intérieur des groupes.

b. Rappeler sa formule puis développer avec les valeurs de cet exemple :

C'est la moyenne des variances des groupes :

$$Vintra = \sum\left(p_g \times v_g\right) = \left(\frac{3}{9}\times 0.67\right) + \left(\frac{2}{9}\times 9.00\right) + \left(\frac{4}{9}\times 3.50\right) = 3.78$$

4. Indiquer la formule du rapport de corrélation *Eta²* :

$$Eta^2 = \frac{Vinter}{Vtotale}$$

5. Calculer la variance des 9 notes (*Var*). Indiquer ci-dessous la procédure utilisée et la valeur obtenue :

Var = Vinter + Vintra = 3.33 + 3.78 = 7.11

Méthodologie (4 pts)

1. Quelle est la règle méthodologique importante à appliquer avant de commencer l'analyse de la relation entre deux variables, quel que soit le type de chacune de ces deux variables ?

Il faut tout d'abord analyser la distribution de chaque variable (analyse univariée), indépendamment de l'autre.

2. Il est essentiel, lorsque l'on analyse la liaison entre deux variables numériques, d'analyser le graphe de corrélation (et ne pas se contenter de calculer le coefficient de corrélation). Indiquer deux arguments en faveur de cette règle méthodologique :

L'étude du graphe de corrélation permet de détecter l'existence éventuelle :
- d'une liaison non linéaire (une liaison peut être forte mais non linéaire, ce que ne peut détecter le coefficient de corrélation linéaire de Bravais-Pearson),
- de sous-groupes hétérogènes,
- d'une (ou plusieurs) valeur(s) atypique(s).

Inférence (10 pts)

1. Utilisation des tables (4 pts)

Extrait de la table des valeurs critiques de la variable χ^2 (Khi²) :

p / *Ddl*	.05 (5%)	.01 (1%)	.001 (0.1%)
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52

Indiquer, pour chacune des valeurs suivantes de la statistique *Khi²* et des degrés de liberté correspondants :

a. le résultat du test, significatif (noté S) ou non significatif (noté NS), en entourant la bonne réponse.

b. une approximation du seuil *p*.

Khi² = 7.00	*ddl* = 3	S ou [NS]	*p > 5 %*
Khi² = 18.0	*ddl* = 1	[S] ou NS	*p < 0.1 %*
Khi² = 3.0	*ddl* = 2	S ou [NS]	*p > 5 %*
Khi² = 11.90	*ddl = 5*	[S] ou NS	1 % < *p* < 5 %

2. Hypothèse nulle (2 pts)

Indiquer précisément en quoi consiste l'hypothèse nulle (notée H0) lorsque l'on procède à un test du *Khi²* d'indépendance sur deux variables *J* et *K* :

Dans la population parente, il n'existe pas de liaison entre les deux variables J et K.
Dans ce cas, les effectifs observés correspondent aux effectifs "théoriques" (effectifs attendus si indépendance) et Phi²par = 0.

3. Seuil *p* (4 pts)

On étudie la relation entre deux variables *J* et *K*. Sur un échantillon de 70 personnes on obtient un *Phi*² = 0.10. Le test du *Khi*² indique *p* = 1 %. Que signifie cette valeur ? Parmi les propositions suivantes (A, B, C, D, E) deux seulement sont correctes. Indiquer lesquelles sont correctes en entourant ci-dessous les deux lettres correspondantes :

[A] **B** **C** **[D]** **E**

A. Si on tirait tous les échantillons possibles de 70 personnes d'une population parente où il n'existe pas de liaison entre les deux variables *J* et *K*, on trouverait 1 % des échantillons avec un *Phi*² supérieur ou égal à 0.10.

B. Si on tirait tous les échantillons possibles de 70 personnes d'une population parente où il n'existe pas de liaison entre les deux variables *J* et *K*, on trouverait 1 % des échantillons avec un *Phi*² inférieur à 0.10.

C. Sachant que *Phi*² = 0.10 dans notre échantillon de 70 personnes, il y a seulement une chance sur 100 que l'hypothèse nulle soit vraie (qu'il n'y ait pas de liaison entre *J* et *K* dans la population parente).

D. Si l'hypothèse nulle est vraie (s'il n'y a pas de liaison entre *J* et *K* dans la population parente) et que l'on tire au hasard un échantillon de 70 personnes, on a une chance sur 100 d'obtenir un échantillon avec un *Phi*² au moins égal à 0.10

E. Sachant que *Phi*² = 0.10 dans notre échantillon de 70 personnes, il y a une chance sur 100 que l'hypothèse nulle soit fausse (qu'il y ait une liaison entre *J* et *K* dans la population parente).

Formules (3 pts)

Soit deux variables numériques *X* et *Y* observées sur un échantillon de *n* individus
La formule de définition de la corrélation de *Bravais-Pearson*, notée ici *r*, est :

$$r = \frac{Cov(X,Y)}{Ety(X) \times Ety(Y)} \qquad \text{avec} \qquad Cov(X,Y) = \frac{\sum (x - Moy(X)) \times (y - Moy(Y))}{n}$$

Démontrer que si *X* et *Y* sont des variables centrées-réduites (ou "scores *z*"), la formule se simplifie et peut s'écrire :

$$r = \frac{\sum (x \times y)}{n}$$

1/ Une variable centrée-réduite a un écart-type égal à 1.

On a donc Ety(X) = 1 et Ety(Y) = 1

D'où $r = \frac{Cov(X,Y)}{1 \times 1} = Cov(X,Y)$

2/ Une variable centrée-réduite a une moyenne égale à 0.

On a donc Moy(X) = 0 et Moy(Y) = 0

D'où : $Cov(X,Y) = \frac{\sum (x-0) \times (y-0)}{n} = \frac{\sum (x \times y)}{n}$

3/ Par conséquent :

$$r = Cov(X,Y) = \frac{\sum (x \times y)}{n}$$

Dossier CES

On a fait passer un ensemble de tests d'aptitudes à 42 adolescents de classes de 5ème. On trouve dans le tableau ci-dessous un extrait des résultats obtenus à un sous-ensemble de ces tests :

- un test d'aptitude spatiale (*S*)
- un test d'aptitude numérique (*N*)
- un test d'aptitude verbale (*V*)

Ces adolescents (garçons et filles) ont été tirés au hasard dans les deux collèges (*CES*) d'une ville de la région parisienne :

- un collège situé en Zone d'Éducation Prioritaire (ZEP) que l'on qualifiera de "Défavorisé" (1=Def).
- un collège non défavorisé que l'on qualifiera, pour simplifier, de "Favorisé" (2=Fav).

L'échantillon comprend 14 adolescents provenant du CES défavorisé et 28 adolescents provenant du collège favorisé.

On s'intéressera ici uniquement aux variables *N*, *V* et *CES.*

On trouve ci-dessous un extrait du tableau des données :

	S	N	V	SEX	CES
AGN	43	12	16	2	1
BAK	3	6	15	1	1
BAR	26	15	23	1	2
BED	20	12	16	2	2
BEN	24	12	13	1	1
BOR	54	10	17	1	1
CAR	21	9	14	2	2
CEM	50	8	13	1	1
CHA	17	10	19	2	2
…	…	…	…	…	…
GRU	16	11	15	2	2
HAL	26	6	14	2	1
HER	20	11	25	2	2
KIR	35	6	10	2	1
LAB	17	13	20	1	2
LAE	9	6	12	2	1
LIO	38	5	9	2	1
LOM	33	15	23	1	2
LOU	0	7	15	2	2
MAR	19	6	8	2	1
MOR	9	10	17	1	2
NOR	0	5	10	2	1
NOU	36	11	19	1	2
PAP	0	13	8	1	2
PIC	6	11	14	2	2
RET	31	12	17	1	2
RIC	46	15	18	1	2
ROB	29	11	21	1	2
SAG	10	13	21	1	2
SCH	33	15	15	1	2
THI	17	6	19	2	2
VIE	4	2	10	2	1
VIT	51	15	21	1	2

Source : Adapté de Rozencwajg, P. (2005). *Pour une approche intégrative de l'intelligence : Un siècle après Binet*. Paris : L'Harmattan - Collection Mouvement des Savoirs.